Informatica®
10.2.1

Data Discovery Guide

Informatica Data Discovery Guide
10.2.1
May 2018

Publication Date: 2019-04-23

Table of Contents

Preface

The Informatica *Data Discovery Guide* is written for Informatica Analyst and Informatica Developer users. It contains information about how you can use profiles to analyze the content, quality, and structure of data sources and perform data discovery to discover the metadata of source systems that include content and structure.

Use profiles to discover data quality issues in a data source and to understand the relationships between columns in one or more data sources.

Informatica Resources

Informatica Network

Informatica Network hosts Informatica Global Customer Support, the Informatica Knowledge Base, and other product resources. To access Informatica Network, visit https://network.informatica.com.

As a member, you can:

- Access all of your Informatica resources in one place.
- Search the Knowledge Base for product resources, including documentation, FAQs, and best practices.
- View product availability information.
- Review your support cases.
- Find your local Informatica User Group Network and collaborate with your peers.

Informatica Knowledge Base

Use the Informatica Knowledge Base to search Informatica Network for product resources such as documentation, how-to articles, best practices, and PAMs.

To access the Knowledge Base, visit https://kb.informatica.com. If you have questions, comments, or ideas about the Knowledge Base, contact the Informatica Knowledge Base team at KB_Feedback@informatica.com.

Informatica Documentation

To get the latest documentation for your product, browse the Informatica Knowledge Base at https://kb.informatica.com/_layouts/ProductDocumentation/Page/ProductDocumentSearch.aspx.

If you have questions, comments, or ideas about this documentation, contact the Informatica Documentation team through email at infa_documentation@informatica.com.

Informatica Product Availability Matrixes

Product Availability Matrixes (PAMs) indicate the versions of operating systems, databases, and other types of data sources and targets that a product release supports. If you are an Informatica Network member, you can access PAMs at https://network.informatica.com/community/informatica-network/product-availability-matrices.

Informatica Velocity

Informatica Velocity is a collection of tips and best practices developed by Informatica Professional Services. Developed from the real-world experience of hundreds of data management projects, Informatica Velocity represents the collective knowledge of our consultants who have worked with organizations from around the world to plan, develop, deploy, and maintain successful data management solutions.

If you are an Informatica Network member, you can access Informatica Velocity resources at http://velocity.informatica.com.

If you have questions, comments, or ideas about Informatica Velocity, contact Informatica Professional Services at ips@informatica.com.

Informatica Marketplace

The Informatica Marketplace is a forum where you can find solutions that augment, extend, or enhance your Informatica implementations. By leveraging any of the hundreds of solutions from Informatica developers and partners, you can improve your productivity and speed up time to implementation on your projects. You can access Informatica Marketplace at https://marketplace.informatica.com.

Informatica Global Customer Support

You can contact a Global Support Center by telephone or through Online Support on Informatica Network.

To find your local Informatica Global Customer Support telephone number, visit the Informatica website at the following link:
http://www.informatica.com/us/services-and-training/support-services/global-support-centers.

If you are an Informatica Network member, you can use Online Support at http://network.informatica.com.

Part I: Introduction to Data Discovery

This part contains the following chapters:

CHAPTER 1

Introduction to Profiling

This chapter includes the following topics:

Profiling Overview

Use profiling to find the content, quality, and structure of data sources of an application, schema, or enterprise. The data source content includes value frequencies and data types. The data source structure includes keys and functional dependencies.

As part of the discovery process, you can create and run profiles. A profile is a repository object that finds and analyzes all data irregularities across data sources in the enterprise and hidden data problems that put data projects at risk. Running a profile on any data source in the enterprise gives you a good understanding of the strengths and weaknesses of its data and metadata.

You can use Informatica Analyst and Informatica Developer to analyze the source data and metadata. Analysts and developers can use these tools to collaborate, identify data quality issues, and analyze data relationships. Based on your job role, you can use the capabilities of either the Analyst tool or Developer tool. The degree of profiling that you can perform differs based on which tool you use.

You can perform the following tasks in both the Developer tool and Analyst tool:

- Perform column profiling. The process includes discovering the number of unique values, null values, and data patterns in a column.
- Perform data domain discovery. You can discover critical data characteristics within an enterprise.
- Curate profile results including data types, data domains, primary keys, and foreign keys.
- Create scorecards to monitor data quality.
- Choose an operating system profile to create and run column profiles, enterprise discovery profiles, and scorecards based on the permissions of the operating system user that you define in the operating system profile.
- Use repository asset locks to prevent other users from overwriting work.
- Use version control system to save multiple versions of a profile.
- Create and assign tags to data objects.

- Look up the meaning of an object name as a business term in the Business Glossary Desktop. For example, you can look up the meaning of a column name or profile name to understand its business requirement and current implementation.

You can perform the following tasks in the Developer tool:

- Discover the degree of potential joins between two data columns in a data source.
- Determine the percentage of overlapping data in pairs of columns within a data source or multiple data sources.
- Compare the results of column profiling.
- Generate a mapping object from a profile.
- Discover primary keys in a data source.
- Discover foreign keys in a set of one or more data sources.
- Discover functional dependency between columns in a data source.
- Run data discovery tasks on a large number of data sources across multiple connections. The data discovery tasks include column profile, inference of primary key and foreign key relationships, data domain discovery, and generating a consolidated graphical summary of the data relationships.

You can perform the following tasks in the Analyst tool:

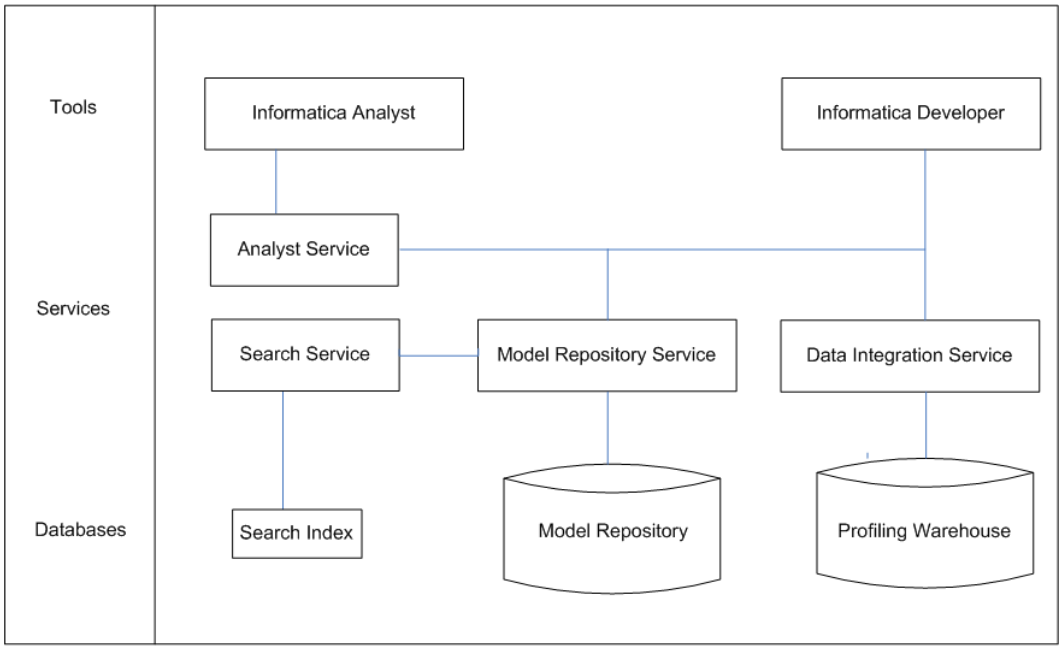
- Perform enterprise discovery on a large number of data sources across multiple connections. You can view a consolidated discovery results summary of column metadata and data domains.
- Perform discovery search to find where the data and metadata exists in the enterprise. You can search for specific assets, such as data objects, rules, and profiles. Discovery search finds assets and identifies relationships to other assets in the databases and schemas of the enterprise.
- View the profile results for a historical profile run.
- Compare the profile results for two profile runs in a column profile.
- View scorecard lineage for each scorecard metric and metric group.
- View the scorecard dashboard.
- Add comments to a profile or columns in a profile.
- Assign tags to a profile or columns in a profile.
- Assign business terms to columns in a profile.

Profiling Architecture

The profiling architecture consists of tools, services, and databases. The tools component consists of client applications. The services component has application services required to manage the tools, perform the

data integration tasks, and manage the metadata of profile objects. The databases component consists of the Model repository and profiling warehouse.

The following image shows the architecture components for profiling:



When you run a profile, the Analyst Service or Developer tool receives the profile definition from the Model Repository Service. Then, the Analyst Service or Developer tool invokes the profiling plug-in in the Data Integration Service. Next, the profiling plug-in processes the profile job and submits the job to the Data Integration Service. The Data Integration Service generates the profiling results. The Data Integration Service then writes the profiling results to the profiling warehouse.

Discovery search uses the Search Service. The Search Service performs each search on a search index instead of the Model repository or profiling warehouse. The Search Service generates the search index based on content in the Model repository and profiling warehouse. The Search Service contains extractors to extract content from each repository.

The following table describes the architecture components:

Component	Description
Informatica Analyst	A web-based client application that you can use to discover, analyze, and report on data and metadata of data sources.
Informatica Developer	A client application that you use to perform advanced data discovery, such as primary key discovery, foreign key discovery, and enterprise discovery.
Analyst Service	An application service that runs the Analyst tool and manages connections between service components and Analyst tool users.
Search Service	An application service that manages search in the Analyst tool. By default, the Search Service returns search results from the Model repository, such as data objects, profiles, mapping specifications, reference tables, rules, and scorecards.
Search Index	A file system in a custom directory that stores indexed content that the Search Service extracts from the Model repository and profiling warehouse.

Component	Description
Model Repository Service	An application service that manages the Model repository.
Data Integration Service	An application service that performs data integration tasks for the Analyst tool, the Developer tool, and external clients.
Model repository	A relational database that stores the metadata for projects created in the Analyst tool or Developer tool.
Profiling warehouse	A database that stores profiling information, such as profile results and scorecard results.

Data Discovery Process

When you begin a data integration project, profiling is often the first step. You can create profiles to analyze the content, quality, and structure of data sources. As a part of the profiling process, you discover the metadata of data sources.

You use different profiles for different types of data analysis, such as a column profile, primary key discovery, foreign key discovery, and data domain discovery. You uncover and document data quality issues. Complete the following tasks to perform data discovery:

1. Find and analyze the content of data in the data sources. Includes data types, value frequency, pattern frequency, and data statistics, such as minimum value and maximum value.
2. Discover the structure of data. Includes keys, functional dependencies, and foreign keys.
3. Review and validate profile results.
4. Drill down on profile results.
5. Curate profile results.
6. Create reference data.
7. Document data issues.
8. Create and run rules.
9. Create scorecards to monitor data quality.

You can use the following tools to manage the discovery process:

Informatica Administrator

Manage users, groups, privileges, and roles. You can administer the Analyst service and manage permissions for projects and objects in Informatica Analyst. You can control the access permissions in Informatica Developer using this tool.

Informatica Developer

Create and run profiles in this tool to find and analyze the metadata of one or more data sources including discovering the relationships between columns. You create profiles using a wizard.

Informatica Analyst

You can run a column profile, perform data domain discovery, and perform enterprise discovery on data objects in the Analyst tool. After you run a profile, you can drill down on data rows in a data source.

CHAPTER 2

Data Discovery

This chapter includes the following topics:

Data Discovery Overview

Data discovery is the process of discovering the metadata of source systems that include content and structure. Content refers to data values, frequencies, and data types. Structure includes candidate keys, primary keys, foreign keys, and functional dependencies. You can create and run profiles to discover the content and structure of data sources.

You can define a profile to analyze data in a single data object or across multiple data objects. Add comments to profiles so that you can track the profiling process effectively.

Run a profile to evaluate the data structure and to verify that data columns contain the types of information you expect. You can drill down on data rows in profiled data. If the profile results reveal problems in the data, you can apply rules to fix the result set. You can create scorecards to track and measure data quality before and after you apply the rules. If the external source metadata of a profile or scorecard changes, you can synchronize the changes with its data object.

Profile and Analysis Types

Create a profile based on the type of analysis that you need to perform. The type of profile that you create corresponds to the type of analysis that you perform. For example, to perform a primary key analysis, you create a primary key profile.

You can create the following profiles to perform data analysis and discovery:

Column Profile

Analyzes data quality in selected columns in a table or file. You can define profiles for column analysis in the Analyst tool and Developer tool.

Data Domain Discovery

Discovers critical data characteristics within an enterprise. Data domain discovery identifies all the data domains associated with a column based on the column value or name. As part of the discovery process, you can manually create data rules and column name rules to verify whether a value or column name belongs to a data domain. You can then associate these rules when you create a data domain. You can also create data domains from the values and patterns in column profile results.

Primary Key Profile

Discovers primary key relationships between columns in a table or file. You can define profiles for primary key analysis in the Developer tool.

Functional Dependency Profile

Discovers functional dependencies between columns in a table or file. You can define profiles for functional dependency analysis in the Developer tool.

Foreign Key Profile

Discovers foreign key relationships between columns across multiple tables or multiple files. You can define profiles for foreign key analysis in the Developer tool.

Join Profile

Determines the degree of potential joins between columns in a data source or across multiple data sources. You can define profiles for join analysis in the Developer tool. The results appear in a Venn diagram.

Overlap Discovery

Determines the percentage of overlapping data in pairs of columns within a data source or multiple data sources. You can run the overlap discovery task from the editor in the Developer tool. You can validate the results and view them in a Venn diagram.

Enterprise Discovery

Runs multiple data discovery tasks on a large number of data sources and generates a consolidated summary of the profile results. Includes running a column profile, data domain discovery, and discovering primary key and foreign key relationships. Enterprise discovery automates the profile process for a large number of data sources.

Note: Changes that you make to profiles in the Analyst tool do not appear in the Developer tool until you refresh the Developer tool connection to the Model repository.

Profiling Components

A profile has multiple components that you can use to effectively analyze the content and structure of data sources.

A profile has the following components:

Filter

Creates a subset of the original data source that meets specific criteria. You can then run a profile on the sample data.

Rule

Business logic that defines conditions applied to data when you run a profile. Add a rule to the profile to validate the data.

Tag

Metadata that defines an object in the Model repository based on business usage. Create tags to group objects according to their business usage. Assign tags to a profile or columns in a profile in the Analyst tool.

Comment

Description about the profile. Use comments to share information about profiles with other Analyst and Developer tool users. Add comments to a profile or columns in a profile in the Analyst tool.

Scorecard

A graphical representation of valid values for a column or the output of a rule in profile results. Use scorecards to measure data quality progress.

Profile Results

You can view the profile results after you run a profile. You can view a summary of values, patterns, and statistics for columns and rules in the profile. You can view properties for the columns and rules in the profile. You can preview profile data.

The following table describes the profile results for each profile type:

Profile Type	Results
Column profile	- Number and percentage of null, distinct, and non-distinct values in columns and the inferred data types for column values. - Frequency and character patterns of data values in a selected column and a statistical summary for the column. - Data types inferred by analyzing column data. - Documented data type for the data. - Maximum and minimum values. - Date and time of the profile run.
Primary key profile	- Number and percentage of unique, duplicate, and null values for inferred primary key candidates. - Number of key violations in the inferred primary key candidates.
Functional dependency profile	- Inferred functional dependencies. - Number of functional dependency violations.
Foreign key profile	- Primary and foreign key columns that meet the primary-foreign key inference criteria you defined. - Number of data values that match between the primary and foreign keys, expressed as a percentage. - Type of relationship defined for the primary and foreign key columns before the profile run.
Join profile	- Venn diagram that shows the relationships between columns. - Number and percentage of orphaned, null, and joined values in columns.
Overlap discovery	- Percentage of ovelap between two columns. - Venn diagram that shows the overlap between columns.

Profile Type	Results
Data domain discovery	- Column name and data that match predefined data domains. - Data domain group that the column belongs to and its data type.
Enterprise discovery	- Column profile results. - Data domain discovery results. - Primary key discovery results. - Foreign key profile results in both graphical and tabular views.

You can use third-party reporting tools to read profile results from the profiling warehouse. Informatica provides a set of profile views that you can customize for the profile statistics that you want to read. These views are based on common types of profile statistics and profile results analysis.

CHAPTER 3

Column Profile Concepts

This chapter includes the following topics:

Column Profile Concepts OverviewProfiles Overview

A column profile determines the characteristics of columns in a data source, such as value frequency, percentages, and patterns.

Column profiling discovers the following facts about data:

- The number of null, distinct, and non-distinct values in each column, expressed as a number and a percentage.
- The patterns of data in each column and the frequencies with which these values occur.
- Statistics about the column values, such as the maximum and minimum lengths of values and the first and last values in each column.
- Documented data types, inferred data types, and possible conflicts between the documented and inferred data types.
- Pattern and value frequency outliers.

You can configure the following options when you create or edit a profile:

- Column profile options. You can select the columns on which you want to run a profile, choose a sampling option, and drill-down option.
- Add, edit, or delete filters and rules.

In the profile results, you can add comments and tags to a profile and to the columns in a profile. You can assign business terms to columns.

The Model repository locks profiles to prevent users from overwriting work with the repository profile locks. The version control system saves multiple versions of a profile and assigns a version number to each version. You can check out a profile and then check the profile in after making changes. You can undo the action of checking out a profile before you check the profile back in.

Create scorecards to periodically review data quality. You create scorecards before and after you apply rules to profiles so that you can view a graphical representation of the valid values for columns.

Use the Scheduler Service to schedule profile runs and scorecard runs to run at a specific time or intervals. The Scheduler Service manages schedules for profiles, scorecards, deployed mappings, and deployed workflows. You can create, manage, and run schedules in Informatica Administrator.

Column Profile Options

When you create a profile, you can use the profile wizard to define filters, rules, drill-down options, sampling options, and connection. These options determine how the profile reads rows from the source data.

You can define the following options in a column profile, data domain discovery profile, or an enterprise discovery profile:

- Filters. You can create and apply filters to a profile.
- Rules. You can add rules when you create a profile. You can reuse the rules that you create in the Analyst tool or Developer tool.
- Drill-down options. You can choose to read current data in the data source or read profile data that is staged in the profiling warehouse.
- Sampling options. You can choose one of the sampling options to determine the number of rows to run a profile on.
- Connection. You can run the profiles in the native or Hadoop run-time environment.

Repository Profile Locks and Versioned Profile Management

The Model repository locks profiles to prevent users from overwriting work. When you begin to edit a profile, the profile is locked to prevent other users from saving changes to it. The lock is released when you save the profile. Versioned profile management creates versions of a profile, and you can view version history.

The Model repository locks a profile when you edit it in the Developer tool or Analyst tool. If the tool stops unexpectedly, the lock is retained, so that when you connect to the Model repository again, you can view the profiles that you have locked. You can continue to edit the profiles, or you can unlock the profiles.

When the Model repository is integrated with a version control system, you can manage versions of a profile. For example, you can check out and check in profiles, undo checkouts, view specific historic versions of the profile, and view the profiles that you have checked out. For information about repository asset locks and versioned asset management in the Analyst tool, see the *Analyst Tool Guide*. For information about repository object locks and versioned object management in the Developer tool, see the *Developer Tool Guide*.

Scorecards

A scorecard is the graphical representation of the valid values for a column or output of a rule in profile results. Use scorecards to measure data quality progress. You can create a scorecard from a profile and monitor the progress of data quality over time.

A scorecard has multiple components, such as metrics, metric groups, and thresholds. After you run a profile, you can add source columns as metrics to a scorecard and configure the valid values for the metrics. Scorecards help the organization to measure the value of data quality by tracking the cost of bad data at the metric and scorecard levels. To measure the cost of bad data for each metric, assign a cost unit to the metric and set a fixed or variable cost. When you run the scorecard, the scorecard results include the cost of bad data for each metric and total cost value for all the metrics.

Use a metric group to categorize related metrics in a scorecard into a set. A threshold identifies the range, in percentage, of bad data that is acceptable to columns in a record. You can set thresholds for good, acceptable, or unacceptable ranges of data.

When you run a scorecard, configure whether you want to drill down on the score metrics on live data or staged data. After you run a scorecard and view the scores, drill down on each metric to identify valid data records and records that are not valid. You can also view scorecard lineage for each metric or metric group in a scorecard. To track data quality effectively, you can use score trend charts and cost trend charts. These charts monitor how the scores and cost of bad data change over a period of time.

The profiling warehouse stores the scorecard statistics and configuration information. You can configure a third-party application to get the scorecard results and run reports. You can also display the scorecard results in a web application, portal, or report, such as a business intelligence report.

CHAPTER 4

Data Domain Discovery Concepts

This chapter includes the following topics:

Data Domain Discovery Concepts Overview

You need to identify and understand the meaning of critical source data so that you can take measures to work effectively on it. Data domain discovery is the process of discovering the functional meaning of data in the data sources based on the semantics of data.

Create a profile to perform data domain discovery and you can identify critical data characteristics within an enterprise. You can then apply further data management policies, such as data quality or data masking, to the data. For example, discover product codes or descriptions to analyze which data quality standardization or parsing rules you need to apply to make the data useful and trustworthy. Another example is to find sensitive customer data, such as credit card numbers, email IDs, and phone numbers. You may then want to mask this information to protect it.

You can create and run a profile to perform data domain discovery in both Analyst and Developer tools. You can define a profile to perform data domain discovery based on the following rules:

- Data rule. Finds columns with data that matches specific logic defined in the rule.
- Column name rule. Finds columns that match column name logic defined in the rule.

You can create data domains from the values and patterns in column profile results. You can then use these data domains to discover critical data across multiple data systems or enterprise.

You can create a profile with a sampling option and filters to perform data domain discovery. When you run the profile, you apply the sampling option and filters on the data source and generate a data set. The data domain discovery process uses the data set to discover data domains.

Data Domains

A data domain is a predefined or user-defined Model repository object based on the semantics of column data or a column name. For example, Social Security number, credit card number, email ID, and phone number can be individual data domains.

A data domain helps you find important data that remains undiscovered in a data source. For example, you may have legacy data systems that contain Social Security numbers in a Comments field. You need to find this information and protect it before you move it to new data systems.

You can choose a minimum percentage of source rows or minimum number of source rows as a conformance criteria for data domain match. You can also exclude null values when you perform data domain discovery in a column profile.

You can group logical data domains into data domain groups. A data domain glossary lists all the data domains and data domain groups. Use the Preferences menu in the Developer tool to import and export data domains to and from the data domain glossary.

You use rules to define data and column name patterns that match source data and metadata. When you create a data domain, the Analyst tool or Developer tool copies associated rules and other dependent objects to the data domain glossary. Use the Developer tool to manage data domains that includes import and export of data domains to and from the data domain glossary. You can also use the Developer tool to manage the rule logic of data domains.

Note: You may want to save all the data domain rules in a single project or folder. This step helps after you export data domains and you have a need to edit the rules and other associated data objects.

Data Domain Groups

Data domain groups help you categorize data domains into specific groups. For example, you can group the data domains first_name, last_name, and account_number under the Personal Health Information (PHI) data domain group.

You can create a Personally Identifiable Information (PII) data domain group that includes the Social Security number, first name, and last name. A data domain can be a part of multiple data domain groups. For example, the Social Security number can belong to both Payment Card Industry (PCI) and PII data domain groups. Data domain groups can contain data domains and not other data domain groups.

Note: If you import the data domain file `Informatica_IDE_DataDomain.xml` after installation, the data domain glossary displays predefined data domain groups and data domains. You can then create more data domain groups as required. To view and change the rules associated with the data domains, import the `Informatica_IDE_DataDomainRule.xml` file.

Data Domain Glossary

The data domain glossary is a container for all the domain groups and data domains. You can use the data domain glossary to create, manage, and remove data domains and data domain groups.

You can search for specific domains and domain groups within the data domain glossary. You can also export data domains to an XML file and import data domains from an XML file to the data domain glossary.

The data domain glossary contains copied rules and all the reference data associated with data domains. You cannot edit the rules in the data domain glossary.

You can view the data domain glossary from the Preferences menu in the Developer tool and from the Manage menu in the Analyst tool. Use the Model Repository Service privilege **Manage Data Domains** to determine who creates, edits, and deletes data domains and data domain groups.

Data Domain Discovery Process

You can define and run a profile to perform data domain discovery in the Analyst tool or Developer tool based on your job role. After you have configured the data domain discovery options and run the profile, you can verify and drill down on the results. If you run data domain discovery from within the editor, you can add the results to a data model.

Complete the following steps to perform data domain discovery:

1. Create or import data domains and domain groups.
2. Optionally, consolidate data domains under appropriate domain groups.
3. Create a profile to perform data domain discovery. You start by choosing whether you want to run a column profile along with data domain discovery or only data domain discovery.
4. Select the columns, domains, and the appropriate sampling options.
5. Run the profile.
6. Verify, drill down on profile results, and add results to a data model as required.

CHAPTER 5

Curation Concepts

This chapter includes the following topics:

Curation Concepts Overview

Curation is the process of validating and managing discovered metadata of a data source so that the metadata is fit for use and reporting.

You can curate the following inferred profile results:

- Data types
- Data domains
- Primary keys
- Foreign keys

You curate inferred profile results to make the metadata about columns, data domains, and data object relationships in the databases and schemas accurate. You can then find the most relevant metadata when you use discovery search to search for information across multiple repositories. You can also find the most relevant metadata when you view the foreign key relationship diagram in the enterprise discovery results.

You can curate specific metadata inferences that a profile generates as part of the profile run. For example, you can approve or reject the inferred data types in the column profile results and data domain discovery results. You can also approve or reject the inferred primary keys and foreign keys in enterprise discovery results.

Curation for Analysts and Developers

As a data analyst or data steward, you can curate the column profile results and data domain discovery results in the Analyst tool. You can curate the profile results to make accurate profile information ready for discovery search and further validation of the data assets.

As a developer or data architect, you can curate column profile results, data domain discovery results, primary key discovery results, and foreign key discovery results in the Developer tool.

Curation Examples

When you perform enterprise discovery as a developer, the Developer tool processes the selected data domains for the entire data set. This action can result in multiple data domain inferences, such as phone number data inferred as the Social Security number data domain. Multiple data domain inferences occur when parts of the data within a column match different data domains. For example, a 10-digit phone number that is missing one digit might have the same pattern as a Social Security number. This occurrence indicates potential data quality issues within a column or a matching pattern across multiple data domains. In this case, the Developer tool might infer both phone number data domain and Social Security number data domain. You can curate the profile results so that you can select the most appropriate data domain and approve it. In the example, phone number is the relevant data domain because the inference of Social Security number data domain occurs due to a data quality issue.

When you run enterprise discovery, the Developer tool might infer multiple data types, such as Date, String, and Varchar, for a date column. As a data architect, you might want to choose and approve the Date data type, which is the most relevant data type for a date column.

Enterprise discovery in the Developer tool might infer all the data object relationships based on the column data. Some of these data object relationships include unwanted data object relationships in the discovered candidate keys. For example, the Developer tool might infer columns that represent a sequence as possible keys and discover relationships with other tables with similar columns. These data object relationships might not form valid relationships in the database. In such cases, you can assess, verify, and approve the most appropriate inferred profile results as part of curation.

Curation Tasks

You can curate profile results after the profile run. You can also reverse a curation decision that you took when you previously ran the profile.

You can perform the following curation tasks in the Analyst tool:

- Approve or reject the inferred data types for multiple columns and data domains.
- Restore approved or rejected data types to the inferred status.
- Restore approved or rejected data domains to the inferred status.
- View or hide rejected result rows.
- Exclude columns from profile runs based on specific metadata preferences, such as approved data types and data domains.

You can perform the following curation tasks in the Developer tool:

- Approve or reject the inferred data types for multiple columns.
- Restore approved or rejected data types to the inferred status.
- Restore approved or rejected data domains to the inferred status.
- View or hide rejected result rows.
- Approve or reject data objects in the primary key discovery results.
- Approve or reject enterprise discovery results, including foreign key discovery results.
- Exclude columns from profile runs based on specific metadata preferences, such as approved data types and data domains.

Part II: Data Discovery with Informatica Analyst

This part contains the following chapters:

CHAPTER 6

Column Profiles in Informatica Analyst

This chapter includes the following topics:

Column Profiles in Informatica Analyst Overview

When you create a profile, you select the columns in the data object on which you want to run a profile. You can configure the sampling and drill-down options for faster profiling. You can choose a run-time environment. When you create a profile, you can add rules and filters to the profile. After you run the profile, you can examine the profiling statistics to understand the data.

You can profile wide tables and flat files that have a maximum of 1000 columns. When you create or run a profile, you can choose to select all the columns or select each column for a profile. You can select all columns to drill down and view value frequencies for these columns.

You can create column profiles with the following methods in Informatica Analyst:

- Right-click the data object in the **Library** workspace to create a profile.
- Use default options to create a default column profile.
- Customize the settings for the profile to create a custom profile.

Note: You can view and run the profile on Avro, JSON, Parquet, and XML data sources. You can create and edit a column profile on Avro, JSON, Parquet, and XML data sources in the Informatica Developer.

Column Profiling Process

As part of the column profiling process, you can choose to either include all the source columns for profiling or select specific columns. You can also accept the default profile options, or configure the sampling options, drill-down options, and run-time environment.

The following steps describe the column profiling process:

1. Choose a name, description, and location for the column profile.
2. Select an imported data object or an external source that you want to run the profile on.
3. Optionally, preview the source data.
4. Select the columns you want to run the profile on.
5. Determine whether you want to create the profile with the default options or change the default options. The options that you can configure include sampling options, drill-down options, and run-time environment.
6. Optionally, add rules and filters when you create the profile.
7. Run the profile.

Note: Consider the following rules and guidelines for column names and profiling multilingual and Unicode data:

- You can profile multilingual data from different sources and view profile results based on the locale settings in the browser. The Analyst tool changes the Datetime, Numeric, and Decimal data types based on the browser locale.
- Sorting on multilingual data. You can sort on multilingual data. The Analyst tool displays the sort order based on the browser locale.
- To profile Unicode data in a DB2 database, set the DB2CODEPAGE database environment variable in the database and restart the Data Integration Service.

Profile Options

Profile options include data sampling options and data drill-down options. You can configure these options when you create or edit a column profile for a data object.

You use the **Discovery** workspace to configure the profile options. You can choose to create a profile with the default options for columns, sampling, and drill-down options. Use the drill-down option to choose between live data and staged data.

Sampling Options

Sampling options determine the number of rows that the Analyst tool chooses to run a profile on. You can configure sampling options when you define a profile or when you run a profile.

The following table describes the sampling options for a profile:

Option	Description
All rows	Chooses all rows in the data object.
Sample first <number> rows	The number of rows that you want to run the profile against. The Analyst tool chooses the rows from the first rows in the source.
Random sample <number> rows	The random sample algorithm chooses the rows at random in the data object to run the profile on.
Random sample (auto)	Random sample size is computed based on the number of rows in the data object.
Exclude approved data types and data domains from the data type and data domain inference in the subsequent profile runs	Excludes the approved data type or data domain from data type and data domain inference from the next profile run.

After you choose to run the profile on a random sample of rows, the random sample algorithm chooses the rows at random in the data object to run the profile on. When you choose a random sampling option for column profiles, the Analyst tool performs drilldown on the staged data. This can impact the drill-down performance. When you choose a random sampling option for data domain discovery profiles, the Analyst tool performs drill down on live data.

Drilldown Options

You can configure drilldown options when you define a profile or when you edit a profile.

The following table describes the drilldown options for a profile:

Options	Description
Live	Drills down on live data to read current data in the data source.
Staged	Drills down on staged data to read profile data that is staged in the profiling warehouse.
Select Columns	Identifies columns for drilldown that you did not select for profiling.

Run-time Environment

You can choose native, Hive, or Hadoop as the run-time environment for a column profile. Informatica Analyst sets the run-time environment in the profile definition after you choose a run-time environment.

Native Environment

When you run a profile in the native run-time environment, the Analyst tool submits the profile jobs to the Profiling Service Module. The Profiling Service Module then breaks down the profile jobs into a set of mappings. The Data Integration Service runs these mappings on the same machine where the Data Integration Service runs and writes the profile results to the profiling warehouse. By default, all profiles run in the native run-time environment.

You can use native sources to create and run profiles in the native environment. A native data source is a non-Hadoop source, such as a flat file, relational source, or mainframe source. You can also run a profile on a mapping specification or a logical data source with a Hive or HDFS data source in the native environment.

Hadoop Environment

You can choose Hive or Hadoop option to run the profiles in the Hadoop run-time environment.

If you choose the Hive option and select a Hadoop connection, the Data Integration Service pushes the profile logic to the Hive engine on the Hadoop cluster to run profiles. The Data Integration service executes only the environment SQL of the Hive connection. If the Hive sources and targets are on different clusters, the Data Integration Service does not execute the different environment SQL commands for the connections of the Hive source or target.

If you choose the Hadoop option and select a Hadoop connection, the Data Integration Service pushes the profile logic to the Blaze engine on the Hadoop cluster to run profiles.

When you run a profile in the Hadoop environment, the Developer tool submits the profile jobs to the Profiling Service Module. The Profiling Service Module then breaks down the profile jobs into a set of mappings. The Data Integration Service pushes the mappings to the Hadoop environment through the Hadoop connection. The Hive engine or Blaze engine processes the mappings and the Data Integration Service writes the profile results to the profiling warehouse.

Note: Hive engine is deprecated, and Informatica will drop support for it in a future release. You can still choose to run profiles on the Hive engine. In a future release, when Informatica drops support for the Hive engine, the Data Integration Service will ignore the Hive engine selection and run the profile on the Blaze engine.

Column Profiles for Sqoop Data Sources

You can run a column profile on data objects that use Sqoop. You can select the Hive or Hadoop run-time environment to run the column profiles.

On the Hive engine, to run a column profile on a relational data object that uses Sqoop, you must set the Sqoop argument **m** to 1 in the JDBC connection. Use the following syntax:

```
-m 1
```

When you run a column profile on a logical data object or customized data object, you can configure the num-mappers argument to achieve parallelism and optimize performance. You must also configure the split-by argument to specify the column based on which Sqoop must split the work units.

Use the following syntax:

```
--split-by <column_name>
```

If the primary key does not have an even distribution of values between the minimum and maximum range, you can configure the split-by argument to specify another column that has a balanced distribution of data to split the work units.

If you do not define the split-by column, Sqoop splits work units based on the following criteria:

- If the data object contains a single primary key, Sqoop uses the primary key as the split-by column.
- If the data object contains a composite primary key, Sqoop defaults to the behavior of handling composite primary keys without the split-by argument. See the Sqoop documentation for more information.
- If a data object contains two tables with an identical column, you must define the split-by column with a table-qualified name. For example, if the table name is CUSTOMER and the column name is FULL_NAME, define the split-by column as follows:
  `--split-by CUSTOMER.FULL_NAME`
- If the data object does not contain a primary key, the value of the m argument and num-mappers argument default to 1.

When you use Cloudera Connector Powered by Teradata or Hortonworks Connector for Teradata and the Teradata table does not contain a primary key, the split-by argument is required.

Operating System Profiles in Informatica Analyst Overview

You can choose an operating system profile in the Analyst tool. After you choose an operating system profile, the Data Integration Service creates and runs the column profiles, enterprise discovery profiles, and scorecards based on the permission of the operating system profile user.

The Analyst tool uses the default profile to run profiles and scorecards. If you have only one operating system profile, the operating system profile is selected by default. If you have multiple operating system profiles, then you can choose one of the operating system profiles.

Selecting an Operating System Profile

You can select an operating system profile in Informatica Analyst. The Data Integration Service uses the permissions of the operating system profile user to run the profiling jobs.

1. In Informatica Analyst header area, click **<Username>** > **Settings**.

   The **Settings** dialog box appears.
2. Select an operating system profile. Click **Save**.

Repository Asset Locks and Team-based Development Overview

The Model repository locks profiles to prevent users from overwriting the work of another user. If the Model repository is integrated with a version control system, it saves multiple versions of assets and assigns a version number to a version. You can check out and check in profiles and undo checkouts. You can view a specific version of a profile that you have checked out.

When you begin to edit a profile in the Analyst tool, the Model repository locks the profile so that other users cannot edit the profile. When you save the profile, you retain the lock. When you close the profile, the Model repository unlocks the profile.

The Model repository protects profiles from being overwritten by other members of the team with versioned asset management. When you try to edit a profile that another user has checked out, you receive a notification indicating the user who has checked out the profile. You can open a checked out profile in read-only mode, or save the profile with a different name.

You can select a version of the profile in the Profile Properties dialog box to view the profile definition for that version. You can access Profile Properties option in the Actions menu. For more information about repository asset locks and versioned asset management, see the *Analyst Tool Guide*.

Creating a Column Profile in Informatica Analyst

You can create a custom profile or default profile. When you create a custom profile, you can configure the columns, sample rows, and drill-down options. When you create a default profile, the column profile and data domain discovery runs on the entire data set with all the data domains.

1. In the **Discovery** workspace, click **Profile**, or select **New** > **Profile** from the header area.

   Note: You can right-click on the data object in the **Library** workspace and create a profile. In this profile, the profile name, location name, and data object are extracted from the data object properties. You can create a default profile or customize the settings to create a custom profile.

   The **New Profile** wizard appears.
2. The **Single source** option is selected by default. Click **Next**.
3. In the **Specify General Properties** screen, enter a name and an optional description for the profile. In the Location field, select the project or folder where you want to create the profile. Click **Next**.
4. In the **Select Source** screen, click **Choose** to select a data object, or click **New** to import a data object. Click **Next**.
   - In the **Choose Data Object** dialog box, select a data object. Click **OK**.
     The Properties pane displays the properties of the selected data object. The Data Preview pane displays the columns in the data object.
   - In the **New Data Object** dialog box, you can choose a connection, schema, table, or view to create a profile on, select a location, and create a folder to import the data object. Click **OK**.
5. In the **Select Source** screen, select the columns that you want to run a profile on. Optionally, select **Name** to select all the columns. Click **Next**.

   All the columns are selected by default. The Analyst tool lists column properties, such as the name, data type, precision, scale, nullable, and participates in the primary key for each column.
6. In the **Specify Settings** screen, choose to run a column profile, data domain discovery, or a column profile and data domain discovery. By default, column profile option is selected.
   - Choose **Run column profile** to run a column profile.
   - Choose **Run data domain discovery** to perform data domain discovery. In the **Data domain** pane, select the data domains that you want to discover, select a conformance criteria, and select the columns for data domain discovery in the **Edit columns selection for data domin discovery** dialog box.
   - Choose **Run column profile** and **Run data domain discovery** to run the column profile and data domain discovery. Select the data domain options in the **Data domain** pane.

     Note: By default, the columns that you select is for column profile and data domain discovery. Click **Edit** to select or deselect columns for data domain discovery.

- Choose Data, Columns, or Data and Columns to run data domain discovery on.
- Choose a sampling option. You can choose **All rows (complete analysis)**, **Sample first**, **Random sample** , or **Random sample (auto)** as a sampling option in the **Run profile on** pane. This option applies to column profile and data domain discovery.
- Choose a drilldown option. You can choose **Live** or **Staged** drilldown option, or you can choose **Off** to disable drilldown in the **Drilldown** pane. Optionally, click **Select Columns** to select columns to drill down on. You can choose to omit data type and data domain inference for columns with an approved data type or data domain.
- Choose **Native**, **Hive (deprecated)**, or **Hadoop** option as the run-time environment. If you choose the Hive or Hadoop option, click **Choose** to select a Hadoop connection in the **Select a Hadoop Connection** dialog box.

7. Click **Next**.

   The **Specify Rules and Filters** screen opens.
8. In the **Specify Rules and Filters** screen, you can perform the following tasks:
   - Create, edit, or delete a rule. You can apply existing rules to the profile.
   - Create, edit, or delete a filter.

     Note: When you create a scorecard on this profile, you can reuse the filters that you create for the profile.
9. Click **Save and Finish** to create the profile, or click **Save and Run** to create and run the profile.

Editing a Column Profile

You can make changes to a column profile after you run it.

1. In the **Library** workspace, select the project that contains the profile, or select the profile in the **Assets** pane.
2. Click the profile name.

   The summary view appears in the **Discovery** workspace.
3. If the version control system is enabled, click **Actions** > **Check Out** to check out the profile.
4. Click **Actions** > **Edit Profile**.

   The **Profile** wizard appears.
5. Based on the changes you want to make, choose one of the following page options:
   - **Specify General Properties**. Change the basic properties such as name, description, and location.
   - **Select Source**. Choose another matching data source and columns to run the profile on.
   - **Specify Settings**. Choose to run column profile or column profile and data domain discovery. Select the data domains that you want to discover and modify the data domain discovery, sampling, and drilldown options.
   - **Specify Rules and Filters**. Create, edit, or delete rules and filters.
6. Click **Save and Finish** to complete editing the profile, or click **Save and Run** to edit and run the profile.
7. If the version control system is enabled, you must perform the following tasks:
   - Click **Save and Finish** to complete editing the profile.

- In the summary view, click **Check In** to check in the profile.
- Click **Actions** > **Run Profile** to run the profile.

Running a Profile

Run a profile to analyze a data source for content and structure and select columns and rules for drill down. You can drill down on live or staged data for columns and rules. You can run a profile only on a column or rule without running the profile on all the source columns after the initial profile run.

1. In the **Library** workspace, select the project or folder that contains the profile in the Projects pane, or select the profile in the **Assets** pane.
2. Click **Actions** > **Open**.

   The summary view appears in the **Discovery** workspace.
3. Click **Actions** > **Run Profile.**

   The Analyst tool performs a profile run and displays the profile results in summary view.
4. In the summary view, click on a column to view the column results.

   The detailed view appears.

Synchronize Option

When you change the metadata of an external data source, the data object metadata in the Model Repository is not updated by default. Use the Synchronize option to synchronize the data object metadata to the data source metadata. You can use the Synchronize option for column profiles, enterprise discovery profiles, and scorecards. The external data source can be a relational data source or flat file data source.

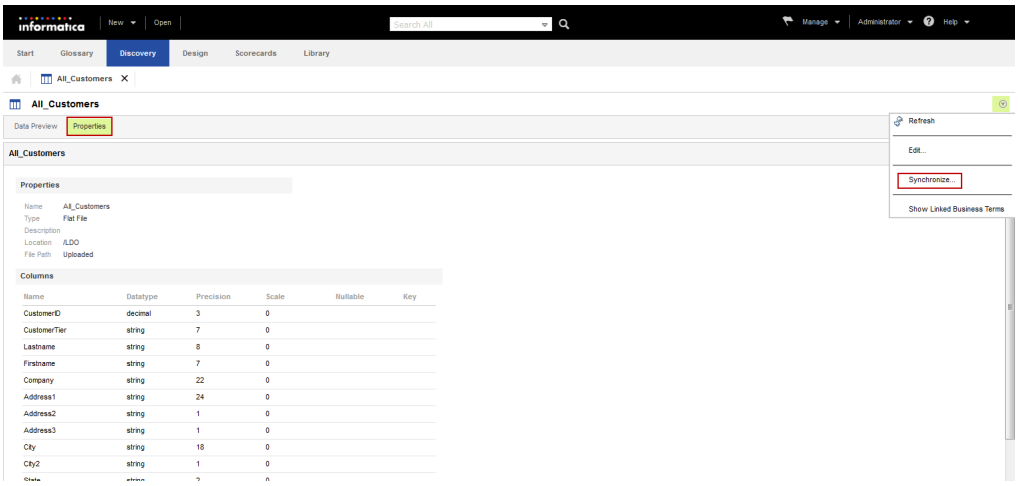
Synchronizing a Flat File Data Object in Informatica Analyst

You can synchronize the changes to an external flat file data source with its data object in the Analyst tool. Use the **Synchronize Flat File** wizard to synchronize the data objects.

1. Open the **Library** workspace.
2. In the **Projects** section, select a flat file data object from a project.

   The Analyst tool displays the data preview for the flat file in the **Data Preview** tab.
3. Click the **Properties** tab.
4. From the Actions menu, click **Synchronize**.

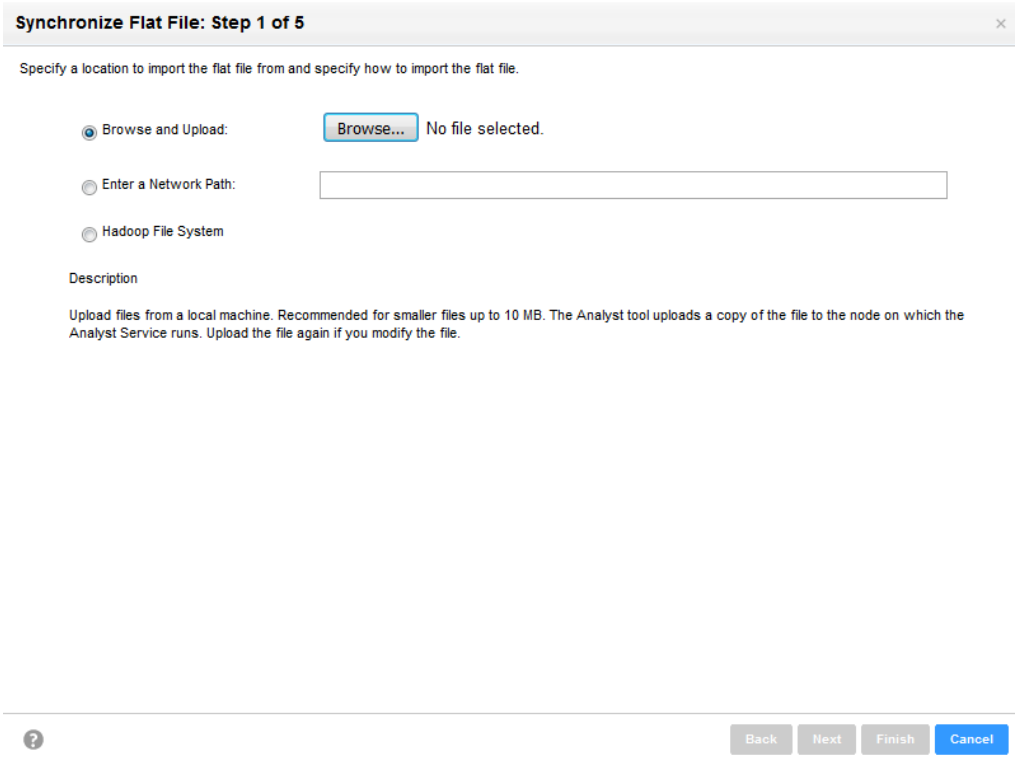
   The following image shows the Properties tab and the Synchronize option in the Actions menu:



The **Synchronize Flat File** wizard appears.

5. Choose to browse for a location or enter a network path to import the flat file.
   - To browse for a location, click **Choose File** to select the flat file from a directory that your machine can access.
   - To enter a network path, select **Enter a Network Path** and configure the file path and file name.

   The following image shows the Synchronize Flat File wizard:



6. Click **Next**.
7. Choose to import a delimited or fixed-width flat file.
   - To import a delimited flat file, accept the **Delimited** option.

- To import a fixed-width flat file, select the **Fixed-width** option.

8. Click **Next**.
9. Configure the flat file options for the delimited or fixed-width flat file.
10. Click **Next**.
11. Optionally, change the column attributes.
12. Click **Next**.
13. Accept the default name or enter another name for the flat file.
14. Optionally, enter a description.
15. Click **Finish**.

    A synchronization message prompts you to confirm the action.
16. Click **Yes** to synchronize the flat file.

    A message that states synchronization is complete appears. To view details of the metadata changes, click **Show Details**.
17. Click **OK**.

Synchronizing a Relational Data Object in Informatica Analyst

You can synchronize the changes to an external relational data source with its table data object. External data source changes include adding, changing, and removing source columns and rule columns.

1. Open the **Library** workspace.
2. In the **Projects** section, select a table data object from a project.

   The Analyst tool displays the data preview for the table on the **Data Preview** tab.
3. Click the **Properties** tab.
4. From the Actions menu, click **Synchronize**.

   A message prompts you to confirm the action.
5. To complete the synchronization process, click **Yes**.

   A synchronization status message appears.
6. A message that states synchronization is complete appears.

   To view details of the metadata changes, click **Show Details**.
7. Click **OK**.

CHAPTER 7

Rules in Informatica Analyst

This chapter includes the following topics:

Rules in Informatica Analyst Overview

A rule is business logic that defines conditions applied to source data when you run a column profile. You can add a rule to the profile to validate data.

You might want to use a rule in different circumstances. You can add a rule to cleanse one or more data columns. You can add a lookup rule that provides information that the source data does not provide. You can add a rule to validate a cleansing rule for a data quality or data integration project.

When you create or edit a column profile, you can create a rule and add it to the profile, or apply an existing rule to the profile. You can use expression rules or predefined rules in a column profile.

After you run the profile, the Analyst tool displays the profile results for the rule column in summary view. You can view the column results for a rule in detailed view. The output of a rule can be one or more virtual columns. The virtual columns exist in the profile results. The Analyst tool runs a profile on the virtual columns. For example, you use a predefined rule that splits a column that contains first and last names into FIRST_NAME and LAST_NAME virtual columns. The Analyst tool runs the profile on the FIRST_NAME and LAST_NAME columns.

Note: If you delete a rule object that other object types reference, the Analyst tool displays a message that lists those object types. Determine the impact of deleting the rule before you delete it.

Predefined Rules

Predefined rules are rules created in the Developer tool or provided with the Developer tool and Analyst tool. Apply predefined rules to the column profiles to modify or validate source data.

Predefined rules use transformations to define rule logic. You can use predefined rules with multiple profiles. In the Model repository, a predefined rule is a mapplet with an input group, an output group, and transformations that define the rule logic.

Predefined Rules Process

Use the **New Rule Wizard** to apply a predefined rule to a profile.

You can perform the following steps to apply a predefined rule:

1. Open a profile.
2. Select a predefined rule.
3. Review the rules parameters.
4. Select the input column. You can select multiple columns if you want to apply the rule to more than one column.
5. Configure the profiling options.

Applying a Predefined Rule

When you apply a predefined rule, you select the rule and configure the input columns and output columns for the rule. Apply a predefined rule to use a rule promoted as a reusable rule or use a rule created by a developer.

1. In the **Library** workspace, select the project that contains the profile, or select the profile in the **Assets** pane.
2. Click **Actions** > **Open** to open the profile.

   The summary view appears in the **Discovery** workspace.
3. Click **Actions** > **Edit Profile**.

   The **Profile Wizard** appears.
4. Click **Specify Rules and Filters**.
5. In the **Specify Rules and Filters** screen, click **Actions** > **Apply an Existing Rule** in the **Rules** panel.

   The **Apply Rule Wizard** dialog box appears.
6. Select a rule, and click **Next**.
7. Click **Add**.

   The **Choose columns for input port** dialog appears.
8. Select a field and an input column. Click **OK**.

   The input columns and output columns appear in the **Apply Rule Wizard** dialog box.
9. In the **Apply Rule Wizard** dialog box, click **OK**.

   The rule appears in the **Specify Rules and Filters** screen.

Expression Rules

Expression rules use expression functions and columns to define rule logic. Create expression rules and add them to a column profile in the Analyst tool.

Use expression rules to change or validate values for columns in a column profile. You can create one or more expression rules to use in a profile. Expression functions are SQL-like functions used to transform source data. You can create expression rule logic with the following types of functions:

- Character
- Conversion
- Data Cleansing
- Date
- Encoding
- Financial
- Numeric
- Scientific
- Special
- Test

You can use the following methods to create an expression rule:

- Profile wizard. When you create or edit a column profile, you can create and apply expression rules in the profile wizard. You can promote the rule to a reusable rule and use it in multiple profiles.
- Rule specification. You can configure a rule specification in the Analyst tool and use the rule specification in the column profile. When you configure a rule specification, you translate the requirements of a business rule into one or more rule statements. The rule statements represent the logic that determines whether a data set conforms to the business rule. Generate a mapplet from the rule specification and use the mapplet in the column profiles that you create in the Developer tool.

You can use the expression editor to add expression functions, configure columns as input to the functions, validate the expression, and configure the return type, precision, and scale. After you create and validate an expression rule, you can edit the precision value of the output rule column. By default, the precision value of the output rule column is set to 10. The precision value is truncated when the output rule column exceeds the set precision value.

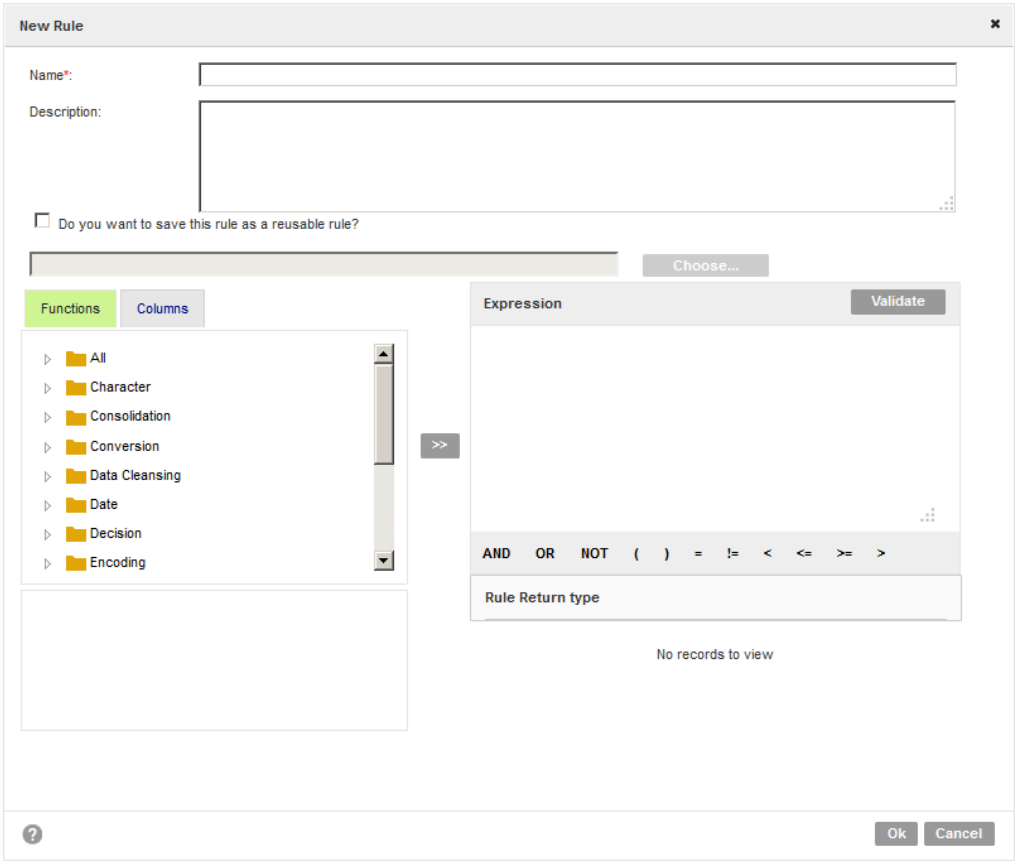
The output of an expression rule is a virtual column that uses the name of the rule as the column name. The Analyst tool runs a column profile on the virtual column. For example, you use an expression rule to validate a ZIP code. The rule returns 1 if the ZIP code is valid and 0 if the ZIP code is not valid. Informatica Analyst runs a column profile on the 1 and 0 output values of the rule.

Creating an Expression Rule

Use the **Profile** wizard to create an expression rule and add it to a profile. Create an expression rule to validate values for columns in a profile.

1. Open a profile.
2. In the summary view, click **Actions** > **Edit Profile** to open the **Profile** wizard.
3. Click **Specify Rules and Filters**.
4. In the Rules pane, click **Actions** > **Add a Rule**.

The **New Rule** dialog box appears.



5. In the **New Rule** dialog box, enter a name and an optional description for the rule. You can create a rule in the Functions panel or Columns panel.
   - In the Functions panel, select a function category, and click the right arrow (>>) button. In the dialog box, specify parameters, and click **OK**.
     The function along with the columns and values appears in the Expression panel.
   - In the Columns panel, select a column, and click the right arrow (>>) button. The column appears in the Expression panel. Add functions, expressions, and values to create a rule.
6. To verify the rule, click **Validate**.
7. Optionally, choose to promote the rule as a reusable rule and configure the project and folder location. If you promote a rule to a reusable rule, you or other users can use the rule in another profile as a predefined rule.
8. Click **OK**.

   The **Specify Rules and Filters** screen appears with the rule in the Rules pane.

Creating an Expression Rule Using Rule Specification

You can use the rule specification to create an expression rule in Informatica Analyst. You can add the rule to column profiles to validate data.

1. In the header area, click **New** > **Rule Specification**.

   The **New Rule Specification** wizard appears.

2. In the **New Rule Specification** wizard, enter a name and an optional description for the rule.
3. In the **Location** field, click **Browse** to select the project or folder where you want to save the rule.
4. Click **Continue**.

   The rule specification appears in the **Design** workspace.
5. To enter the properties for the rule, select the top-level octagonal shape in the rule, and click **Properties**.
6. To configure a primary rule set, click the next-level rectangle shape in the rule.
7. To enter the inputs for the rule set, click **Properties** > **Inputs**.

   The **Inputs Management** dialog box appears.
8. In the **Inputs Management** dialog box, click **Add Input**, and enter a name, data type, maximum length, and a description for the input. Optionally, you can enter multiple inputs.
9. Click **OK**.

   The inputs appear in **Properties** section.
10. To define a rule logic, click **Rule Logic**, and enter an operator, condition, and choose an action in the **Action** list.
11. Optionally, enter multiple rule sets as necessary.
12. To validate the rule, click the **Validate** ( ) icon.
13. To save and use the rule specification in column profiles, click **Save and Finish**.
14. To save and continue working on the rule, click **Save and Continue**.
15. To use the rule specification in the Developer tool, click the **Generate rule** ( ) icon to generate a mapplet.

    The Analyst tool creates a mapplet in the Model repository. Validate the mapplet as rule and then use the mapplet in the column profiles that you create in the Developer tool.

CHAPTER 8

Filters in Informatica Analyst

This chapter includes the following topics:

Filters in Informatica Analyst Overview

You can create a filter so that you can make a subset of the original data source that meets the filter criteria. You can then run a profile on the filtered data.

You can create a filter to view the profile results that meets the filter criteria. You can view the profile results with the default filters that are available in the summary view.

Creating a Filter

You can create a filter so that you can make a subset of the original data source that meets the filter criteria.

1. Open a profile.
2. In the summary view, click **Actions** > **Edit Profile**.

   The **Profile** wizard appears.
3. Click **Specify Rules and Filters**.
4. In the **Filters** pane, click **Actions** > **Add a Filter**.

   The **New Filter** dialog box appears.
5. Create a simple, advanced, or an SQL filter.

   Note: For a simple or advanced filter on a date column, provide the condition in the YYYY/MM/DD HH:MM:SS format.

   The **Data Preview** pane displays the subset of the original data source that meets the filter criteria.
6. Click **OK**.

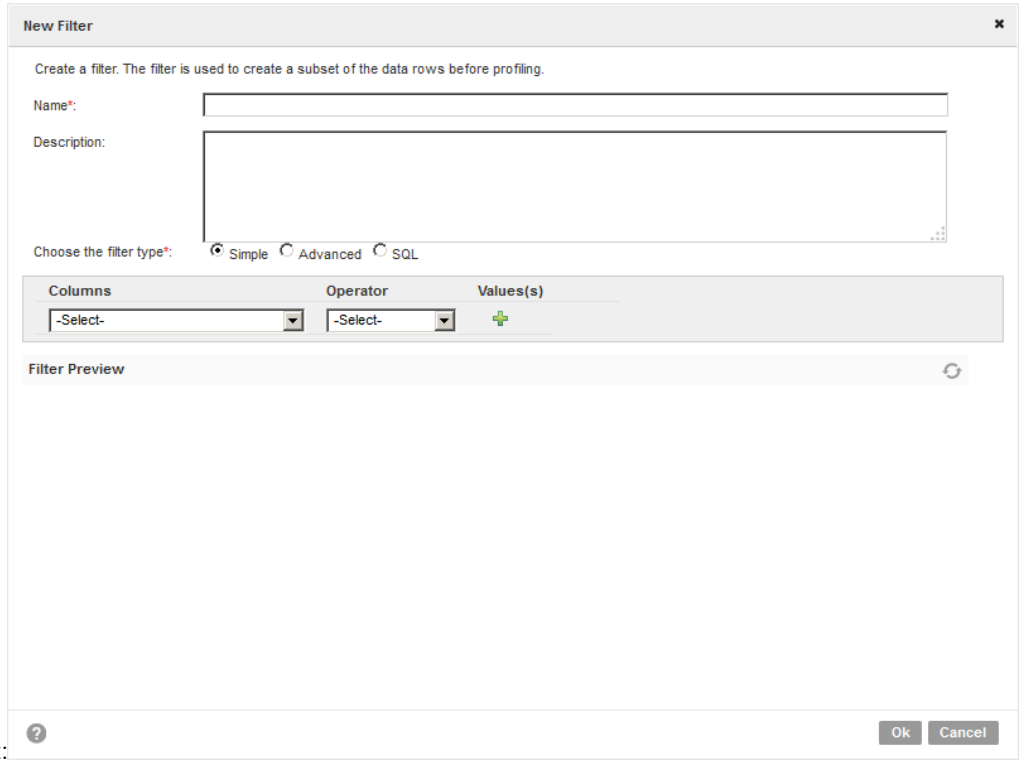
   The **Specify Rules and Filters** screen appears with the filter in the **Filters** pane.

Creating a Simple Filter

You can create a simple filter with conditional operators, such as =, !=, >, <. Use the filter to create a subset of the original data source.

1. In the **New Filter** dialog box, click **Simple**.

   The following image shows the options that you can use to create a simple filter in the **New Filter** dialog box:



2. Enter a name and an optional description.
3. Select a column.
4. Select a conditional operator.
5. Enter a value.
6. Optionally, click the plus (+) icon to add more filters.
7. Click **OK**.

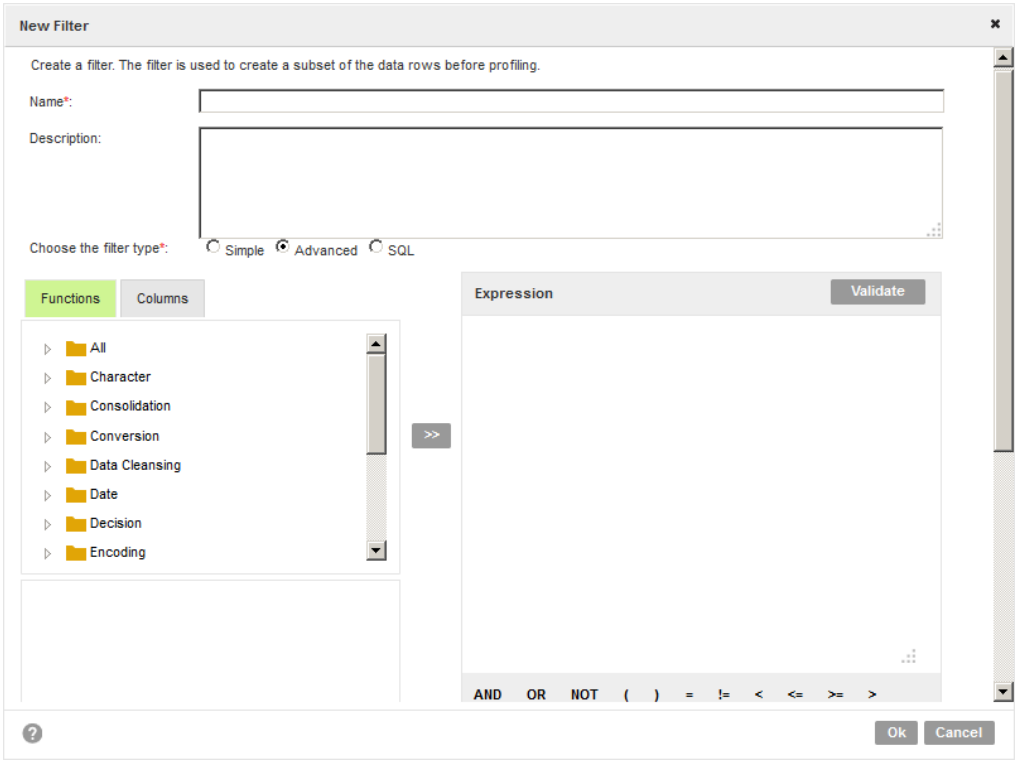
   The **Specify Rules and Filters** page appears with the filter in the Filters pane.

Creating an Advanced Filter

You can create an advanced filter with expressions, such as AND, OR, and NOT to make a subset of the original data source.

1. In the **New Filter** dialog box, click **Advanced**.

   The following image shows the advanced filter options in the **New Filter** dialog box.



2. Enter a name and an optional description for the advanced filter.
3. You can create an advanced filter with the Functions panel or Columns panel.
   - In the Functions panel, select a function category, and click the right arrow (>>) button.
     In the dialog box, specify the parameters and click **OK**. The function along with the columns and values appears in the Expression panel.
   - In the Columns panel, select a column, and click the right arrow (>>) button. The column appears in the Expression panel.
     Add functions, expressions, and values to create an advanced filter.
4. To verify the advanced filter, click **Validate**.
5. Click **OK**.

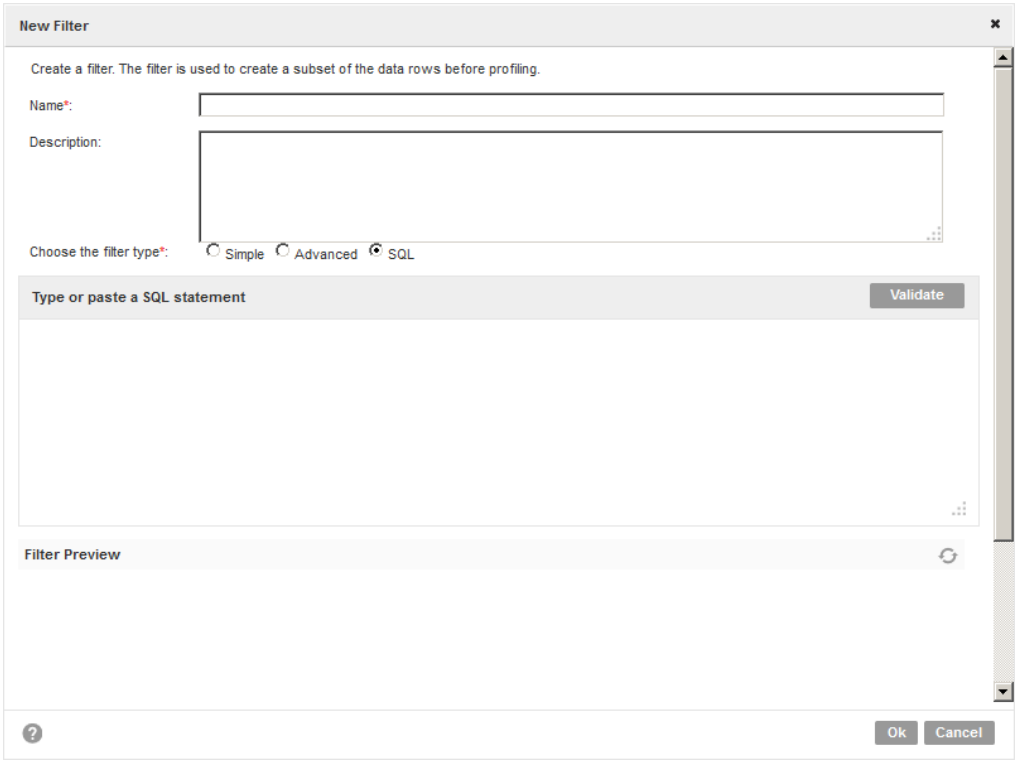
   The **Specify Rules and Filters** screen appears with the filter in the Filters pane.

Creating an SQL Filter

You can create an SQL filter with SQL queries. You can create an SQL filter for relational data sources.

1. In the **New Filter** dialog box, click **SQL**.

   The following image shows the SQL filter options in the **New Filter** dialog box:



2. Enter a name and an optional description for the SQL filter.
3. In the text box, type in or paste an SQL query.
4. Click **Validate** to verify the SQL query.
5. Click **OK**.

   The **Specify Rules and Filters** page appears with the SQL filter in the Filters pane.

Managing Filters

You can edit and delete filters.

1. In the **Library** workspace, select the project that contains the profile, or select the profile in the **Assets** pane you want to filter.
2. Open a profile.
3. In the summary view, click **Actions** > **Edit Profile** to open the **Profile** wizard.
4. Click **Specify Rules and Filters**.
5. In the Filters pane, select a filter, and click **Actions** > **Edit Filter**.

   The **Edit Filter** dialog box appears.

6. Edit the filter settings, and click **OK**.
7. To delete a filter, select a filter, and click **Actions** > **Delete Filter**.

CHAPTER 9

Column Profile Results in Informatica Analyst

This chapter includes the following topics:

Column Profile Results in Informatica Analyst Overview

View profile results to understand and analyze the content, structure, and quality of data. You can view all the columns and rules in a profile in summary view. You can view the properties of a column or rule in detail in the detailed view.

You can view the profile results under the **Discovery** workspace. The view header displays the type of profile, the number of columns in the profile, number of rules in the profile, sampling data, and date and time of creation.

In summary view, you can view the properties of each column as a value, horizontal bar chart, or as a percentage. You can view column properties, such as null, distinct, non-distinct values, patterns, data types, and data domains. You can view the profile results in summary view based on the default filters.

In detailed view, you can view null, distinct, and non-distinct values, inferred data types, inferred data domains, inferred patterns, values, business terms, and preview the data in panes.

You can view profile results for the latest run, historical run, and consolidated run. You can compare profile results for two profile runs and view the results in summary view and detailed view. You can view profile statistics and curate the data. The profile statistics include values, patterns, data types, outliers, and statistics for columns and rules. You can perform data discovery and drill down on data.

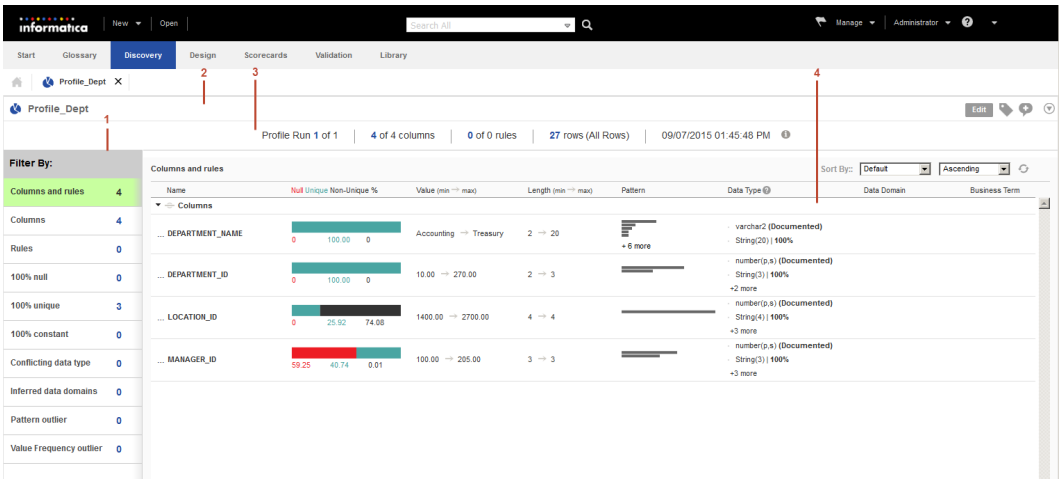
Note: You can view and run a profile on Avro, JSON, Parquet, and XML data sources. You can view profile results for the latest run, historical run, and consolidated run and compare profile results for two profile runs.

You can export value frequencies, pattern frequencies, drill-down data, comments, tags, and business terms to a CSV file. You can export the profile summary information to a Microsoft Excel file so that you can view all data in a file for further analysis. You can view the rule information in the profile results. The profile results that appear depend on the profile configuration and sampling options.

Summary View

The summary of profile results appear in a grid format in the summary view. You can use the default filters in the summary view to view specific statistics. For example, when you choose Rules, the summary view displays all the rules in the profile.

The following image shows a sample graphical view of summary view:



1. Default filters. You can view the profile results in the summary view based on the default filters.
2. Profile header. You can view the profile name in the header. You can use the Edit button to edit the profile, use the tag and comments icons to add or edit tags and comments, and choose the options from the Actions menu.
3. Summary view header. You can view profile-specific information in the summary view header. You can view the profile run number, total number of profile runs, number of columns and rules, and the number of rows in the profile.
4. Summary view. You can view the properties for all the columns and rules in the profile.

In the summary view, you can run or edit the profile, detect pattern or value frequency outliers, add columns to a scorecard, choose a profile run, compare two profile runs, export profile results or data domain discovery results to a Microsoft Excel spreadsheet, verify the inference results of multiple columns, add or delete comments and tags, or view profile properties.

Summary View Properties

The summary view displays the properties for all the columns and rules in a profile. The summary view includes a visual representation of the properties. You can click each summary property to sort the values of the property.

The following table describes the profile results summary properties:

Property	Description
Name	Displays the name of the column or rule in the profile.
Null Distinct Non-Distinct %	Displays the null values, distinct values, and non-distinct values in percentages for a column or rule output. You can view the values in a horizontal bar chart.
Pattern	Displays the multiple patterns in the column as horizontal bar charts. You can view the pattern characters and the number of similar patterns in a column as a percentage when you hover the mouse over the bar chart.
Value	Displays the minimum and maximum values in the column or rule output.
Length	Displays the minimum and maximum length of the values in the column or rule output.
Data Type	Displays the documented data type of the column or rule. Displays the inferred data types when you hover the mouse over the field. The Analyst tool can infer the following data types: - String - Varchar - Decimal - Integer - Date You can also view the percentage of conformance based on the inferred data types. **Note:** The Analyst tool cannot derive the data type from the values of a numeric column that has a precision greater than 38. The Analyst tool cannot derive the data type from the values of a string column that has a precision greater than 255. If you have a date column on which you create a column profile with a year value earlier than 1800, the inferred data type might show up as fixed length string. Change the default value for the year-minimum parameter in the InferDateTimeConfig.xml, as required.
Data Domain	Displays the names of the data domains associated with the column along with the percentage of conformance and the number of conforming rows.
Business Term	Displays the business term assigned to the column.

Default Filters in Summary View

You can view the profile results in summary view based on the default filters.

The summary view displays the profile results for all source columns, virtual columns, and rule columns by default. The Filter By pane displays the number of columns on which you can apply the default filters.

In the summary view, you can view the profile results by using the following default filter options:

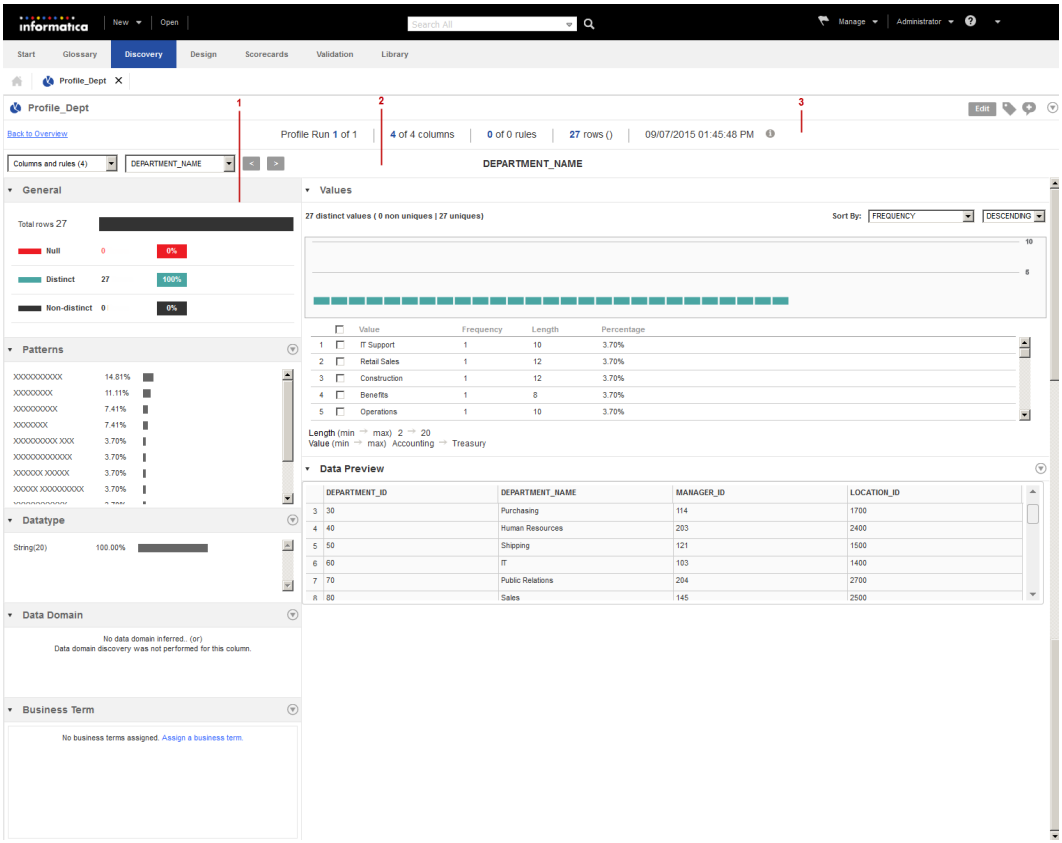
Default Filter Option	Description
Columns and rules	Displays the profile results for the source columns, and rule columns. You can expand and collapse the source columns and rule columns to view the results.
Columns	Displays the profile results for the source columns.
Rules	Displays the profile results for the rule columns.
100% null	Displays the profile results for the columns that have 100% null values.
100% distinct	Displays the profile results for the columns that have 100% distinct values.
100% constant	Displays the profile results for the column that have the same value for all records. For example, 100% constant filter includes the profile results of a Country column if it contains only a "USA" value.
Conflicting data types	Displays the profile results for columns where the documented data type and inferred data type do not match. For example, the filter displays the column CustomerTier because the documented data type for the column is Integer (2) and the inferred data type is string.
Inferred data domains	Displays the profile results for the columns where the inferred data domain is the same as the configured data domain.
Pattern outlier	Displays the profile results for the columns that have pattern outliers.
Value frequency outlier	Displays the profile results for the columns that have value or frequency outliers.

Detailed View

Column results appear in the detailed view. You can view the column properties in detail.

The detailed view for a column appears after you click on the column in summary view.

The following image shows a sample graphical view of column properties in detailed view:



1. Panes. You can view general properties, values in the column, data preview, inferred patterns, inferred data types, inferred data domains, and business terms in panes.
2. Column details header. You can view column results by selecting the column in the dropdown list or by using the navigation buttons.
3. Summary view header. You can view profile specific information in the summary view header. You can view the profile run, number of columns, rules, and rows in the profile run, and the time and date of the profile run.

In the detailed view, you can run or edit the profile, add the column to a scorecard, choose a profile run, compare two profile runs, export the profile results to a Microsoft Excel spreadsheet, export value frequencies, pattern frequencies, data types, drilldown data for selected values, or drilldown data for selected patterns to a csv file, add or delete comments and tags to the column, and view profile properties.

Use the Actions menu in each pane to perform further actions on the column properties. You can collapse or expand the panes.

Detailed View Panes

The detailed view displays the column properties, such as the number and percentage of distinct, non-distinct, and null values, patterns, inferred data types, inferred data domains, values, data preview, and linked business terms in panes.

When you can click the column or rule, the detailed view for the column or rule opens.

The following table describes the panes in detailed view:

Panes	Description
General	Displays the number of rows with null values, distinct values, and non-distinct values in different colors. You can view the values in percentages. You can view the increase and decrease of the general values in every consecutive profile run as a sparkline. A sparkline displays the variation in the number of null values, distinct values, or non-distinct values across the latest five consecutive profile runs in a line chart. You can view the number of values and the percentage of values when you move the pointer over the sparkline for each profile run. You can add tags and comments to the column.
Patterns	Displays the patterns for the column values. The frequency in which the patterns appear in a column appears as a horizontal bar chart and in percentages. You can drill down on a pattern, add a pattern to a reference table, or create a data domain with the selected pattern.
Data type	Displays the inferred data types for the column. The frequency of the data types in a column appears as a horizontal bar chart and in percentages. You can drill down on a data type, approve, reject, or reset the selected inferred data type. The **Show Rejected** option displays rejected inferred data types.
Data Domain	Displays the inferred data domains for the column. You can drill down on a data domain for conforming rows, non-conforming rows, or rows with null values. You can approve, reject, or reset the data domain value. The **Show Rejected** option displays rejected data domains. You can verify the data domain value.
Business Term	Displays the assigned business term for the column. You can assign or unassign a business term to a column.
Values	Displays all the values in the column in a graphical representation along with the frequency, length, and percentage. You can drill down on each value. You can add the value to a reference table, create a value frequency rule, and create a data domain.
Data Preview	Displays the drilldown data for the selected pattern, data type, data domain, or value.

Statistics

You can view statistics, such as values, patterns, data types, data domain, and outliers for the columns and rules in a profile.

You can view profile statistics in summary view, and view column statistics in summary view and detailed view. You can view statistics for the latest profile run, historical profile run, and consolidated profile run. You can compare profile results for two profile runs, and view the statistics for the profile and columns in summary view and detailed view.

Data Preview

You can view the drill-down data for the selected pattern, data type, data domain, or value in the Data Preview pane.

You can view the Data Preview pane in the detailed view. When you click a column in summary view, the detailed view appears and the Data Preview pane is collapsed by default. To view the column data, you can click **Actions** > **Show Preview**.

The following table describes the options in the **Actions** menu in the Data Preview pane:

Option	Description
Add to Filter	Create a drill-down filter to filter the drill-down data so that you can analyze data irregularities on the subsets of profile results.
Save Filter	Saves the drill-down filter.
Show Preview	Displays the source rows.
Export Data	Exports the drilldown results to a CSV file or Microsoft Excel file.

Data Types

The data types include all the inferred data types for each column in the profile results.

You can view the data types in summary view and detailed view. In the summary view, you can view the documented data type and the inferred data types. The **Conflicting data type** filter displays the columns where a conflict between the documented data type and inferred data type exists. In the detailed view, you can view the inferred data types for the column. The frequency of the data types in a column appears as a horizontal bar chart and in percentages. You can drill-down, approve, reject, or reset the selected inferred data type. The Show Rejected option displays rejected inferred data types.

The following table describes the properties for the data types:

Property	Description
Data type	Displays the list of documented and inferred data types for the column in the profile.
Frequency	Displays the number of times a data type appears for a column, expressed as a number.
Percent	Displays the percentage that a data type appears for a column.

Property	Description
Drill down	Drills down to specific source rows based on a column data type. **Note:** You cannot perform a drill-down action if you select multiple inferred data types.
Status	Indicates the status of the data type. The statuses are Inferred, Approved, or Rejected. **Inferred** Indicates the data type of the column that the Analyst tool inferred. **Approved** Indicates an approved data type for the column. When you approve a data type, you commit the data type to the Model repository. **Rejected** Indicates a rejected data type for the column.

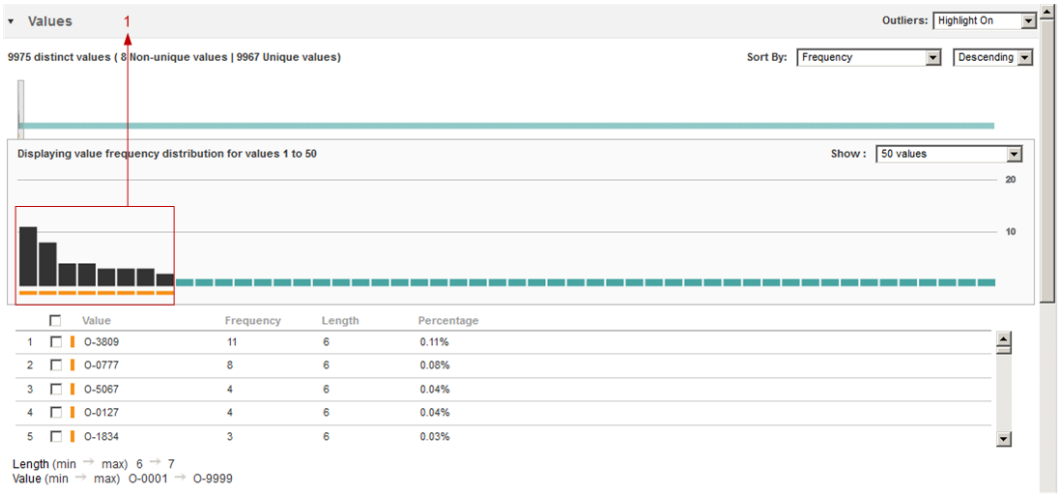
Outliers

An outlier is a pattern, value, or frequency for a column in the profile results that does not fall within an expected range of values.

The profiling plug-in the Data Integration Service runs an algorithm to identify the values that do not fall within the range of the majority of values in the column. Any pattern, value, or frequency that does not fall within the expected range of these majority values in the column is an outlier.

By default, the Analyst tool does not determine outliers in the profile results. In the summary view, you can run the outlier to view the outlier results. The Pattern outlier filter displays the outliers based on the patterns in the column. The Value Frequency outlier filter displays the outliers based on the values or frequencies in the column. The outlier detection occurs in the background so that you can perform other actions in the summary view.

In the detailed view, you can view the outlier values in the Values pane when you select the **Highlight On** option from the list. The outlier value appears as a vertical bar with an orange underline. To view only the outlier value, you must select the **Filter** option from the list.



1. Outlier values. An outlier value appears as a vertical bar with an orange underline.

Running an Outlier

Run an outlier to identify patterns, values, or frequencies in a column that do not fall within an expected range of values.

1. In the summary view, click **Actions** > **Detect Outlier**.

   The Pattern outlier and Value Frequency outlier in the **Filter By** pane changes from N/A to the number of outliers detected.
2. In the **Filter By** pane, click **Pattern outlier**.

   The columns with pattern outliers appear in the summary view.
3. In the **Filter By** pane, click **Value Frequency Outlier**.

   The columns with value or frequency outliers appear in the summary view.
4. In the detailed view, select **Highlight On** from the outlier drop-down list.

   In the Values pane, the outliers appear as vertical bars with orange underlines.
5. Click **Filter** in the Outliers dropdown list to view only outlier values.

Patterns

You can view the patterns for the column values and the frequency in which the patterns appear in summary view and detailed view.

In the summary view, you can view the multiple patterns in the column as horizontal bar charts. You can view the pattern characters and the number of similar patterns in a column as a percentage when you hover the mouse over the bar chart. In the detailed view, you can view the frequency with which the patterns appear in a column as a horizontal bar chart and in percentages. You can drilldown, add the pattern to a reference table, or create a data domain with the selected pattern.

The profiling warehouse stores a maximum of 16,000 unique highest frequency values including NULL values for profile results by default. If there is at least one NULL value in the profile results, the Analyst tool can display NULL values as patterns.

Note: The Analyst tool cannot derive the pattern for a numeric column that has a precision greater than 38. The Analyst tool cannot derive the pattern for a string column that has a precision greater than 255.

The following table describes the properties for the column patterns:

Property	Description
Pattern	Displays the pattern for the column in the profile.
Frequency	Displays the number of times a pattern appears for a column, expressed as a number.
Percentage	Displays the percentage that a pattern appears for a column.

The following table describes the pattern characters and what they represent:

Character	Description
9	Represents any numeric character. Informatica Analyst displays up to three characters separately in the "9" format. The tool displays more than three characters as a value within parentheses. For example, the format "9(8)" represents a numeric value with eight digits.
X	Represents any alphabetic character. Informatica Analyst displays up to three characters separately in the "X" format. The tool displays more than three characters as a value within parentheses. For example, the format "X(6)" might represent the value "Boston." **Note:** The pattern character X is not case sensitive and might represent uppercase characters or lowercase characters from the source data.
p	Represents "(", the opening parenthesis.
q	Represents ")", the closing parenthesis.

Note: Column patterns can also include special characters. For example, ~, [,], =, -, ?, =, {, *, -, >, <, and $.

Values

You can view values for columns and the frequency in which the values appear in the column.

View minimum and maximum values in a column in the summary view. In the detailed view, you can view the value properties for a column.

Values in Summary View

You can view the minimum and maximum values for all the columns and rules for the latest profile run, historical profile run, and consolidated profile run in the summary view.

Example

A retail store database has a column named Employee ID in the Employee table populated with employee IDs ranging from 100 through 250 and has names, such as Bob and Robert as well. When you run a column profile on the Employee table, the Value column for Employee ID in summary view displays 100 --> Robert

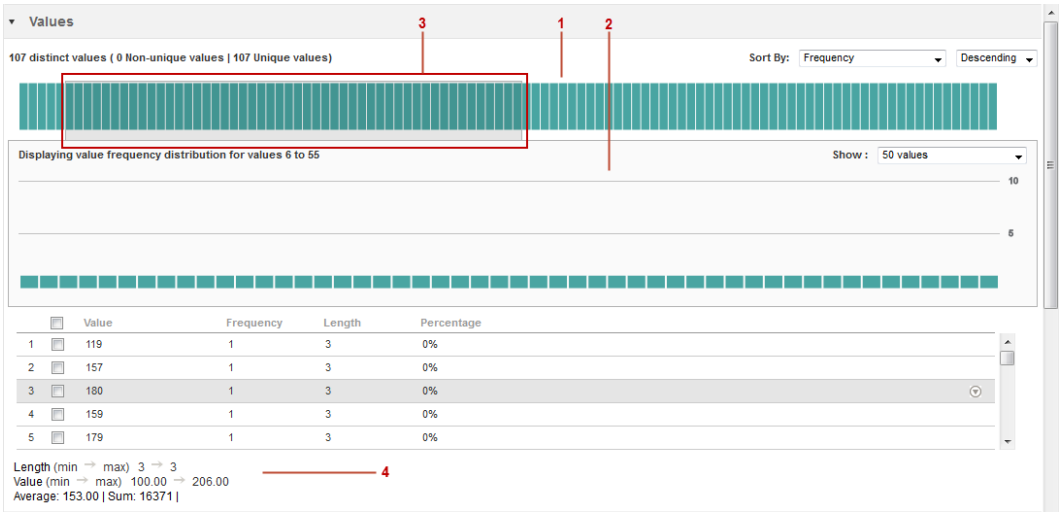
Values in Detailed View

The column values in detailed view include values for a column and the frequency in which the values appear in the column.

The **Values** pane displays the column values in a graphical representation. You can view the frequency, length, and percentage of each value. You can sort the values based on value or frequency. You can drill down on the data, add the values to a reference table, create a value frequency rule, or create a data domain. You can view the null values as a red vertical bar, the frequency of values as a black vertical bar, and the outlier values as vertical bars with orange highlight. You can highlight the outliers, disable outliers, or filter the results to display only outlier values in the column.
The Values pane contains the graphical layout and value sections.

The graphical layout is divided into two panels.

The following image shows the Values pane in the detailed view:



1. Upper panel. You can view the values as a vertical bar chart. You can sort the values by frequency and value. You can sort the value in ascending or descending order. You can view the outlier values as vertical bars with orange highlights.
2. Lower panel. You can view the values in the slider in the lower panel where each value is represented by a vertical bar. You can drill down on the value, add the value to a reference table, create a value frequency rule, and create a data domain on the value. You can view 50, 75, or 100 values at a time.
3. Slider. You can slide the slider over the values in the upper panel. The lower panel displays the values in the slider.
4. Value properties. The value properties section displays the values and properties.

The following table describes the panels in the graphical layout:

Panel	Description
Upper panel	Displays all the values as a vertical bar chart. You can view a maximum of 16,000 values in the upper panel. You can use the slider to view a batch of values.
Lower panel	Displays the values for the batch that you select in the upper panel. By default, the Analyst tool displays 50 values. You can choose to view 75 or 100 values at a time.

The following table describes the properties for the column values in the value section:

Property	Description
Value	Displays a list of values for the batch that you select in the upper panel. **Note:** The Analyst tool excludes the CLOB, BLOB, Raw, and Binary data types in column values.
Frequency	Displays the number of times a value appears in the column, expressed as a number.
Length	Displays the length of the column value.
Percentage	Displays the percentage that a value appears in the column.

The following table describes the statistics for the selected column:

Statistics	Description
Length (min - max)	Displays the length of the shortest value and longest value for the column.
Value (min - max)	Displays the minimum and maximum values in the column.
Average	Displays the average of the values for the column.
Sum	Displays the sum of all the values in the column.

Values in Detailed View for Profile Results Comparison

The Values pane in detailed view for profile results comparison displays value properties, such as number of distinct values, minimum value, maximum value, maximum and minimum length, average, standard deviation, and sum of values.

The detailed view of a column for profile results comparison displays value properties, value, and the frequency of the value with a horizontal bar chart.
The following table describes the properties for the column values in the detailed view when you compare the results of two profile runs.:

Property	Description
No. of distinct values	Displays the number of distinct values in the column.
Min value	Displays the minimum value in the column.
Max value	Displays the maximum value in the column.
Length (Min - Max)	Displays the length of the shortest value and longest value for the column.
Average	Displays the average of the values for the column.
Standard Deviation	Displays the standard deviation or variability between column values for all values of the column.
Sum	Displays the sum of all the values in the column.

Types of Profile Run

You can view the profile results for the latest profile run, historical profile run, and consolidated profile run. You can view the profile run results in the summary view.

Latest Profile Run

View profile results for the latest profile run on the profile in summary view.

You can view the profile results for the latest profile run in summary view when you:

- Create, save, and run a profile.
- Open a profile that you have run previously from the **Library** workspace.
- Click **Back to Latest Profile Run** link in the summary view or detailed view for the consolidated profile run.
- Click **Back to Latest Profile Run** link in the summary view or detailed view for a historical profile run.
- Select the latest profile run in the **Select Profile Run** dialog box, and click **OK**.

Historical Profile Run

View the profile results for a previous profile run in the summary view.

The profiling warehouse saves the profile results of all the profile runs of a profile. You can choose to view the results from a previous version of the profile run by selecting the profile run in the Select Profile Run dialog box.

Consolidated Profile Run

View the latest profile results for each column in the profile in summary view.

In the consolidated profile run, you can view the latest results for each column in the profile. When you choose the Consolidated profile run in the **Select Profile Run** dialog box, the profiling warehouse retrieves the latest column results from all the profile runs of the profile. You can view the results in summary view, and the summary view header displays Incremental profile run.

Example

As a data analyst, you can view the latest results for each column in a profile. For example, you can choose columns 1, 2 and 3 to perform profile run A and choose columns 3, 4 and 5 for profile run B. To view the latest results for all the columns, you can choose Consolidated profile run in the Select Profile Run dialog box. The summary view displays results for columns 1 and 2 from run A and displays results for columns 3, 4, and 5 from run B.

Selecting a Profile Run

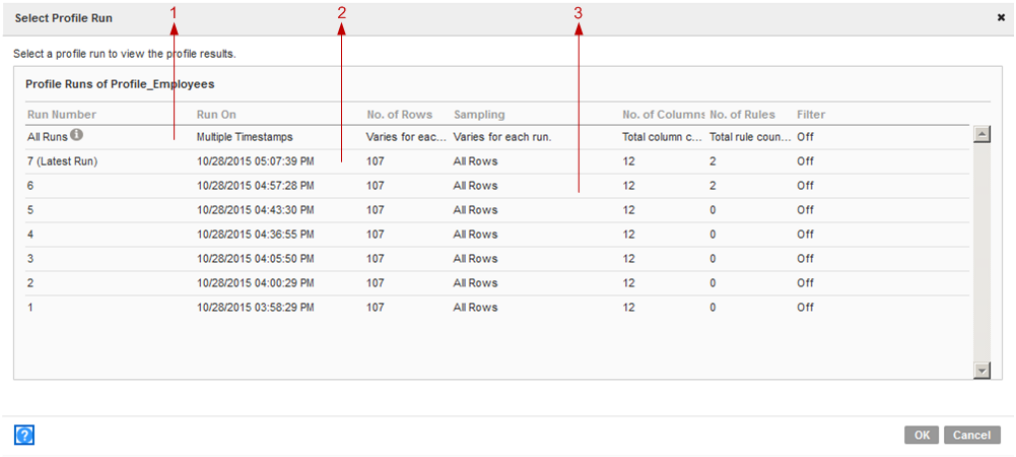
You can select a historical profile run, latest profile run, or consolidated profile run to view the profile results. You can view the profile results in summary view, and view the column results in detailed view.

1. In the **Library** workspace, select the project or folder that contains the profile, or select the profile in the **Assets** pane.
2. Click **Actions** > **Open** to open the profile.

   The summary view appears in the **Discovery** workspace.

3. In the summary view, click **Actions** > **Choose Profile Run**.

   The **Select Profile Run** dialog box appears.
   The following image shows the **Select Profile Run** dialog box.



   1. Consolidated profile run. When you choose this profile run, you can view the latest profile results for each column in summary view.
   2. Latest profile run. When you choose this profile run, you can view the latest profile results for the profile in summary view.
   3. Historical profile run. When you choose this profile run, you can view the historical profile results for a previous profile run in summary view.

4. In the **Select Profile Run** dialog box, select one of the profile runs to view its profile results:
   - To view the profile results for the latest profile run, select the latest profile run, and click **OK**.
   - To view the profile results for a historical profile run, select a profile run other than latest, and click **OK**.
   - To view the profile results for a consolidated profile run, select **All Runs**, and click **OK**. The latest profile results for each column is displayed in the summary view.

   The Analyst tool performs a profile run and displays the profile results in the summary view.
5. In the summary view, click a column to view the column results.

   The detailed view appears.

Compare Multiple Profile Results Overview

You can compare profile results for two profile runs. You can view the compare results in summary view, and column results in detailed view.

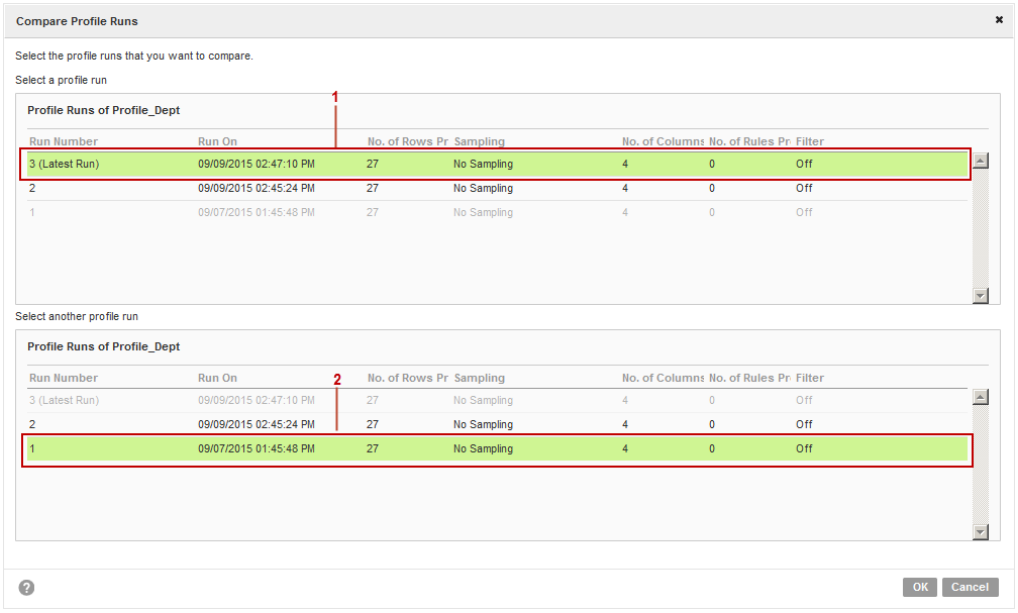
In the summary view, you can view the compare results for all the columns in both the profile runs.

Comparing Multiple Profile Results

When you compare two profile runs, you can view the profile results comparison in summary view.

1. In the summary view, click **Actions** > **Compare Profile Run**.

The following image shows the **Compare Profile Runs** dialog box.



1. Run A. Choose a profile run as Run A.
2. Run B. Choose a profile run as Run B.

The **Compare Profile Runs** dialog box appears.

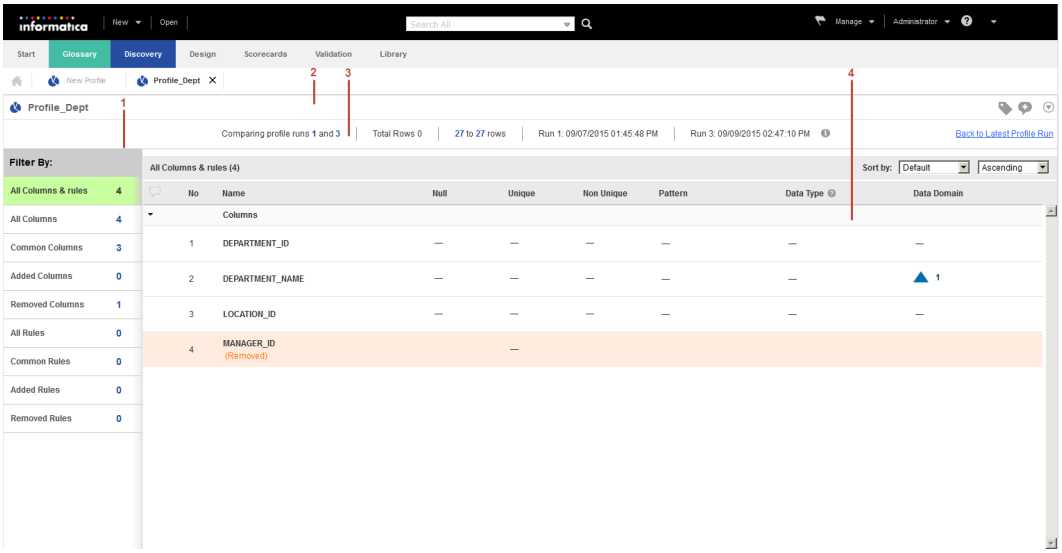
2. Select a profile from the **Run A** pane, and select another profile from the **Run B** pane.
3. Click **OK**.

   The summary view displays a consolidated view of the profile results.

Summary View of Compare Profile Results

When you compare two profile runs, you can view the results in a grid format in the summary view. You can use the default filters in the summary view to view specific statistics.

The following image shows the compare profile results for two profile runs in summary view:



1. Default filter. You can view the profile comparison results in the summary view based on the default filters.
2. Profile header. You can view the profile name in the header.
3. Summary view header. You can view profile specific information in the summary view header. You can view the profile runs that is compared, increase or decrease in rows between the profile runs, number of rows in the profile, and the time and date of the profile runs.
4. Summary view. You can view the comparison between the columns in both the profile runs.

Summary View Properties for Profile Results Comparison

The summary view properties for compare profile results includes the number and percentage of distinct, non-distinct, and null values, patterns, inferred data types, inferred data domains, and linked business terms. The summary view includes a visual representation of the properties. You can click each summary property to sort on values of the property.

In the summary view, the Data Integration Service assigns a number in ascending order to all the columns and rules

Note: An up arrow with a numeric count displays an increase in values of a property from one profile run to another. A down arrow with a numeric count displays a decrease in values of a property from one profile run to another.

The following table describes the summary properties for compare profile results:

Property	Description
No	Displays the number of the column or rule.
Name	Displays the name of the column or rule in the profile.
Null	Displays the increase or decrease in null values.
Distinct	Displays the increase or decrease in distinct values.
Non-distinct	Displays the increase or decrease in non-distinct values.

Property	Description
Pattern	Displays the variation in patterns between the profile runs.
Data type	Displays the variation between the inferred data types for the column or rule in the two profile runs.
Data Domain	Displays the variation between the inferred data domains associated with the column or rule in the two profile runs.

Default Filters for Profile Results Comparison in Summary View

You can view the profile results based on the default filters in the summary view.

In the summary view, you can view source columns and virtual columns. The output for a rule appears as a virtual column in the summary view. When you change the output port for a rule and compare the profile run with a historical run, the historical rule output column appears in the **Removed Rules** filter and the new rule output column appears in the **Added Rules** filter. If you change the rule logic for a single output rule, or if you change the inputs for a multiple rule output in a profile run and compare it with a historical run, the **Added Rules** and **Removed Rules** filter output does not change. The filter output does not change because the filters consider only name changes to the columns as valid inputs to the filter.

You can use the following default filter options to view the profile results that meet specific conditions:

Default Filter Option	Description
All Columns & rules	Displays the profile results for the source columns, virtual columns, and rule columns. You can expand and collapse the source columns and rule columns to view the results.
All Columns	Displays the profile results for the source columns and virtual columns.
Common Columns	Displays the columns available in both the profile run results.
Added Columns	Displays the columns available in the latest profile run. For example, when you compare run 5 with run 3, the Added Columns displays the columns available in run 5 and not run 3.
Removed Columns	Displays the columns available in the historical profile run. For example, when you compare run 5 with run 3, the Removed Columns displays the columns available in run 3 and not run 5.
All Rules	Displays the profile results for all the rule columns.
Added Rules	Displays the rules available in the latest profile run. For example, when you compare run 5 with run 3, the Added Rules displays the rules available in run 5 and not run 3.
Removed Rules	Displays the rules available in the historical profile run. For example, when you compare run 5 with run 3, the Removed Rules displays the rules available in run 3 and not run 5.

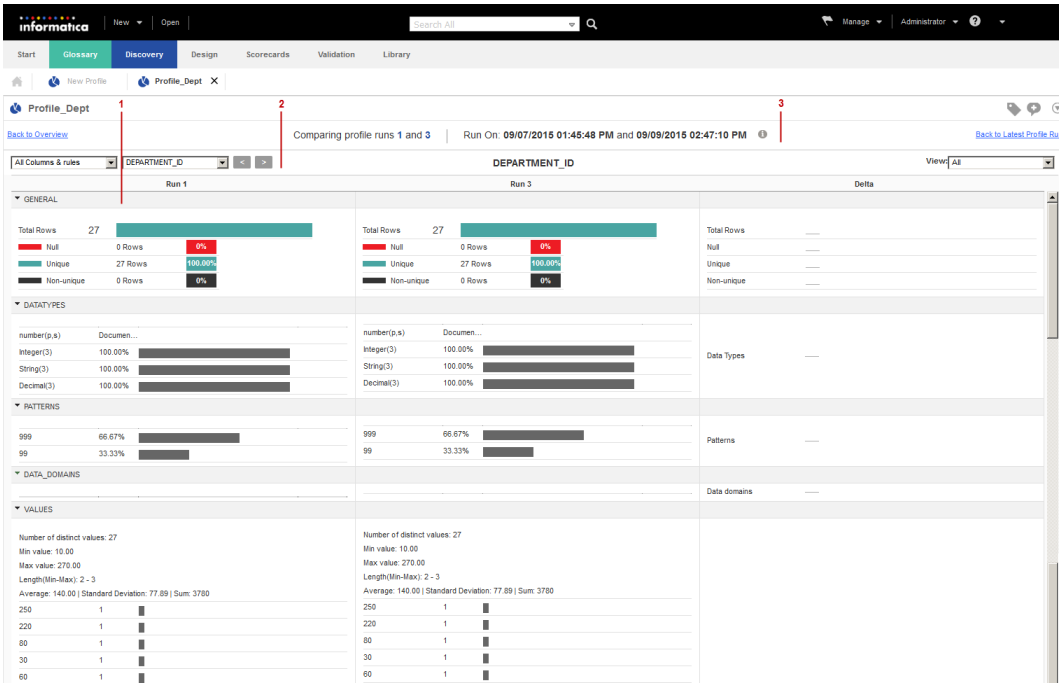
The summary view displays the profile results for all source columns and virtual columns by default.

Detailed View of Compare Profiles Results

Column results appear in a grid format in the detailed view. Column details include general information such as distinct, non-distinct, and null values, patterns, data types, data domains, business terms, values, and data preview.

The detailed view for a column appears when you click the column name. You can view the column results in run A and run B as separate columns, and the comparison of data is available in the delta column.

The following image shows the compare profile results for a column in detailed view:



1. Panes. You can view the profile results and statistics for the column in the two profile runs, and view the delta information for the column in the two profile runs in panes.
2. Profile header. You can view column results by selecting the column in the drop-down list or by using the navigation buttons. You can view the column name, and can view specific results by using the options in the View drop-down list.
3. Summary view header. You can view profile specific information in the summary view header. You can view the profile runs that is compared and the time and date of the profile runs.

Detailed View Panes for Profile Results Comparison

The detailed view displays the profile results and comparison results for a column in the two profile runs in detail.

The detailed view displays the column results for run A and run B, and the comparison of data is available in the delta column. To view other column results, you can select a filter from the filter drop-down list or select the column from the column drop-down list.

Column Profile Drilldown

Use the drill-down options in a column profile to drill down to specific rows in the data source based on a column value. You can choose to read the current data in a data source for drill-down or read profile data staged in the profile warehouse. When you drill-down to a specific row on staged profile data, the Analyst tool creates a drill-down filter for the matching column value. After you drill down, you can edit, recall, reset, and save the drill-down filter.

You can select columns for drill-down even if you did not choose those columns for profiling. You can choose to read the current data in a data source for drill-down or read profile data staged in the profiling warehouse. After you perform a drill-down on a column value, you can export drill-down data for the selected values or patterns to a CSV file at a location you choose. Though Informatica Analyst displays the first 200 values for drill-down data, the tool exports all values to the CSV file.

Drilling Down on Row Data

After you run a profile, you can drill down to specific rows that match the column value, data type, or pattern.

1. Run a profile.

   The profile results appear in the summary view.
2. In the summary view, click a column name.

   The column results appear in detailed view.
3. In the detailed view, right-click on a value in the **Values** pane, and select **Drilldown**.

   The **Data Preview** pane displays the drilldown data.

Applying Filters to Drilldown Data

You can filter the drilldown data iteratively so that you can analyze data irregularities on the subsets of profile results.

1. Select a column value on the **Values** tab.
2. Right-click and select **Drilldown**.

   The drilldown results appear in the **Data Preview** pane.
3. To add a filter condition, right-click a column value in the **Data Preview** pane, and select **Add to Filter**.

   The **Drilldown Filter** dialog box appears with the filter condition.
4. Add the required filter conditions, and click **OK**.

   You cannot apply drill-down filters to inferred data types.
5. To save the filter, click **Actions** > **Save Filter**.
6. To clear the drilldown filters, click **Actions** > **Refresh**.
7. To export drilldown data to a Microsoft Excel spreadsheet, click **Actions** > **Export Data**.

Curation in the Analyst tool

Curation is the process of validating and managing discovered metadata of a data source so that the metadata is fit for use and reporting. When you curate metadata in the Analyst tool, you can approve, reject, and reset the inferred data types or data domains in profile results.

You can approve one data type and one data domain for a column. You can hide the rejected data types or data domains for a column. After you approve or reject an inferred data type or data domain, you can reset the data type or data domain to restore the inferred status.

Approving Data types and Data Domains

The profile results include the inferred data types and data domains for each column in the data source. You can choose and approve a single data type and a single data domain for each column in the Analyst tool.

1. Run a profile.

   The profile results appear in the summary view.
2. In the summary view, click a column name.

   The column results appear in detailed view.
3. In the detailed view, select a data type in the **Data type** pane or a data domain in the **Data Domain** pane.
4. Click **Actions** > **Approve**.
5. To restore the inferred status of the data type or data domain, select the data type or data domain, and click **Actions** > **Reset**.

Rejecting Data types and Data Domains

In the detailed view, you can reject a data type or data domain. You can show or hide the rejected data types and data domains.

1. Run a profile.

   The profile results appear in the summary view.
2. In the summary view, click a column name.

   The column results appear in detailed view.
3. In the detailed view, select a data type in the **Data type** pane or a data domain in the **Data Domain** pane.
4. Click **Actions** > **Reject**.

   The Analyst tool removes the rejected data type from the list of data types.
5. To view the rejected data types, click **Actions** > **Show Rejected**.

Column Profile Export Files in Informatica Analyst

You can export column profile results to a CSV file or a Microsoft Excel file based on whether you choose a part of the profile results or the complete results summary.

You can export value frequencies, pattern frequencies, data types, or drilldown data to a CSV file for selected values and patterns. You can export the profiling results summary for all columns to a Microsoft Excel file.

Use the Data Integration Service privilege **Drilldown and Export Results** to determine, by user or group, who exports profile results.

Profile Export Results in a CSV File

You can export value frequencies, pattern frequencies, data types, or drilldown data to view the data in a file. The Analyst tool saves the information in a CSV file.

When you export inferred column patterns, the Analyst tool exports a different format of the column pattern. For example, when you export the inferred column pattern X(5), the Analyst tool displays the following format of the column pattern in the CSV file: XXXXX.

Profile Export Results in Microsoft Excel

When you export the complete profile results summary, the Analyst tool saves the information to multiple worksheets in a Microsoft Excel file. The Analyst tool saves the file in the "xlsx" format.

The following table describes the information that appears on each worksheet in the export file:

Tab	Description
Column Profile	Summary information exported from the summary view after the profile runs. Examples are column names, rule names, number of distinct values, number of null values, inferred data types, and date and time of the last profile run.
Values	Values for the columns and rules and the frequency in which the values appear for each column.
Patterns	Value patterns for the columns and rules you ran the profile on and the frequency in which the patterns appear.
Data Types	All the data types for the column, frequency of each data type, percentage value, and status of the data type, such as Inferred, Approved, or Rejected.
Statistics	Statistics about each column and rule. Examples are average, length, top values, bottom values, and standard deviation.
Properties	Properties view information, including profile name, type, sampling policy, and row count.

Exporting Profile Results from Informatica Analyst

You can export the results of a profile to a ".csv" or ".xlsx" file to view the data in a file.

1. In the **Library** workspace, select the project or folder that contains the profile.
2. Click the profile to open it.

   The profile results appears in summary view.
3. In the summary view, click **Actions** > **Export Data**.

   The **Export data to a file** dialog box appears.
4. In the **Export data to a file** dialog box, enter a file name. Optionally, use the default file name.
5. Select **All (Summary, Values, Patterns, Statistics, Properties)** or **Data doman discovery results**, and select a **Code Page**. Click **OK**.

   The data is exported to a Microsoft Excel spreadsheet.

6. Click a column in the summary view.

   The column results appear in detailed view.
7. In the detailed view, click **Actions** > **Export Data**.

   The **Export data to a file** dialog box appears.
8. In the **Export data to a file** dialog box, enter a file name. Optionally, use the default file name.
9. Select one of the following options:
   - All (Summary, Values, Patterns, Statistics, Properties)
   - Value frequencies for the selected column.
   - Pattern frequencies for the selected column.
   - Data types for the selected column.
   - Drilldown data for the selected values.
   - Drilldown data for the selected patterns.
   - Drilldown data for the selected data types.
10. Enter a file format. The format is **Excel** for the **All** option and **CSV** for the rest of the options. You can choose to export the field name as a first row in the file.
11. Select the code page of the file.
12. Click **OK**.

    The data is exported to the file.

CHAPTER 10

Business Terms, Comments, and Tags in Informatica Analyst

This chapter includes the following topics:

Business Terms, Comments, and Tags in Informatica Analyst Overview

You can add business terms, comments, and tags to a profile or to columns in the profile. You can assign, view, and edit the business terms, comments, and tags in summary view and detailed view.

Business Terms

You can assign business terms to columns in a profile in the Analyst tool. You can edit an asset link or delete a business term for a column. A business glossary is a set of terms that use business language to define concepts for business users. A business term provides the business definition and usage of a concept.

You can assign, view, or delete business terms in a column in the summary view and detailed view. To view the business term in the **Glossary** workspace, click the business term in the detailed view.

You can edit the properties of an asset link for a business term. You can export business terms as a virtual column to a CSV file along with the profile results.

Assigning Business Terms to Columns

Assign business terms to columns in a profile in summary view and detailed view. You can delete a business term for a column in the **Business Terms** panel. You can edit the properties of an asset link for a business term in the **Edit Asset Link** dialog box.

1. In summary view, right-click the column name and select **Manage Business Terms**. In detailed view, select **Manage Business Terms** from the **Actions** menu in the **Business Term** tab.

   The **Business Terms** panel appears.
2. Click plus (+) icon, or click the **Assign business term** link to add a business term.

   The **Assign Business Term** panel appears.
3. Select a business term from the list of business terms in the **Assign Business Term** panel. Enter an asset name. Optionally, you can add the context and description for the asset. Click **OK**.

   The business term appears in the **Business Terms** panel.

Comments

You can add comments to a profile and the columns in the profile so that you can provide additional information for further collaboration and analysis.

At the profile level, you can add comments about the profile, profile definition, or profile metadata. You can view the comments for a profile in the summary view.

You can add and view column comments in summary view and detailed view.

You can perform the following tasks with comments:

- Export comments as a virtual column to a CSV file along with the profile results. The CSV file contains all the comments for a profile and columns in the profile.
- Search profile results using the keywords in the comments column.
- Add comments to both source columns and virtual columns in a profile.

Note: When you do not select any column or add any column comment, the comments panel in summary view displays profile comments.

Adding Comments to a Profile or Columns

You can add or view a comment in the **Comments** panel.

1. You can add comments in summary view or detailed view.
   - In summary view, to add profile comments, click **Actions** > **Show Comments**.
   - In summary view, to add column comments, right-click on a column, and select **Show Comments**.
   - In detailed view, click **Add Comment** in the **General** pane.

   The **Comments** panel appears.
2. Click **Add Comment**.

   A text box appears in the **Comments** panel.

3. Add a descriptive comment text, and click **Save**.

   The comment appears in the **Comments** panel with the current user name and the date and time of creation.

Tags

You can assign tags to a profile or columns in the profile to group objects according to their business usage.

You can view or assign profile tags in the summary view. You can view or assign column tags in summary view and detailed view.

You can perform the following tasks with column tags:

- Export tags as a virtual column to a CSV file along with the profile results. The CSV file contains all the tags for a profile and the columns in the profile.
- Assign tags to both source columns and virtual columns in the profile.

Note: When you do not select any column or when you do not add any column tag, the tags panel displays profile tags in summary view.

Assigning Tags to a Profile or Columns

Add tags to a profile in summary view. Add tags to a column in summary view and detailed view.

1. You can add tags in summary view or detailed view.
   - In summary view, to assign tags to a profile, click **Actions** > **Show Tags**.
   - In summary view, to add tags to a column, right-click on a column, and click **Show Tags**.
   - In detailed view, click **Add Tag** in the **General** pane.

   The **Tags** panel appears.
2. Click the plus (+) icon, or click the **Assign Tags** link to assign a tag.

   The **Assign Tags** dialog box appears.
3. Select one or more tags to assign to a profile or column. Click **OK** to open the**Tags** panel.

   Note: To create a tag, click **Add New Tag** in the **Assign Tags** panel.

CHAPTER 11

Scorecards in Informatica Analyst

This chapter includes the following topics:

Scorecards in Informatica Analyst Overview

A scorecard is the graphical representation of valid values for a column in a profile. You can create scorecards and drill down on live data or staged data.

Use scorecards to measure data quality progress. For example, you can create a scorecard to measure data quality before you apply data quality rules. After you apply data quality rules, you can create another scorecard to compare the effect of the rules on data quality.

Scorecards display the value frequency for columns as scores. The scores reflect the percentage of valid values in the columns. After you run a profile, you can add columns from the profile as metrics to a scorecard. You can create metric groups so that you can group related metrics to a single entity. You can define thresholds that specify the range of bad data acceptable for columns in a record and assign metric weights for each metric. When you run a scorecard, the Analyst tool generates weighted average values for each metric group. To further assess data quality, you can also assign a fixed or variable cost to each metric.

When you run the scorecard, the Analyst tool computes the sum of cost of bad data for each metric and displays the total cost.

When you create or edit a scorecard, you can create scorecard filters based on the source data. The scorecard filters enable you to recalculate metric scores based on the filter condition. To identify valid data records and records that are not valid, you can drill down on each metric. You can use trend charts to track how metric scores and cost of bad data in metrics change over a period of time. You can reuse the profile filters in a scorecard.

When version control system is enabled in the Analyst tool, you can create multiple versions of a scorecard and view version history for a scorecard. By default, the scorecard is checked out after you create a scorecard. You must check in the scorecard so that the other users can edit the scorecard.

You can view the scorecard dashboard in the **Scorecards** workspace. In the scorecard dashboard, you can view the data objects that have scorecards, scorecards in a project, scorecard run trend in the past six months, and the aggregate of good, acceptable, and unacceptable metrics for all the scorecard runs in a month.

You can configure and manage email notifications for scorecards in Informatica Analyst. Use the Email Service to manage the email notifications. The Email Service is a system service that you can configure in Informatica Administrator.

Informatica Analyst Scorecard Process

You can create and edit a scorecard in the Developer tool and Analyst tool. You can run a scorecard in the Analyst tool. You can run the scorecard on current data in the data object or on data staged in the profiling warehouse.

You can view a scorecard in the **Scorecards** workspace. After you run the scorecard, you can view the scores on the **Scorecard** panel. You can select the data object and navigate to the data object from a score within a scorecard. The Analyst tool opens the data object in another tab.

You can perform the following tasks when you work with scorecards:

1. Create a scorecard in the Developer tool or Analyst tool, and add columns from a profile.
2. Open the scorecard in the Analyst tool.
3. After you run a profile, add profile columns as metrics to the scorecard.
4. Optionally, create scorecard filters based on the source data.
5. Optionally, configure the cost of invalid data for each metric.
6. Run the scorecard to generate the scores for columns.
7. View the scorecard to see the scores for each column in a record.
8. Drill down on the columns for a score.
9. Edit a scorecard.
10. Set thresholds for each metric in a scorecard.
11. Create a group to add or move related metrics in the scorecard.
12. Edit or delete a group, as required.
13. View the score trend chart for each score to monitor how the score changes over time.
14. Optionally, view the cost trend chart for each metric to monitor the value of data quality.
15. View scorecard lineage for each metric or metric group.

16. View consolidated information about the scorecards for which you have read access.

Creating a Scorecard in Informatica Analyst

Create a scorecard and add columns from a profile to the scorecard. You must run a profile before you add columns to the scorecard.

1. In the **Library** workspace, select the project or folder that contains the profile.
2. Click the profile to open the profile.

   The profile results appear in the summary view in the **Discovery** workspace.
3. Click **Actions** > **Add to scorecard**.

   The **Add to Scorecard** wizard appears.
4. In the **Add to Scorecard** screen, you can choose to create a new scorecard, or edit an existing scorecard to add the columns to a predefined scorecard. The **New Scorecard** option is selected by default. Click **Next**.
5. In the **Step 2 of 8** screen, enter a name for the scorecard. Optionally, you can enter a description for the scorecard. Select the project and folder where you want to save the scorecard. Click **Next**.

   By default, the scorecard wizard selects the columns and rules defined in the profile. You cannot add columns that are not included in the profile.
6. In the **Step 3 of 8** screen, select the columns and rules that you want to add to the scorecard as metrics. Optionally, click the check box in the left column header to select all columns. Optionally, select **Column Name** to sort column names. Click **Next**.
7. In the **Step 4 of 8** screen, you can add a filter to the metric.

   You can apply the filter that you created for the profile to the metrics, or create a new filter. Select a metric in the **Metric Filters** pane, and click the **Manage Filters** icon to open the **Edit Filter: column name** dialog box. In the **Edit Filter: column name** dialog box, you can choose to perform one of the following tasks:

   - Choose a filter that you created for the profile. Click **Next**.
   - Select an existing filter. Click the edit icon to edit the filter in the **Edit Filter** dialog box. Click **Next**.
   - Click the plus (+) icon to create filters in the **New Filter** dialog box. Click **Next**.

   Optionally, you can choose to apply the selected filters to all the metrics in the scorecard.

   The filter appears in the **Metric Filters** pane.
8. In the **Step 4 of 8** screen, click **Next**.
9. In the **Step 5 of 8** screen, select each metric in the **Metrics** pane to perform the following tasks:
   - Configure valid values. In the **Score using: Values** pane, select one or more values in the **Available Values** pane, and click the right arrow button to move them to the **Valid Values** pane. The total number of valid values for a metric appears at the top of the **Available Values** pane.
   - Configure metric thresholds. In the **Metric Thresholds** pane, set the thresholds for **Good**, **Acceptable**, and **Unacceptable** scores.
   - Configure the cost of invalid data. To assign a constant value to the cost for the metric, select **Fixed Cost**. To attach a numeric column as a variable cost to the metric, select **Variable Cost**, and click **Select Column** to select a numeric column. Optionally, click **Change Cost Unit** to change the unit of cost. If you do not want to configure the cost of invalid data for the metric, choose **None**.

10. Click **Next**.
11. In the **Step 6 of 8** screen, you can select a metric group to which you can add the metrics, or create a new metric group, To create a new metric group, click the group icon. Click **Next**.
12. In the **Step 7 of 8** screen, specify the weights for the metrics in the group and thresholds for the group.
13. In the **Step 8 of 8** screen, select **Native** or **Hadoop** run-time environment option to run the scorecard. If you choose the Hadoop option, click **Browse** to choose a Hadoop connection to run the profile on the Blaze engine.
14. Click **Save** to save the scorecard, or click **Save & Run** to save and run the scorecard.

    The scorecard appears in the **Scorecard** workspace.

Add Columns to Existing Scorecards

After you run a profile, you can add columns in the profile results to an existing scorecard. You can add metrics or metric groups, configure valid values for the columns, and add the cost of invalid data for each metric. If you add a column to a scorecard from a profile with a sampling option other than **All Rows**, the scorecard might not reflect the profile results.

When you can add columns to an existing scorecard. you cannot edit the existing metrics or metric groups of the scorecard in the **Add to Scorecard** wizard. To modify the existing metrics in the scorecard, navigate to the Scorecard workspace, edit the scorecard, and update the metrics or metric groups as required.

Adding Columns to an Existing Scorecard

After you run a profile, you can add columns to an existing scorecard.

1. Click a profile to open it.

   The profile results appear in the summary view.
2. Select a column. Click **Actions** > **Add to scorecard**.

   The **Add to Scorecard** wizard appears.

   Note: Use the following rules and guidelines before you add columns to a scorecard:

   - You cannot add a column to a scorecard if both the column name and scorecard name match.
   - You cannot add a column twice to a scorecard even if you change the column name.
3. Select **Existing Scorecard** to add the columns to a predefined scorecard. Click **Next**.
4. In the **Step 2 of 7** screen, select the scorecard that you want to add the columns to. Click **Next**.

   You can view the existing metrics and metric groups associated with the scorecard.
5. In the **Step 3 of 7** screen, select the columns and rules that you want to add to the scorecard as metrics. Optionally, click the check box in the left column header to select all columns. Click **Column Name** to sort column names. Click **Next**.
6. In the **Step 4 of 7** screen, you can create filters for the metrics. You can also apply the filter that you created for the profile to the metrics.
7. In the **Step 5 of 7** screen, you can perform the following tasks:

- In the **Metrics** pane, select each metric and configure metric values in the other panes.
- In the **Score using: Values** pane, select multiple values in the **Available Values** pane, click the right arrow button to move the values to the **Valid Values** pane.
  The total number of valid values for a metric appears at the top of the **Available Values** pane.
- In the **Metric Thresholds:** pane, you can set thresholds for **Good**, **Acceptable**, and **Unacceptable** scores.
- In the **Cost of invalid data**, you can:
  - Select each metric and configure the cost of invalid data for the metric.
  - Select **Fixed Cost** option to assign a constant value to the cost for the metric. You can click **Change Cost Unit** to change the unit of cost.
  - Select **Variable Cost** option to attach a numeric column as a variable cost to the metric. You can click **Select Column** to select a numeric column.

8. Click **Next**.
9. In the **Step 6 of 7** screen, you can perform the following tasks:
   - Select the metric group to which you want to add the metrics.
   - In the **Default - Metrics** pane, you can double-click the default metric weight of 0 to change the value.
   - In the **Metric Thresholds:** pane, you can set thresholds for **Good**, **Acceptable**, and **Unacceptable** scores.
10. Click **Next**.
11. In the **Step 7 of 7** screen, select a run-time environment.
12. Click **Save** to save the scorecard, or click **Save & Run** to save and run the scorecard.

Running a Scorecard

Run a scorecard to generate scores for columns.

1. In the **Assets** panel, choose the scorecard that you want to run.
2. Click the scorecard to open it.
   The scorecard appears in the **Scorecards** workspace.
3. Click **Actions** > **Run Scorecard.**
4. Select a score from the **Metrics** pane and select the columns from the **Columns** pane to drill down on.
5. In the **Drilldown** option, choose to drill down on live data or staged data.
   For optimal performance, drill down on live data.
6. Click **Run**.

Viewing a Scorecard

Run a scorecard to see the scores for each metric. A scorecard displays the score as a percentage and bar. View data that is valid or not valid. You can also view scorecard information, such as the metric weight, metric group score, score trend, and name of the data object.

1. Run a scorecard to view the scores.
2. Select a metric that contains the score you want to view.
3. Click **Actions** > **Drilldown** to view the rows of valid data or rows of data that is not valid for the column.

   The Analyst tool displays the rows of data that is not valid by default in the **Drilldown** section.

Editing a Scorecard

Edit valid values for metrics in a scorecard. You must run a scorecard before you can edit it.

1. In the **Library** workspace, click the scorecard you want to edit in the **Assets** pane.

   The scorecard appears in the **Scorecards** workspace.
2. If the version control system is enabled, click **Actions** > **Check Out**.
3. Click **Actions** > **Edit** > **General**.

   The **Edit Scorecard** dialog box appears.
4. In the **General**tab, you can edit the name and description of the scorecard as required.
5. Click the **Metrics** tab.
6. Select a score in the **Metrics** pane and configure the valid values from the list of all values in the **Score using: Values** pane.
7. In the **Metric Thresholds** pane, you can make changes to the score thresholds as required.
8. Review the cost of invalid data for each metric and make changes as required.
9. Click the **Scorecard Filters** tab.
10. You can add, edit, or delete filters.
11. Click the **Metric Groups** tab.
12. You can create, edit, or remove the metric groups.

    You can also edit the metric weights and metric thresholds in the **Metric Groups** tab.
13. Click the **Notifications** tab.
14. You can make changes to the scorecard notification settings as required.

    You can set up global and custom settings for metrics and metric groups.
15. Click the **Run-time Environment** tab.

    You can select **Native** or **Hadoop** as the run-time environment.
16. Click **Save** to save changes to the scorecard, or click **Save & Run** to save the changes and run the scorecard.
17. Click **Check In**.

Metrics

A metric is a column of a data source or output of a rule that is part of a scorecard. When you create a scorecard, you can assign a weight to each metric. Create a metric group to categorize related metrics in a scorecard into a set.

Metric Weights

When you create a scorecard, you can assign a weight to each metric. The default value for a weight is 0.

When you run a scorecard, the Analyst tool calculates the weighted average for each metric group based on the metric score and weight you assign to each metric.

For example, you assign a weight of W1 to metric M1, and you assign a weight of W2 to metric M2. The Analyst tool uses the following formula to calculate the weighted average:

```
(M1 X W1 + M2 X W2) / (W1 + W2)
```

Value of Data Quality

A measure of data quality in the source data is critical information in the management of the data assets in the organization. The cost of invalid data in metrics represented in a scorecard helps organizations derive value in monitoring data quality of the source data. As a data analyst, you might want to associate a value, such as a currency unit or any custom unit, to metrics and metric groups. You can then run the scorecard to view the total cost of invalid data in the source data.

You can define the cost unit for a metric based on the business needs. You can also configure a variable or fixed cost for each metric when you create a scorecard or edit it.

Fixed Cost

Fixed cost is a constant value that you can assign to a metric in a scorecard. You can choose a predefined cost unit or create a custom cost unit that meets the business needs.

Variable Cost

Variable cost is a value that you assign to a metric based on the values in a numeric column of a data source. The Data Integration Service calculates the variable cost for the metric based on the column or virtual column that you assign to the cost.

Example

As a mortgage loan officer, you need to provide your customers with payment books so that the customers can submit the mortgage payments. You can use a scorecard to measure the accuracy of your customer addresses to ensure the delivery of the payment books. You might want to set the variable cost to the Monthly Payment Amount column for the Address Accuracy metric. Run the scorecard to compute the total cost that the mortgage organization loses if customers did not pay the monthly amount on time.

Defining Thresholds

You can set thresholds for each score in a scorecard. A threshold specifies the range in percentage of bad data that is acceptable for columns in a record. You can set thresholds for good, acceptable, or unacceptable

ranges of data. You can define thresholds for each column when you add columns to a scorecard, or when you edit a scorecard.

Complete one of the following prerequisite tasks before you define thresholds for columns in a scorecard:

- Open a profile and add columns from the profile to the scorecard in the **Add to Scorecard** dialog box.
- Optionally, click a scorecard in the **Library** workspace and select **Actions** > **Edit** to edit the scorecard in the **Edit Scorecard** dialog box.

1. In the **Add to Scorecard** dialog box or the **Edit Scorecard** dialog box, select each metric in the **Metrics** pane.
2. In the **Metric Thresholds** pane, enter the thresholds that represent the upper bound of the unacceptable range and the lower bound of the good range.

   You can set thresholds for up to two decimal places.
3. Click **Next** or **Save**.

Metric Groups

Create a metric group to categorize related scores in a scorecard into a set. By default, the Analyst tool categorizes all the scores in a default metric group.

After you create a metric group, you can move scores out of the default metric group to another metric group. You can edit a metric group to change its name and description, including the default metric group. You can delete metric groups that you no longer use. You cannot delete the default metric group.

Creating a Metric Group

Create a metric group to add related scores in the scorecard to the group.

1. In the **Library** workspace, click the scorecard you want to edit in the **Assets** pane.

   The scorecard appears in the **Scorecards** workspace.
2. Click **Actions** > **Edit**.

   The **Edit Scorecard** window appears.
3. Click the **Metric Groups** tab.

   The default group appears in the **Metric Groups** panel and the scores in the default group appear in the **Metrics** panel.
4. Click the **New Group** icon to create a metric group.

   The **Metric Groups** dialog box appears.
5. Enter a name and optional description.
6. Click **OK**.
7. Click **Save** to save the changes to the scorecard.

Moving Scores to a Metric Group

After you create a metric group, you can move related scores to the metric group.

1. In the **Library** workspace, click the scorecard you want to edit in the **Assets** pane.

   The scorecard appears in the **Scorecards** workspace.
2. Click **Actions** > **Edit**.

   The **Edit Scorecard** window appears.
3. Click the **Metric Groups** tab.

   The default group appears in the **Metric Groups** panel and the scores in the default group appear in the **Metrics** panel.
4. Select a metric from the **Metrics** panel and click the **Move Metrics** icon.

   The **Move Metrics** dialog box appears.

   Note: To select multiple scores, hold the Shift key.
5. Select the metric group to move the scores to.
6. Click **OK**.

Editing a Metric Group

Edit a metric group to change the name and description. You can change the name of the default metric group.

1. In the **Library** workspace, click the scorecard you want to edit in the **Assets** pane.

   The scorecard appears in the **Scorecards** workspace.
2. Click **Actions** > **Edit**.

   The **Edit Scorecard** window appears.
3. Click the **Metric Groups** tab.

   The default metric group appears in the **Metric Groups** panel and the metrics in the default metric group appear in the **Metrics** panel.
4. On the **Metric Groups** panel, click the **Edit Group** icon.

   The **Edit** dialog box appears.
5. Enter a name and an optional description.
6. Click **OK**.

Deleting a Metric Group

You can delete a metric group that is no longer valid. When you delete a metric group, you can choose to move the scores in the metric group to the default metric group. You cannot delete the default metric group.

1. In the **Library** workspace, click the scorecard you want to edit in the **Assets** pane.

   The scorecard appears in the **Scorecards** workspace.
2. Click **Actions** > **Edit**.

   The **Edit Scorecard** window appears.
3. Click the **Metric Groups** tab.

The default metric group appears in the **Metric Groups** panel and the metrics in the default metric group appear in the **Metrics** panel.

4. Select a metric group in the **Metric Groups** panel, and click the **Delete Group** icon.

   The **Delete Groups** dialog box appears.

5. Choose the option to delete the metrics in the metric group or the option to move the metrics to the default metric group before deleting the metric group.
6. Click **OK**.

Drilling Down on Columns

Drill down on the columns for a score to select columns that appear when you view the valid data rows or data rows that are not valid. The columns you select to drill down on appear in the **Drilldown** panel.

1. Run a scorecard to view the scores.
2. Select a column that contains the score you want to view.
3. Click **Actions** > **Drilldown** to view the rows of valid or invalid data for the column.
4. Click **Actions** > **Drilldown Columns.**

   The columns appear in the **Drilldown** panel for the selected score. The Analyst tool displays the rows of valid data for the columns by default. Optionally, click **Invalid** to view the rows of data that are not valid.

Trend Charts

Use trend charts to monitor how the metric scores and cost of invalid data in metrics change over a period of time.

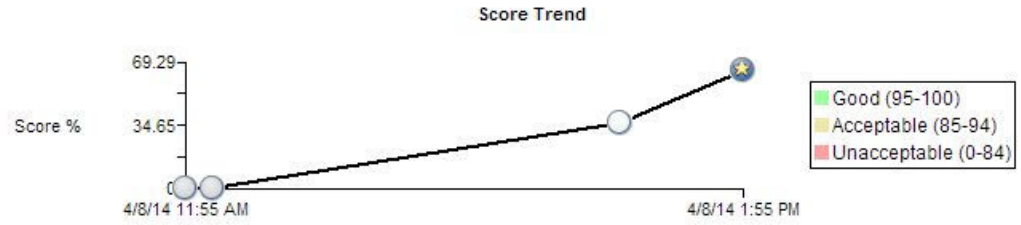
The trend charts contain both score and cost graphs that plot the score or cost values in the vertical axis against all the scorecard runs in the horizontal axis. By default, the trend chart shows data from the last 10 scorecard runs. You can view the number of total rows and invalid rows for the metric in the trend chart. The trend chart also displays whether the score and cost trends remained constant or moved up or down based on the last scorecard run.

The Analyst tool uses the historical scorecard run data for each date and latest valid score values to calculate the score. The Analyst tool uses the latest threshold settings in the chart to depict the color of the score points. You can view the Good, Acceptable, and Unacceptable thresholds for the score. The thresholds change each time you run the scorecard after editing the values for scores in the scorecard. When you export a scorecard, the Analyst tool includes the trend chart information including the score and cost information in the exported file.

Score Trend Chart

A score trend chart is a graphical representation of how the metric scores change over multiple profile runs. The score trend chart plots the metric score values in the vertical axis against all the scorecard runs in the horizontal axis.

The following image shows a sample score trend chart:



Example

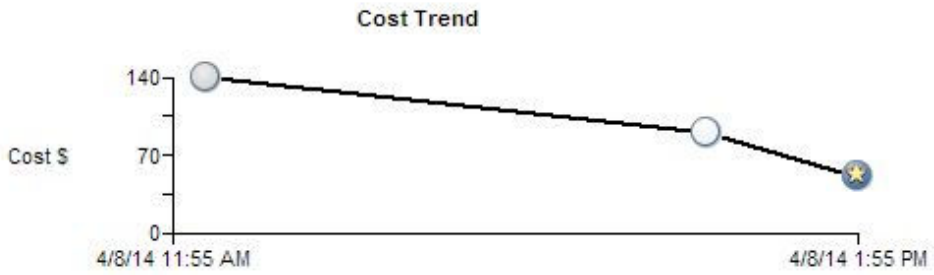
As a data analyst, you can monitor the data quality to analyze whether the mappings and other process changes result in increasing the data quality score. After you measure the change in data quality, you can report back the data quality change for the organization to analyze and use. For example, at the end of multiple scorecard runs, the percentage of valid values in a Social Security number column might have moved from 84 to 90. You can report this change in data quality as a visual chart for a quick analysis.

Cost Trend Chart

A cost trend chart is a graphical representation of how the cost of invalid data in metrics change over multiple profile runs. The cost trend chart can measure the impact of data quality in an organization. The cost trend chart plots the cost values in the vertical axis against all the scorecard runs in the horizontal axis. You can also view the total cost of invalid data and the valid values for the metric in a grid under the cost trend chart.

A cost trend chart helps you track the impact of invalid data on high-value records. Occasionally, when you use a fixed cost to compute invalid data, you might miss out on the impact of invalid data on high-value records. This issue happens because the trend charts might show an improvement in the score and decrease in the overall cost over multiple scorecard runs. However, the fewer data quality issues represented in the scorecard might exist on high-value records.

The following image shows a sample cost trend chart:



Example

In a financial institution, you have multiple high-balance customers with large deposits and investments, such as $10 million, in the bank. You also have a large number of low-balance customers. The score trend chart might show an improvement in scores over a period of time. However, an incorrect address or gender on a few high-balance customer accounts might impact the relationships with the most valuable customers of the organization. You can set the Account Balance column as the variable cost column for computing invalid

data. If the cost of invalid data due to the column is high, you can consider the total value at risk and take immediate, corrective action.

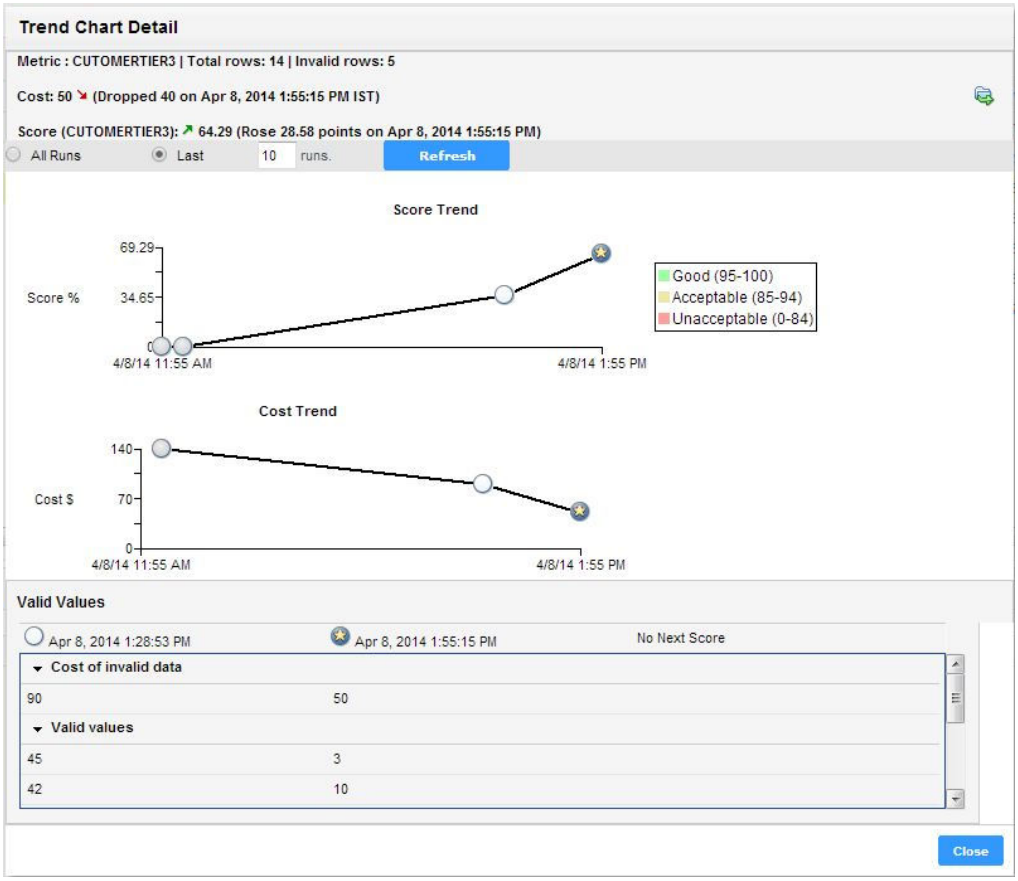
Viewing Trend Charts

You can view trend charts for each metric to monitor how the score or cost of invalid data changes over time.

1. In the **Library** workspace, select the project or folder that contains the scorecard.
2. Click the scorecard to open it.

   The scorecard appears in the **Scorecards** workspace.
3. In the **Scorecard** view, select a metric.
4. Click **Actions** > **Show Trend Chart**.

   The **Trend Chart Detail** dialog box appears.
   The following image shows the **Trend Chart Detail** dialog box:



   You can view score and cost values that have changed over time. At the top of the dialog box, you can view the total number of rows and the number of invalid rows. The Analyst tool uses historical scorecard run data for each date and the latest valid score values to calculate the score. Under the score and cost trend charts, you can view the valid values for the metric and the cost of invalid data.

Exporting Trend Charts

You can export the score and cost trend charts to a ".xlsx" file to view the data in a file.

1. Open a scorecard.
2. Select a metric, and click **Actions** > **Show Trend Chart**.

   The **Trend Chart Details** dialog box appears.
3. Click the **Export Data** icon.

   The **Export data to a file** dialog box appears.
4. Enter a file name. Optionally, use the default file name.

   The default file format is Microsoft Excel.
5. Select the code page of the file.
6. Click **OK**.

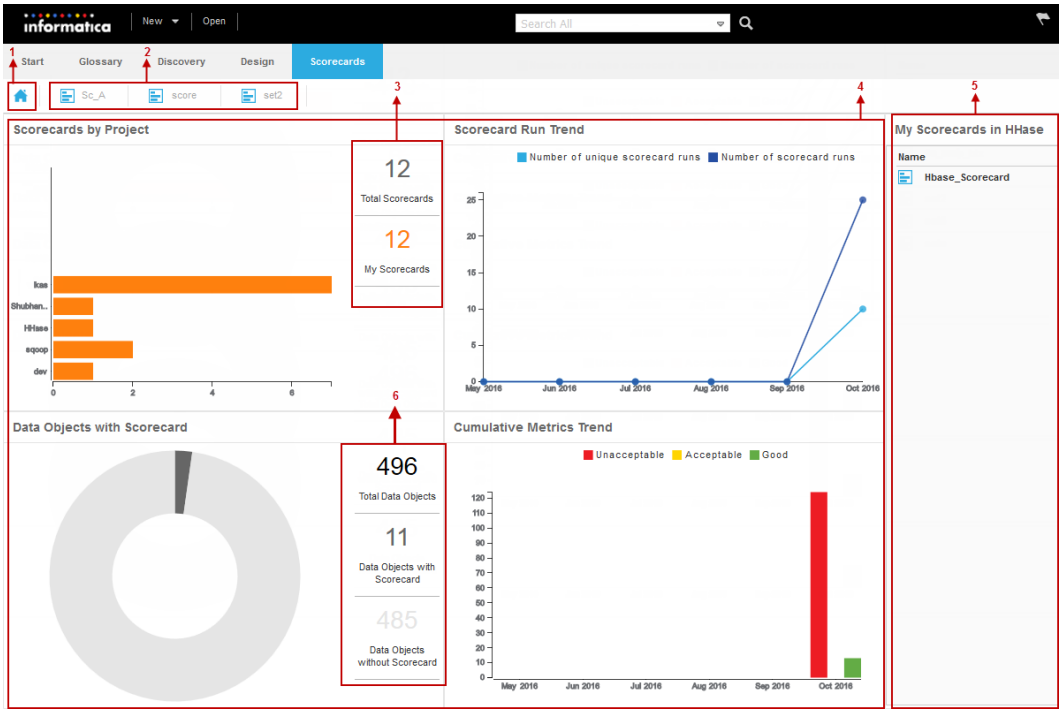
Scorecard Dashboard in Informatica Analyst

The **Scorecards** workspace in Informatica Analyst displays the scorecard dashboard. In the scorecard dashboard, you can view the data objects with scorecards, scorecard run trend for the past six months, scorecards in a project, an aggregate of good, acceptable, and unacceptable metrics for all the scorecard runs in a month, and the assets list pane.

The scorecard dashboard on your machine is not refreshed automatically when the scorecards are modified by other users. Use the F5 function key or toggle between workspaces or scorecard result tabs to refresh the scorecard dashboard.

You can view the data as a data series or as data points in the panes. The data points appear as small opaque circles and the data series appears as horizontal bars, vertical bars, or slices in the charts.

The following image shows the scorecard dashboard and the assets pane in the **Scorecards** workspace:



1. Scorecard dashboard icon. Displays the scorecard dashboard.
2. Scorecard result tabs. Displays the scorecard results for the open scorecards.
3. Legend in the Scorecards by Project pane. Displays the total number of scorecards in all the projects and the total number of scorecards that you have read access to in all the projects.
4. Scorecard dashboard. Displays the scorecards by project, scorecard run trend, data objects with scorecard, and cumulative metrics trend panes in the dashboard.
5. Assets list pane. Displays the list of scorecards or data objects associated with a legend, data series, or data point in the chart.
6. Legend in the Data Objects with Scorecard pane. Displays the total number of data objects, number of data objects with scorecards, and the number of data objects without scorecards

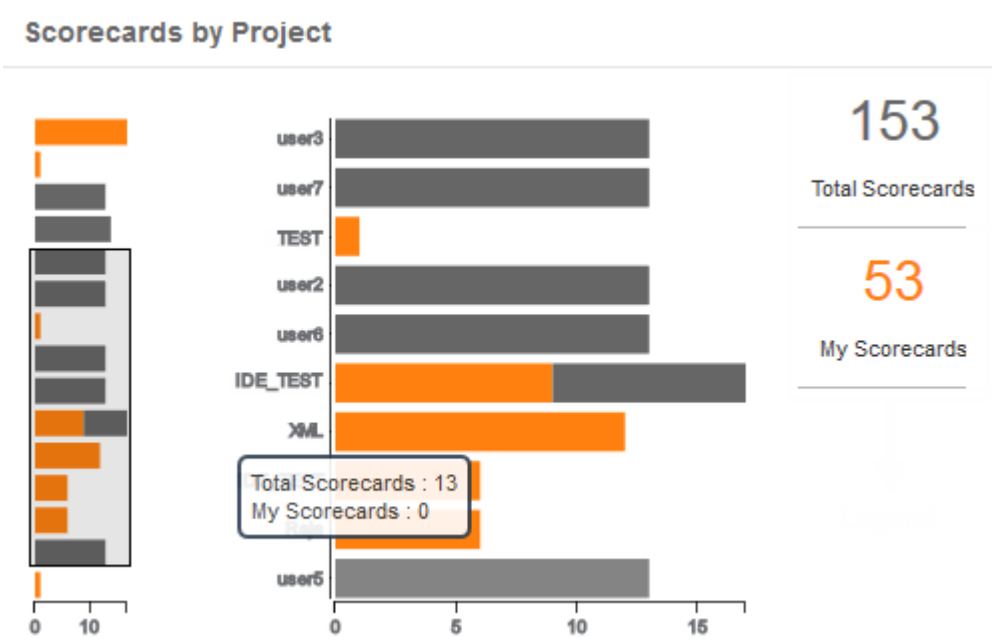
After you click a data point or data series in the scorecard dashboard, the scorecards that map to the data point or data series appear in the assets list pane. After you click a scorecard in the assets list pane, the scorecard results appear in a tab in the **Scorecards** workspace. The assets lists pane displays the scorecards for which you have read access.

Scorecards by Project

The **Scorecards by Project** pane displays the projects in the Model repository along with the number of scorecards in each project as a bar chart. A bar in the bar chart represents a project. The x-axis in the chart displays the number of scorecards and the y-axis displays the projects that has scorecards.

The scorecards in a project appear in Grey color in the bar chart and the scorecards that you have read access to appears in Orange color in the bar chart. The **Total Scorecards** section in the legend displays the total number of scorecards in the Model repository. The **My Scorecards** section in the legend displays the number of scorecards that you have read access to in the Model repository.

The following image shows the **Scorecards by Project** pane in the scorecard dashboard:



You can view the following charts in the pane:

- Detailed chart. Displays all the projects in the Model repository with scorecards and the number of scorecards in each project. If the number of projects is greater than 10, the **Scorecards by Project** pane displays a slider.
- Miniature chart. Displays all the projects and the number of scorecards in each project within the slider in the detailed chart.

When you move the pointer over the miniature chart, the total number of scorecards and the number of scorecards that you have access to in a project appears in a data label.

To view the scorecards in a project for which you have read access to, click the Orange part of the horizontal bar. To view all the scorecards for which you have read access to in the Model repository, click **My Scorecards** in the bar chart. The scorecards appear in the assets list pane. Click a scorecard in the assets pane to view the scorecard results.

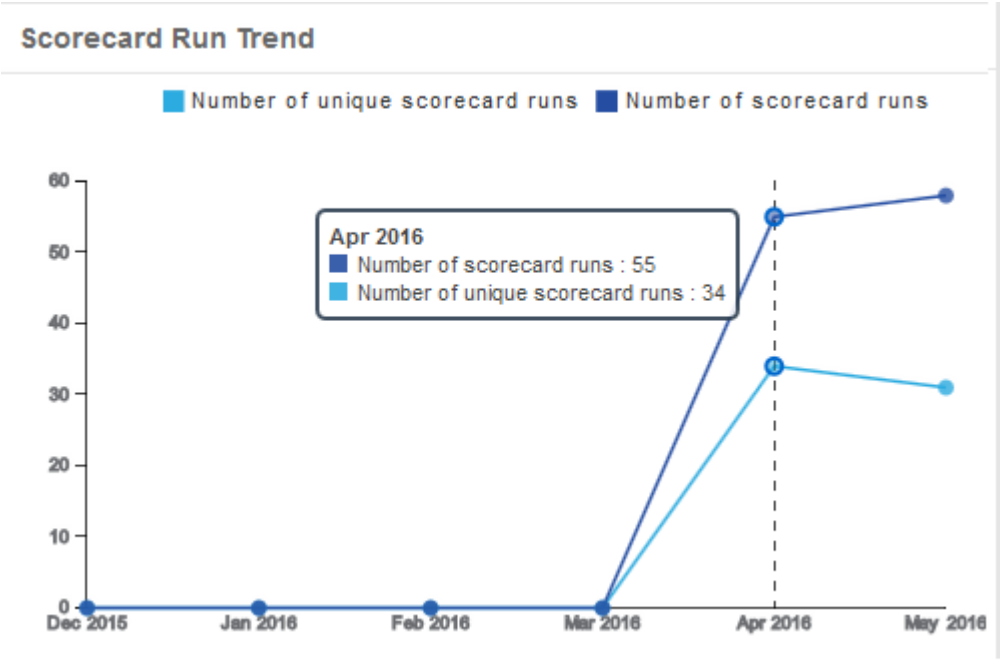
Scorecard Run Trend

The **Scorecard Run Trend** pane displays the scorecard run trend for the current month and the past five months as line charts with markers. The x-axis in the chart displays the current month and the past five months and the y-axis displays the number of scorecards. A marker is a data point in the line chart. When you move the pointer over a marker in the chart, the scorecard run summary for the month appears in a data label.

You can view the following markers in the pane:

- **Number of scorecard runs**. The marker displays the total number of scorecard runs in the month.
- **Number of unique scorecard runs**. The marker displays the total number of unique scorecard runs in the month.

The following image shows the **Scorecard Run Trend** pane in the scorecard dashboard:



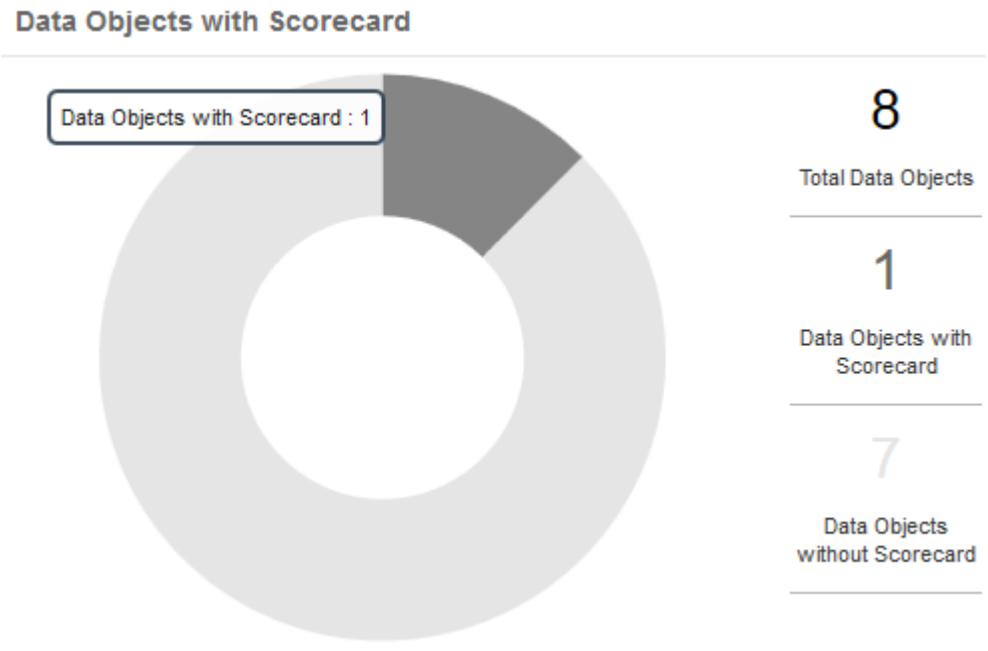
When you click a marker in the pane, the scorecards that map to the marker appear in the assets list pane. You can view the scorecards for which you have read access. Click a scorecard in the assets list pane to view the scorecard results in the **Scorecards** workspace.

Data Objects with Scorecards

The **Data Objects with Scorecards** pane displays a doughnut chart. You can view the number of data objects with scorecards and the number of data objects without scorecards as slices in the chart.

When you move the pointer over the doughnut chart, the data that map to the slice appears in a data label.

The following image shows **Data Objects with Scorecards** pane in the scorecard dashboard:



The legend in the **Data Objects with Scorecards** pane displays the following data statistics:

- Total Data Objects. Displays the total number of data objects in the **Data Object** folder in the **Assets** pane in the **Library** workspace. The data objects includes logical data objects and customised data objects.
- Data Objects with Scorecard. Displays the number of data objects with scorecards.
- Data Objects without Scorecard. Displays the number of data objects without scorecards.

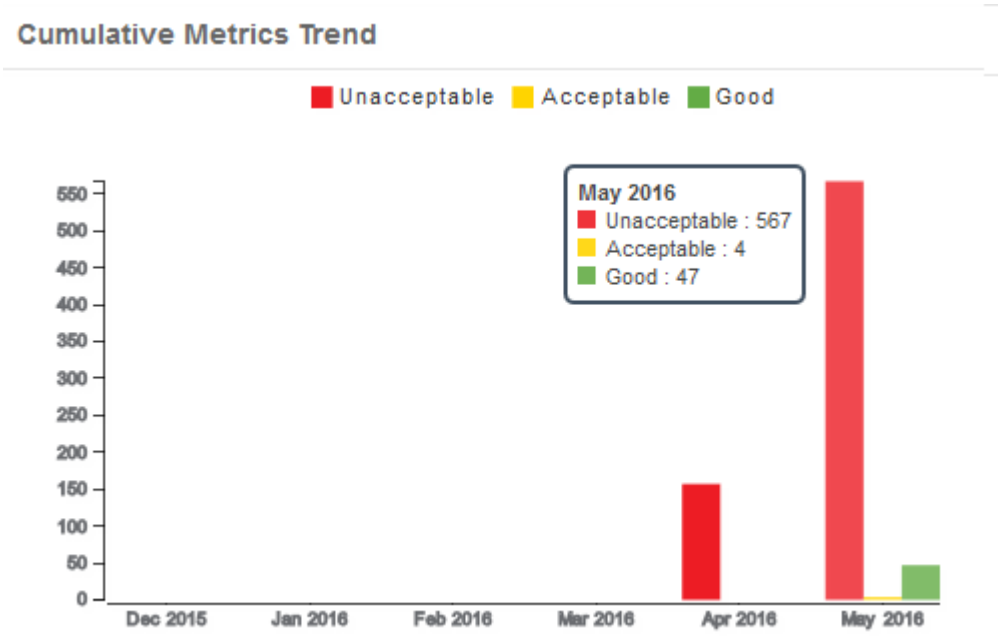
After you click the slices in the doughnut chart or the **Data Objects with Scorecard** and **Data Objects without Scorecard** legend, the scorecards that map to the slice in the doughnut chart or the legend appear in the assets list pane.

Cumulative Metrics Trend

The **Cumulative Metrics Trend** pane displays column charts. You can view the aggregate of good, acceptable, and unacceptable metrics for all the scorecard runs in a month in the chart as vertical bars. You can use the **Cumulative Metrics Trend** pane to view and analyze the metrics trend for the current month and the past five months.

When you move the pointer over a vertical bar, the metric summary for the month appears in a data label. When you click a vertical bar in the pane, the relevant scorecards appear in the assets list pane. You can view the scorecards for which you have read access. Click a scorecard in the assets list pane to view the scorecard results.

The following image shows the **Cumulative Metrics Trend** pane in the scorecard dashboard:



The assets list pane might display a few scorecards in the unacceptable metric list and good metric list when the scorecard trend changes with time in a month. To analyze the metrics, open the scorecard to view the scorecard results.

Example

You are the regional manager for a string of retail stores in the state of California. You create the Sales_SC scorecard on the Sales table.

You set the following thresholds for the Sales_amt metric in the Sales_SC scorecard:

- Unacceptable = 0% to 40%
- Acceptable = 41% to 89%
- Good = 90% to 100%

To capture the sales data on a daily basis, you use the scheduler service to run the scorecards every night on the Sales table. You prepare a monthly report for the management for the month of May and you use the scorecard dashboard to verify your report. For the month of May, when you click the vertical bars for Unacceptable metrics and Good metrics in the chart, the Sales_SC scorecard appears in the assets list pane.

When you analyse the sales across the month, you find the following trend:

1. From 1st May through 25th May, the Sales_amt metric is computed below 40% and is marked as an unacceptable metric.
2. In the last week of May, due to a rise in sales, the Sales_amt metric is computed above 98% and is marked as a good metric.

Scorecard Export Files in Informatica Analyst

You can export scorecard results to a Microsoft Excel file. The Analyst tool saves the file in the "xlsx" format.

When you export a scorecard, you can choose to export the scorecard summary, trend charts, rows that are not valid, and scorecard properties to the Microsoft Excel file. Alternatively, you can export only the scorecard summary, trend charts, and scorecard properties to the Microsoft Excel file.

Scorecard Export Results in Microsoft Excel

When you export the scorecard results, the Analyst tool saves the information to multiple worksheets in a Microsoft Excel file. The scorecard summary, trend charts, invalid rows, and scorecard properties appear as worksheets in the file. The Analyst tool saves the file in the "xlsx" format.

The following table describes the information that appears on each worksheet in the export file:

Tab	Description
Scorecard Summary	Summary information of the exported scorecard results. The information includes scorecard name, total number of rows for each column, number of rows that are not valid, score, and metric weight.
Trend Chart	Trend charts for scores.
Invalid Rows	The details of rows that are not valid for each column. The Analyst tool exports a maximum of 100 rows to the worksheet. The Invalid Rows worksheet appears when you choose the **Data** > **All** option in the **Export data to a file** dialog box.
Properties	Scorecard properties, such as name, type, description, and location.

Exporting Scorecard Results from Informatica Analyst

You can export scorecard results to a ".xlsx" file to view the data in a file.

1. Open a scorecard.
2. Click **Actions** > **Export Data.**

   The **Export Data to a file** dialog box appears.
3. Enter a file name. Optionally, use the default file name.

   The default file format is Microsoft Excel.
4. Select the code page of the file.
5. Click **OK**.

Scorecard Notifications

Configure scorecard notification settings so that the Analyst tool sends emails when specific metric scores, metric group scores, or metric costs move across thresholds. Metric scores or metric group scores might move across thresholds or remain in specific score ranges, such as Unacceptable, Acceptable, and Good. Metric cost values can move beyond the upper and lower cost thresholds that you set.

You can configure email notifications for individual metric scores, metric groups, and metric costs. If you use the global settings for scores, the Analyst tool sends notification emails when specific metric scores cross the threshold from the score ranges Good to Acceptable and Acceptable to Bad. You also get notification emails for each scorecard run if the score remains in the Unacceptable score range across consecutive scorecard runs. If you use the global settings for metric costs, the Analyst tool sends notification emails when the cost of invalid data in selected metrics crosses the upper and lower thresholds.

You can customize the notification settings so that scorecard users get email notifications when the scores move from the Unacceptable to Acceptable and Acceptable to Good score ranges. You can choose to send email notifications if a metric score or metric cost remains within specific ranges for every scorecard run. You can view the current cost of invalid data for each metric in the notification settings based on which you can set the cost thresholds.

Before you configure scorecards to send email notifications, an administrator must configure the Email Service in the Administrator tool.

Notification Email Message Template

You can set up the message text and structure of email messages that the Analyst tool sends to recipients as part of scorecard notifications. The email template has an optional introductory text section, read-only message body section, and optional closing text section.

The following table describes the tags in the email template:

Tag	Description
ScorecardName	Name of the scorecard.
ObjectURL	A hyperlink to the scorecard. You need to provide the username and password.

Tag	Description
MetricGroupName	Name of the metric group that the metric belongs to.
CurrentWeightedAverage	Weighted average value for the metric group in the current scorecard run.
CurrentRange	The score range, such as Unacceptable, Acceptable, and Good, for the metric group in the current scorecard run.
PreviousWeightedAverage	Weighted average value for the metric group in the previous scorecard run.
PreviousRange	The score range, such as Unacceptable, Acceptable, and Good, for the metric group in the previous scorecard run.
MetricName	Name of the metric.
MetricGroupName	Name of the metric group.
CurrentScore	Score based on the latest scorecard run.
CurrentRange	Score range in which the current score remains based on the latest scorecard run.
PreviousScore	Score based on the previous scorecard run.
PreviousRange	Score range based on the previous scorecard run.
CurrentCost	Cost of invalid data in the metric based on the latest scorecard run.
PreviousCost	Cost of invalid data in the metric based on the previous scorecard run.
ColumnName	Name of the source column that the metric is assigned to.
ColumnType	Type of the source column.
RuleName	Name of the rule.
RuleType	Type of the rule.
DataObjectName	Name of the source data object.

Setting Up Scorecard Notifications

You can set up scorecard notifications at both metric and metric group levels. Global notification settings apply to those metrics and metric groups that do not have individual notification settings.

1. Run a scorecard in the Analyst tool.
2. Click **Actions** > **Edit**.
3. Click the **Notifications** tab.
4. Select **Enable notifications** to start configuring scorecard notifications.
5. Select a metric or metric group.
6. Click the **Notifications** check box to enable the global settings for the metric or metric group.

7. Select **Use custom settings** to change the settings for the metric or metric group.

   You can choose to send a notification email when the score is in **Unacceptable**, **Acceptable**, and **Good** ranges and moves across thresholds. You can also send a notification email when the metric cost crosses the upper or lower thresholds.

8. To edit the global settings for scorecard notifications, click the **Edit Global Settings** icon.

   The **Edit Global Settings** dialog box appears where you can edit the settings including the email template.

Configuring Global Settings for Scorecard Notifications

If you choose the global scorecard notification settings, the Analyst tool sends emails to target users when the score is in the **Unacceptable** range. You can also configure the notification settings to send emails when the metric scores or metric costs move across thresholds. You can configure the email template including the email addresses and message text for a scorecard.

1. Run a scorecard in the Analyst tool.
2. Click **Actions** > **Edit** > **Notifications** to open the **Edit Scorecard** dialog box.
3. Select **Enable notifications** to start configuring scorecard notifications.
4. Click the **Edit Global Settings** icon.

   The **Edit Global Settings** dialog box appears where you can edit the settings, including the email template.

5. Choose when you want to send email notifications for metric scores using the **Score in** and **Score moves** check boxes.
6. Choose when you want to send email notifications for metric costs using the **Cost goes** check boxes.
7. In the **Email to** field, enter the email ID of the recipient. Use a semicolon to separate multiple email IDs.

   The default sender email ID is the **Sender Email Address** that is configured in the domain SMTP properties.

8. Enter the text for the email subject.
9. In the **Body** field, add the introductory and closing text of the email message.
10. To apply the global settings, select **Apply settings to all metrics and metric groups**.
11. Click **OK**.

Scorecard Lineage

Scorecard lineage shows the origin of the data, describes the path, and shows how the data flows for a metric or metric group. You can use scorecard lineage to analyze the root cause of an unacceptable score variance in metrics or metric groups. View the scorecard lineage in the Analyst tool.

Complete the following tasks to view scorecard lineage:

1. In Informatica Administrator, associate a Metadata Manager Service with the Analyst Service.
2. Select a project and export the scorecard objects in it to an XML file using the Export Resource File for Metadata Manager option in the Developer tool or infacmd oie exportResources command.
3. In Metadata Manager, use the exported XML file to create a resource and load it.

Note: The name of the resource file that you create and load in Metadata Manager must use the following naming convention: <MRS name>_<project name>. For more information about how to create and load a resource file, see *Informatica PowerCenter Metadata Manager User Guide*.

4. In the Analyst tool, open the scorecard and select a metric or metric group.
5. View the scorecard lineage.

Viewing Scorecard Lineage in Informatica Analyst

You can view a scorecard lineage diagram for a metric or metric group. Before you can view scorecard lineage diagram in the Analyst tool, you must load the scorecard lineage and metadata in Metadata Manager.

1. In the **Library** workspace, click the scorecard you want to view in the **Assets** pane.

   The scorecard appears in the **Scorecards** workspace.
2. In the **Scorecard** view, select a metric or metric group.
3. Right-click and select **Show Lineage**.

   The scorecard lineage diagram appears in a new window.

   Important: If you do not create and load a resource in Metadata Manager with an exported XML file of the scorecard objects, you might see an error message that the resource is not available in the catalog. For more information about exporting an XML file for scorecard lineage, see "Exporting a Resource File for Scorecard Lineage" on page 168.

CHAPTER 12

Data Domain Discovery in Informatica Analyst

This chapter includes the following topics:

Data Domain Discovery in Informatica Analyst Overview

When you create a profile to perform data domain discovery, you can select the source columns, data domains with which you want to match column data and column name, and sampling options. You can choose a conformance criteria for data domain discovery and exclude null values during data domain discovery.

You can create a profile with a sampling option and filters to perform data domain discovery. When you run the profile, you apply the sampling option and filters on the data source and generate a data set. The data domain discovery process uses the data set to discover data domains.

Data Domain Glossary in Informatica Analyst

The data domain glossary lists data domains and data domain groups. You can sort the list by either data domains or data domain groups. Use the data domain glossary to search, add, edit, and remove data domains and data domain groups. You can view and make changes to the rule logic associated with data domains in the Developer tool.

Creating a Data Domain Group in Informatica Analyst

A data domain group organizes data domains into specific groups such as Personal Health Information (PHI), Personally Identifiable Information (PII), or any other conceptual group that is relevant for the project.

1. Click **Manage** > **Data Domain Glossary**.

   The data domain glossary opens in a tab that lists the current data domains and data domain groups.
2. In the Navigator, click **Actions** > **New** > **Data Domain Group**.

   The **Create Data Domain Group** dialog box appears.
3. Enter a name and description for the data domain group.
4. Click **Next**.
5. In the **Available Data Domains** pane, select the data domains that you want to add to the data domain group and click **Add**.

   The Analyst tool moves the selected data domains to the **Selected Data Domains** pane.
6. Click **Finish**.

   The Analyst tool adds the data domain group to the data domain glossary.

Creating a Data Domain in Informatica Analyst

You can create data domains, add them to the data domain glossary, and group data domains into one or more data domain groups. To create a data domain, you can use predefined data rules and column name rules or generate data domains from the values and patterns in the column profile results.

When you create a data domain, the Analyst tool copies rules and other dependent objects associated with the data domain to the data domain glossary. To edit a rule associated with a data domain, you must go to the original rule and make changes to it. You can then associate the modified rule to the data domain again.

1. Click **Manage** > **Data Domain Glossary**.

   The data domain glossary opens in a tab that lists the current data domains and data domain groups.
2. In the Navigator, click **Actions** > **New** > **Data Domain**.

   The **Create Data Domain** dialog box appears.
3. Enter a name and description for the data domain.
4. Click the **Data Rule** check box to discover data domains based on column data. You can also select the **Column Name Rule** check box to discover data domains based on column titles in the data source.

   The **Select** button is enabled.
5. Click **Select** to open the **Select Rule** dialog box.
6. Select an appropriate rule and click **OK**.

   The rules that you selected appear in the **Data Rule** and **Column Name Rule** fields.
7. Click **Next**.
8. In the **Available Data Domain Groups** pane, select the data domain groups you want to include the data domain in and click **Add**.

   The Analyst tool moves the selected data domain groups to the **Selected Data Domain Groups** pane.
9. Click **Finish**.

   The Analyst tool adds the data domain to the data domain glossary.

Creating a Data Domain from Profile Results in Informatica Analyst

Run a column profile to view the values and patterns of source data. You can then verify the profile results and create a data domain from them.

1. Run a column profile to view its results.

   The profile results appear in summary view.
2. In the summary view, click a column to view column results in the detailed view.
3. In the **Values** pane or **Patterns** pane, select the value or pattern based on which you want to create a data domain.
4. Right-click the value or pattern, and then select **Create Data Domain**.

   The **Create Data domain** dialog box appears.
5. Enter the data domain name and an optional description.
6. Click **Create**.

   The data domain is added to the data domain glossary.

Find Data Domains and Data Domain Groups in Informatica Analyst

You can search for specific data domains and data domain groups in the data domain glossary. You can choose between the **Data Domain** view and **Data Domain Group** view to view the list of data domains in the data domain glossary.

For example, you might have a data domain **Zipcode** that you added to the **PII** data domain group. You can find more information about Zipcode and its data domain group PII in the following ways:

Search for data domains.

Type in part of the data domain name such as **zip** or **code** in the text field on top of the Navigator. If you are in the **Data Domain Groups View**, the Analyst tool lists **PII**, which is the data domain group that contains **Zipcode**. If you are in the **Data Domain View**, the Analyst tool lists all data domains that contain the search string, **Zip** or **code**.

Note: Search is not case-sensitive.

View all data domain groups and data domains within them.

In the Navigator, click **Show Data Domain Group View**.

View all data domains.

In the Navigator, click **Show Data Domain View**.

View properties of a data domain.

Verify that you are in **Show Data Domain View**. In the Navigator, click **Zipcode** to view its properties on the right pane. You can view the name, type, description, associated rules, and domain groups it belongs to, in this case, **PII**.

View properties of a data domain group.

Verify that you are in **Show Data Domain Group View**. In the Navigator, click **PII** to view its properties on the right pane. You can view the name, type, description, and the list of data domains, including **Zipcode**, within **PII**.

Refresh the data domain glossary.

In the Navigator, click **Actions** > **Refresh**. The data domain glossary displays the current list of data domains or data domain groups based on the view you are in.

Data Domain Discovery Options in Informatica Analyst

Use the data domain discovery options to choose the columns, data domains, and inference options for data domain discovery. Inference options include choosing whether you want to run data domain discovery based on a rule on column data, column name, or both.

Data Domain Column Selection in Informatica Analyst

You can click **Edit** in the **Specify Settings** screen to choose the columns you want to run as a part of data domain discovery. You can view all the columns in the data source in the **Select Source** screen in the profile wizard. You can choose different columns for column profile and data domain discovery.

The following table describes the **Edit** dialog box properties for data domain discovery:

Option	Description
Name	Displays the column name.
Type	Displays the documented data type of the column.
Precision	Displays the maximum precision for the column.
Scale	Displays the scale of the column.
Nullable	Indicates a column that can have null values.
Key	Indicates whether the column is documented as a primary key or foreign key.

Data Domain Selection in Informatica Analyst

The **Data Domain** pane in the **Specify Settings** screen lists all the data domains from the data domain glossary. You can choose the data domains you want to run as a part of data domain discovery.

The following table describes the **Data Domain** properties for data domain discovery:

Option	Description
Name	Displays the data domain name. You can choose one or more data domains or data domain group.
Description	Displays the description for the data domain.
DomainGroups	Displays the name of the data domain group to which the data domain belongs.

Data Domain Inference Options in Informatica Analyst

Inference options determine whether data domain discovery must run on column data, column name, or both. You can specify the maximum number of source rows the profile can analyze. You can choose a conformance criteria for data domain discovery. You can exclude null values from data domain discovery. You can set the data domain inference options in the **Specify Settings** screen in the profile wizard.

The following table describes the inference options for data domain discovery:

Option	Description
Data	Runs the profile on column data.
Columns	Runs the profile on column titles.
Data and Columns	Runs the profile on both column data and column titles.
Minimum percentage of rows	The minimum conformance percentage of rows in the data set required for a data domain match.
Minimum number of rows	The minimum number of rows in the data set required for a data domain match.
Exclude null values for data domain discovery	Excludes the null values from the data set for data domain discovery.
Edit	Select the columns for data domain discovery.
All Rows	Runs the profile on all rows from the source.
Sample first	Choose maximum number of rows the profile can run on. The Analyst tool chooses the rows starting from the first row in the source.
Random sample	Choose a random sample of rows from the data source.
Random sample (auto)	The Analyst tool chooses a random sample of rows based on the size of the data source.
Exclude approved data types and data domains from the data type and data domain inference in the subsequent profile runs	Excludes the approved data type or data domain from data type and data domain inference from the next profile run.

Minimum Conformance Percentage

You can choose a minimum percentage rows in the data set as a conformance criteria for data domain discovery.

The conformance percentage is the ratio of the number of matching rows divided by the total number of rows.

Note: The Analyst tool considers null values as nonmatching rows. Columns containing a high number of null values might not result in data domain inference unless you specify a low value for minimum conformance percentage.

Example

You have a data source with 10,000 rows where the Comments column has Social Security Numbers in 2,500 rows. You create a column profile and data domain discovery and set a minimum percentage of rows to 30% as the conformance criteria. When you run the profile, the profile results do not display the Social Security Numbers as an inferred data domain because the minimum conformance criteria is 30% of rows or 3,000 rows in the data source.

Minimum Conforming Rows

You can choose a minimum number of rows in the data set as a conformance criteria for data domain discovery.

Example

You have a data source with 10,000 rows where the Comments column has email address in three rows. You create a column profile and data domain discovery profile and set the minimum number of rows to 1 as the conformance criteria. When you run the profile, the profile results display the email address as an inferred data domain with three conforming rows along with the other inferred data domains.

Exclude Null Values in Data Domain Discovery

You can exclude null values when you perform data domain discovery on a data source. When you select the minimum percentage of rows with the exclude null values option, the conformance percentage is the ratio of number of matching rows divided by the total number of rows minus the null values in the column.

The data domain discovery process differs when you choose the **Exclude null values from data domain discovery** option and the multiple sampling options or filters.

The following scenarios explain the data domain discovery results when you choose the exclude null values option along with a sampling option and filters:

- With **All rows** as the sampling option and no filters. Data domain discovery ignores all the null values in the column.
- With a sampling option and no filters. Data domain discovery ignores all the null values in the sampled data and runs on the rest of the sampled data.
- With **All rows** as the sampling option and with filters. Data domain discovery ignores all the null values in the filtered data and runs on the rest of the filtered data.
- With a sampling option and filters. Data domain discovery ignores the null values in the filtered data in the sample and runs on the rest of the filtered data.

Example

You have a data source with 10,000 rows where 3,000 rows have Social Security Numbers in the Comments column. You create a column profile and data domain discovery and choose the following options:

- Select the **Exclude null values from data domain discovery** option.
- Select **All rows** as the sampling option.
- Select the **Minimum percentage of rows** option and configure the option to 12%.

When you run the profile, the profile runs on the data set and ignores the null values during data domain discovery.

Creating a Column Profile to Perform Data Domain Discovery in Informatica Analyst

You need to create at least one data domain before you can create a column profile to perform data domain discovery in the Analyst tool. The profile can discover both column name and column data that match predefined data domains.

1. In the **Discovery** workspace, click **Profile**, or select **New** > **Profile** from anywhere in the Analyst tool.

   The **New Profile** wizard appears.
2. The **Single source** option is selected by default. Click **Next**.
3. In the **Specify General Properties** screen, enter a name and an optional description for the profile. In the Location field, select the project or folder where you want to create the profile. Click **Next**.
4. In the **Select Source** screen, click **Choose** to select a data object, or click **New** to import a data object. Click **Next**.
5. In the **Specify Settings** screen, choose to run a column profile, data domain discovery, or a column profile and data domain discovery. By default, column profile option is selected.
   - Choose **Run data domain discovery** to perform data domain discovery. Select the data domain options in the **Data Domain** pane.
   - Choose **Run column profile** and **Run data domain discovery** to run the column profile and data domain discovery. Select the data domain options in the **Data domain** pane.

     Note: By default, the columns that you select for column profile is also applicable to data domain discovery. Click **Edit** to select or deselect columns for data domain discovery irrespective of the columns that you select for column profile.
   - Choose Data, Columns, or Data and Columns to run data domain discovery on.
   - Choose a sampling option in the **Run profile on** pane.
   - Choose a drilldown option in the **Drilldown** pane. Optionally, click **Select Columns** to select columns to drill down on. You can choose to omit data type and data domain inference for columns with approved data type or data domain.
   - Choose a conformance criteria, and you can select **Exclude null values from data domain discovery** option.
   - Choose **Native**, **Hive (deprecated)**, or **Hadoop** as the run-time environment. If you choose the Hive or Hadoop option, click **Choose** to select a Hadoop connection in the **Select a Hadoop Connection** dialog box.
6. In the **Specify Rules and Filters** screen, you can add, edit, or delete rules and filters for the profile.
7. Click **Save and Finish** to create the profile, or click **Save and Run** to create and run the profile.

Editing a Column Profile and Data Domain Discovery in Informatica Analyst

You can change the properties of a profile after you run it. If you have run a column profile as part of data domain discovery, you can change the column profile settings.

1. In the **Library** workspace, select the project that contains the profile, or select the profile in the **Assets** pane.

2. Click the profile name.

   The summary view appears in the **Discovery** workspace.
3. If the version control system is enabled, click **Actions** > **Check Out** to check out the profile.
4. Click **Actions** > **Edit Profile**.

   The **Profile** wizard appears.
5. Based on the changes you want to make, choose one of the following page options:
   - **Specify General Properties**. Change the basic properties such as name, description, and location.
   - **Select Source**. Choose another matching data source and columns to run the profile on.
   - **Specify Settings**. Choose to run column profile or column profile and data domain discovery. Edit the data domain options, sampling option, and drilldown option.
   - **Specify Rules and Filters**. Create, edit, or delete rules and filters.
6. Click **Save and Finish** to edit the profile, or click **Save and Run** to edit and run the profile.
7. If the version control system is enabled, you must perform the following tasks:
   - Click **Save and Finish** to complete editing the profile.
   - In the summary view, click **Check In** to check in the profile.
   - Click **Actions** > **Run Profile** to run the profile.

Running a Profile to Perform Data Domain Discovery

Run a profile as part of data domain discovery to view the columns that match data domain rule patterns.

1. In the **Library Navigator**, select the project or folder that contains the profile in the Projects pane, or select the profile in the Assets pane.
2. Click **Actions** > **Open**.

   The summary view appears in the **Discovery** workspace.
3. Click **Actions** > **Run Profile.**

   The Analyst tool performs a profile run and displays the profile results in summary view.
4. In the summary view, click a column to view the column results.

   The detailed view appears.

Data Domain Discovery Results in Informatica Analyst

You can view data domain discovery results in the summary view and detailed view.

The data domain field displays statistics about columns that match data domains. In the summary view, you can view the inferred data domains along with the percentage of conforming rows and the number of conforming rows.

In the detailed view, you can perform the following tasks:

- View the inferred data domains with the percentage of conforming rows and number of conforming rows in a horizontal bar chart.
- Drill down the results to conforming rows, non conforming rows, and null values.
- Approve, reject, or reset the data domain.
- Show or hide rejected data domains.
- Run data domain discovery on all the rows in the data source to discover inferred data domain.

Approving Data Domains

You can approve multiple data domains in the Analyst tool.

1. In the **Library** workspace, select the project or folder that contains the profile.
2. Click the profile to open it.

   The profile results appear in summary view.
3. Click the column for which you want to approve the data domain.

   The column results appear in detailed view.
4. In the detailed view, select the data domain in the **Data Domain** pane. Click **Actions** > **Approve**.

   The status of the column or data domain changes to Approved.
5. To restore the inferred status of the column or data domain, select the data domain and click **Actions** > **Reset**.

Rejecting Data Domains

When you open the profile results, the Analyst tool displays approved data domains by default. You can show or hide the rejected data domains.

1. In the Library Navigator, select the project or folder that contains the profile.
2. Click the profile to open it.

   The profile results appear in summary view.
3. Click the column for which you want to reject the data domain.

   The column results appear in detailed view.
4. To reject inferred data domains, click **Actions** > **Reject**.

   The Analyst tool removes the rejected data domain from the data domain discovery results.
5. To view the rejected data domains, click **Actions** > **Show Rejected**.
6. To hide the rejected data domains, click **Actions** > **Hide Rejected**.

Data Domain Discovery Export Files in Informatica Analyst

When you export data domain discovery results from the Analyst tool, you can specify the file name and code page value. You can export data domain discovery results to a Microsoft Excel file.

The Microsoft Excel file contains multiple worksheets that separate discovery results based on columns, data domains, and data domain groups. The Properties worksheet displays profile properties, such as name, description, type, location, date and time last changes were made to the profile, and a link to the profile.

Data Domain Discovery Results in Microsoft Excel

When you export the data domain discovery results to Microsoft Excel, the Analyst tool saves the column names, names of matching data domains, conformance criteria, and null values. The Excel file also contains the data domain group names for each data domain and documented data types for the columns.

The following table describes each worksheet in the export file:

Tab	Description
View by Columns	Data domain discovery results sorted by data source column.
View by Data Domains	Data domain discovery results sorted by data domain.
View by Data Domain Groups	Data domain discovery results sorted by data domain group.
Properties	Basic profile properties, such as name, description, type, location, date and time last changes were made to the profile, and a link to the profile.

Exporting Data Domain Discovery Results from Informatica Analyst

You can export the data domain discovery results to an `.xlsx` file so that you can view the data in a file and distribute it within the enterprise for further use.

1. Run a profile to perform data domain discovery.
2. In the summary view or detailed view, click **Actions** > **Export Data.**

   The **Export data to a file** dialog box appears.
3. Enter the file name. Optionally, use the default file name.
4. Select the code page of the file.
5. Click **OK**.

CHAPTER 13

Enterprise Discovery in Informatica Analyst

This chapter includes the following topics:

Enterprise Discovery in Informatica Analyst Overview

Enterprise discovery is the process of discovering column metadata and data domains in multiple data sources from many schemas and external relational connections. You can perform enterprise discovery on both data sources that you imported into the Model repository and data sources from external relational connections.

As a data analyst, you can perform enterprise discovery in the Analyst tool to infer specific metadata characteristics across a large number of data sources. You might also want to view the source data that matches predefined data domains. You can then curate the inferred enterprise discovery results and make the data ready for discovery search and data quality initiatives. Enterprise discovery in the Analyst tool generates a consolidated results summary of the profile results.

Enterprise discovery results include column profile statistics, such as patterns, unique values, and columns with data type conflicts. Data domain discovery identifies source columns that match predefined data domains.

You can choose an operating system profile in Informatica Analyst. After you choose an operating system profile, the Data Integration Service creates and runs the enterprise discovery profiles based on the permission of the operating system user that you define in the operating system profile.

Enterprise Discovery Process in Informatica Analyst

You can create, edit, and delete enterprise discovery profiles. You can run an enterprise discovery profile in the Discovery workspace. You need to configure the inference options for column profile and data domain discovery before you run the enterprise discovery profile.

Complete the following steps to perform enterprise discovery in the Analyst tool:

1. Configure the general properties for the enterprise discovery profile.
2. Select the data objects from the Model repository that you want to include in enterprise discovery profile.
3. Import relational data sources from external database connections.
4. Configure the data inference options and discovery options for the enterprise discovery profile.
5. Save the changes, and run the enterprise discovery profile.
6. Monitor the profile run and if required, view the statuses of profile tasks that the Analyst tool runs.
7. Review the enterprise discovery results summary. The results appear in the **Summary** and **Profiles** panels.

Configuration Options for Enterprise Discovery

The configuration options for enterprise discovery include data domain discovery options, column profile sampling options, and general profile properties such as name and description.

You can choose to run a column profile or a profile to perform data domain discovery. You can also choose to run both column profile and a profile to perform data domain discovery as part of the configuration.

Data Domain Discovery Settings

The data domain discovery settings include choosing whether data domain discovery must run on column data, column name, or both column data and column name. You can choose data domains and specify whether data domain discovery needs to process all the rows in the data source. You can choose a conformance criteria for data domain discovery. You can exclude nulls from data domain discovery.

The following table describes the data domain discovery settings that you can configure for enterprise discovery in the Analyst tool:

Option	Description
Enable data domain discovery	Performs data domain discovery as part of enterprise discovery.
Run data domain discovery on data	Performs data domain discovery on column data.
Run data domain discovery on column name	Performs data domain discovery on the name of each column.

Option	Description
Minimum conformance percentage	The minimum conformance percentage of rows in the data set required for a data domain match. The conformance percentage is the ratio of number of matching rows divided by the total number of rows. **Note:** The Analyst tool considers null values as nonmatching rows.
Minimum conforming rows	The minimum number of rows in the data set required for a data domain match.
Exclude null values from data domain discovery	Excludes the null values from the data set for data domain discovery.
Exclude columns with approved data domains	Excludes columns with approved data domains from the data domain inference of the profile run.
All rows	Performs data domain discovery on all source rows.
First	The maximum number of rows the profile can run on. The Analyst tool chooses rows starting from the first row in the source.

Column Profile Settings

The sampling options determine whether the Analyst tool runs a column profile on all rows of the data sources or limited number of rows.

The following table describes the column profile settings that you can configure for an enterprise discovery profile:

Option	Description
Enable column profiling	Runs a column profile as part of enterprise discovery.
Exclude approved data types and data domains from the data type and data domain inference in the subsequent profile runs	Excludes the approved data type or data domain from data type and data domain inference from the next profile run.

The following table describes the sampling options that you can configure for an enterprise discovery profile:

Option	Description
All Rows	Runs a column profile on all rows in the data source.
First <number> Rows	The number of rows that you want to run the column profile on. The Analyst tool chooses the rows starting from the first row in the data source.

The following table describes the run-time environment option that you can configure for an enterprise discovery profile:

Option	Description
Native	The Analyst tool submits the profile jobs to the Profiling Service Module. The Profiling Service Module then breaks down the profile jobs into a set of mappings. The Data Integration Service runs these mappings and writes the profile results to the profiling warehouse.
Hadoop	The Data Integration Service pushes the profile logic to the Blaze engine on the Hadoop cluster to run profiles.

Creating an Enterprise Discovery Profile in Informatica Analyst

You can run column profile and data domain discovery as part of enterprise discovery in Informatica Analyst.

1. In the **Discovery** workspace, select **New** > **Profile**.

   The **New Profile** wizard appears.
2. Select **Enterprise Discovery** . Click **Next**.

   The **Specify General Properties** tab appears.
3. In the **Specify General Properties** tab, enter a name for the enterprise discovery profile and an optional description. In the Location field, select the project or folder where you want to create the profile. Click **Next**.

   The **Select Data Objects** tab appears.
4. In the **Select Data Objects** tab, click **Choose**.

   The **Choose Data objects** dialog box appears.
5. In the **Choose Data objects** dialog box, choose one or more data objects to add to the profile. Click **Save**.

   The data objects appear in the **Data Objects** pane.
6. Click **Next**.

   The **Select Resources** tab appears.
7. In the **Select Resources** tab, click **Choose** to open the **Select Resources** tab.

   You can import data from multiple relational data sources.
8. In the **Select Resources** tab, select the connections, schemas, tables, and views that you want to include in the profile. Click **Save**.

   The left pane in the dialog box lists all the internal and external connections, schemas, tables, and views under the Informatica domain.

   The resources appear in the **Resource** pane.
9. Click **Next**.

   The **Specify Settings** tab appears.
10. In the **Specify Settings** tab, you can configure the column profile options and data domain discovery options. Click **Save and Finish** to save the enterprise discovery profile, or click **Save and Run** to run the profile.

You can perform the following tasks in the **Specify Settings** tab.

- Enable data domain discovery. Click **Choose** to select data domains that you want to discover from the **Choose Data Domains** dialog box. The selected data domains appear in the **Data Domains for Data Domain Discovery** pane.
- Run data domain on data, column name, or on both data and column name.
- Select all the rows in the data source, or choose a maximum number of rows to run domain discovery on.
- Choose a minimum conformance percentage or specify the minimum number of conforming rows for data domain discovery.
- Enable column profile settings and select all rows or first few rows in the data source for the column profile. You can exclude data type inference for columns with approved data types in the column profile.
- Choose **Native** or **Hadoop** as the run-time environment.

You can view the enterprise discovery results under the **Summary** and **Profiles** tabs.

Editing Enterprise Discovery Options

You can make changes to the enterprise discovery options after you perform enterprise discovery. You can rename the profile and change the data object selection, data domain selection, and inference options.

1. Open a profile that you ran to perform enterprise discovery.

   The profile results appear in the **Discovery** workspace.
2. If the version control system is enabled, click **Actions** > **Check Out** to check out the profile.
3. Click **Edit Profile**.
4. In the **Specify General Properties** tab, update the profile properties as required.
5. To make changes to the data object selection, click **Select Data Objects** tab.
6. To make changes to the external data sources of enterprise discovery, click **Select Resources** tab.
7. To make changes to the data domain inference options and column profile settings, click **Specify Settings** tab.
8. To apply the configuration changes to all the data domain profile tasks and column profile tasks in the enterprise discovery profile, select **Use global settings for all profiles**. If you do not select this option, the changes you make to the profile settings apply only to those data objects or resources that you newly added to the profile.

   By default, the changes you make apply to the newly added data objects in the enterprise discovery profile.
9. To undo the changes, click **Cancel**.
10. Click **Save and Run** to save the changes and run the profile again.
11. If the version control system is enabled, you must perform the following tasks:
    - Click **Save and Finish** to complete editing the profile.
    - In the summary view, click **Check In** to check in the profile.
    - Click **Actions** > **Run Profile** to run the profile.

CHAPTER 14

Enterprise Discovery Results in Informatica Analyst

This chapter includes the following topics:

Enterprise Discovery Results in Informatica Analyst Overview

You can view the results of enterprise discovery in the **Summary** and **Profiles** views.

The **Summary** view displays column profile results and data domain discovery results. The **Data Domain Discovery** section lists the data domains you included in the profile run and the number of columns that have a data domain match. The **Column Profiling** section displays statistics on the source columns. You can click each profile results row to view the details in the right pane of the **Summary** view.

Summary View

The **Summary** view displays a summary of the column profile results and data domain discovery results. You can view the data domain names that have matches in columns and number of columns with data domain matches. The column statistics include the number of columns for the top 10 pattern matches, all unique

values, and all null values. The column statistics also include the number of columns that have data type conflicts between inferred data types and documented data types.

Summary View Profile Results

The Summary view displays enterprise discovery results in Data Domain Discovery and Column Profiling sections.

Data Domain Discovery

The following table describes the columns in the data domain discovery results:

Column name	Description
Name	Name of the data domain.
Found in Columns	The total number of columns that have a data domain match.
Profile	Name of the profile that contains the matching column.
Column Name	Name of the matching column.
%Data Conformance Percentage	The minimum conformance percentage of rows required for a data domain match.
Connection Name	Name of the relational database connection.
Source Name	Name of the data source.
Inference Status	Data domain inference status. The statuses are **Accepted**, **Rejected**, and **Inferred**.
%Null	The percentage of null values for the column.
Total Rows	The total number of rows.
Conforming Rows	The minimum number of rows required for a data domain match.
Column Name Match	Indicates whether the column name matches a data domain name.
Documented Data Type	Data type declared for the column in the profile object.
Verified	Indicates the validation of the data domain match on all rows of the data source.
Last Run Time	Date and time of the last profile run.

Column Profiling

The following table describes the columns in the column profile results:

Column name	Description
Name	Name of the profile result type, such as pattern, 100% Nulls, and 100% Unique.
Found in Columns	Total number of columns that have matching profile result type.
Profile	Name of the profile that contains the matching column.
Connection	Name of the relational database connection.
Data source	Data source of the profile.
Number of columns	Number of columns in the profile that have matching profile result type.

Viewing Data Domain Discovery Results

You can click a data domain name to view its data domain discovery results. You can open specific profiles from the data domain discovery results.

1. Run a profile to perform enterprise discovery.
2. Verify that you are in the **Summary** view.
3. Click a data domain under the **Data Domain Discovery** section to view its discovery results.

   A list of profiles that contain the data domain appear in the right pane.
4. Select a row in the right pane, if required.

   The hyperlinks to the profile appear in Blue color.
5. Click the profile name link or column name link to open the profile.

   The profile opens and displays the data domain discovery results. The Analyst tool highlights the row with the data domain in the results. If required, you can curate the profile results for further effective use, such as Discovery Search.
6. To go back to the **Summary** view, click **Back to Enterprise Discovery.**.

Viewing Column Profile Results

You can view the column profile results of enterprise discovery in the **Summary** view. You can open specific profiles from the data domain discovery results.

1. Run a profile to perform enterprise discovery.
2. Verify that you are in the **Summary** view.
3. To view the details of inferred patterns, click one of the top 10 patterns under the **Column Profiling** section.

   A list of profiles that contain the inferred pattern results appear in the right pane.
4. To view information such as all null values, all unique values, or data type conflicts, click **100% Nulls**, **100% Unique**, or **Conflicting inferred vs. documented datatypes**.

   The matching list of profiles appears in the right pane.
5. Click the profile name link or column name link to open the profile.

The profile opens and displays the column profile results.

6. To go back to the **Summary** view, click **Back to Enterprise Discovery.**.

Data Type Conflict

Enterprise discovery identifies data type conflicts in columns. A data type conflict is a mismatch of inferred and documented data types of a column after you run enterprise discovery. An inferred data type is the data type that the Analyst tool derives for a data source column based on the column data. A documented data type is the declared data type for a column in the source database.

Enterprise discovery might infer a different data type for a column based on the column data compared to the documented data type of a column. For example, the enterprise discovery can infer a column with a documented string data type as a date data type. You can review the data type conflict, choose the most appropriate data type date for the column, and approve it.

Viewing Data Type Conflicts

When you open a profile with data type conflicts from the **Summary** view, the Analyst tool highlights the data type conflicts in Red.

1. Run a profile to perform enterprise discovery.
2. Verify that you are in the **Summary** view.
3. Under the **Column Profiling** section, click **Conflicting inferred vs. documented data types** to view the data type conflicts in column profile results.

   A list of profiles that contains columns with data type conflicts appear in the right pane.
4. Select a row in the right pane, if required.

   The hyperlinks to the profile appear in Blue color.
5. Click the profile name link or column name link to open the profile.

   The profile opens and displays the data type conflicts in Red color. You can choose to curate the inferred data types to resolve the data type conflict.
6. To curate the data types, select a row with the conflicting data types and click **data types** view.
7. Click **Actions** and then select **Approve** or **Reject**.
8. To go back to the **Summary** view, click **Back to Enterprise Discovery.**.

Profiles View

The **Profiles** view displays a list of all the single data object profiles that the Analyst tool runs as part of enterprise discovery. The profile list also displays the run status of each profile. You can open each profile to view the column profile results and data domain discovery results.

Viewing Profile Properties

You can view the list of profiles that are part of the enterprise discovery in the **Profiles** view. You can open each profile and curate the profile results, if required.

1. Run a profile to perform enterprise discovery.
2. Verify that you are in the **Profiles** view.
3. To view the profile properties of a profile, click the profile name.

   The profile properties appear in the right pane. The profile properties include name of the source data objects, connection name, and row count.
4. To view the profile results, click **Open Profile**.

   The profile displays the column profile results.
5. To go back to the **Profiles** view, click the folder or project name link at the upper left corner of the **Discoveries** workspace.

CHAPTER 15

Discovery Search in Informatica Analyst

This chapter includes the following topics:

Discovery Search in Informatica Analyst Overview

Discovery search finds assets and identifies relationships to other assets in the databases and schemas of the enterprise. Enterprise users can use discovery search to find where the data and metadata exists in the enterprise. You can search for specific assets, such as data objects, rules, and profiles.

If you perform a global search, the Analyst tool performs a text-based search for data objects, data sources, and folders. If you perform discovery search, in addition to the text matches, search results include objects with relationships to the objects that match the search criteria. Discovery search also includes matches based on profile metadata, such as data types and data patterns. For example, you can find objects that contain a specific data pattern and have names that contain a specific keyword.

Discovery search includes the following types of information in the search results:

Objects in the Model repository

Finds primary objects related to the objects that match the discovery search criteria. For example, when you search for a profile, the profile results include the data object of the profile.

Profile warehouse results

Includes inference results from profiles, such as a data domain or data pattern.

Business Glossary terms

Based on the license, includes metadata in the search, such as a business term associated with a rule.

Discovery Search Example

You are a data steward in the enterprise who is responsible for ensuring that sensitive enterprise data is appropriately masked. You might want to identify Personally Identifiable Information (PII) across the schemas and databases on which you or the data architect has performed enterprise discovery. You might have created data domains to identify important data that remains undiscovered in data sources. You perform the search on the "SSN" string. The Analyst tool displays the Social Security data domain and all matching columns from the data sources. In addition, the discovery search string might find additional columns or tables that have "SSN" in their descriptions or names. To narrow down your search, you can filter on mapping specifications to show those mapping specifications that reference the matching data objects. You can apply additional filters to filter additional mapping specifications based on projects or users. You might then want to open the mapping specifications in the results to verify that the mapping specifications meet the privacy policies of the enterprise.

Discovery Search Prerequisites

Before you can perform an effective discovery search across databases in the enterprise, perform enterprise discovery on the databases and schemas in the enterprise.

After you perform enterprise discovery, the Analyst tool stores all the profile results in the profiling warehouse. Verify that all the required data sources are in Model repository. Optionally, verify that the appropriate assets in the Model repository have business terms associated with the assets. When you perform discovery search, the Search Service retrieves search index information based on the Model repository assets and profiling warehouse results. The Search Service then uses the indexed information to display search results based on the appropriate object metadata and relationships.

Discovery Search Process in Informatica Analyst

You can search for assets based on criteria, such as text, patterns, and data types in the profile results. The search returns a list of assets related to the search string.

Complete the following steps to perform discovery search in the Analyst tool:

1. Perform enterprise discovery and run the required single data object profiles on data sources in the enterprise. When you perform discovery search, the Analyst tool searches for information in the profile results and Model repository objects.
2. Choose what type of information you want to find. For example, you might want to find all the assets associated with a data domain definition for sensitive data or specific pattern of data.
3. Perform the search.
4. Analyze the search results to identify the assets and their relationships to other assets.
5. If required, verify that the discovered data is compliant with the business requirements.

Discovery Search Options

You can perform a global search or discovery search to find assets and identify relationships to other assets. Global search retrieves results from the Model repository and optionally, Business Glossary. Discovery search retrieves results from the Model repository and Business Glossary, in addition to profiles based on profile results in the profiling warehouse.

You can search for assets, such as data objects, profiles, and mapping specifications. Enter a search string to search for assets that match the search string and have an association with the search string. You can use wildcard characters when you search for assets.

You can use the following wildcard characters when you search for assets:

*** (Asterisk)**

Add to the end of the search string to find all asset names that start with the search string. For example, to search for all asset names starting with the "emp" string, you can type in "emp*" in the search field.

? (Question mark)

Include in the search string to represent an alphanumeric character.

Note: You cannot start a search string with a wildcard character when you search for assets. The search is not case sensitive.

To search for two or more words together as a phrase, include the words in double quotation marks. Use the character `+` to represent the AND operator and search for a term that must appear in the search results. For example, if the search string is `+sensitive +data`, Search Service finds metadata that includes both the terms. Use a blank space to represent the OR operator. For example, if the search string is `sensitive data`, the Search Service finds metadata that contains either one of the terms.

If the search string contains a hyphen (-), underscore (_), or camel case, the Search Service finds the whole word and part words separated by the delimiter. For example, if you search for `Profile_Customer`, the search engine finds `Profile`, `Customer`, and `Profile_Customer` in the repositories. To include special characters, such as `*` and `?`, in your search string, include the search string that contains special characters in double quotes.

You can perform a discovery search that includes a keyword search and discovery filter. For example, you might want to find employee ID columns that use the format `<FirstNameInitial><LastnameInitial>-<SSN>` so that you can identify data security risks. To search for the employee ID columns, enter `Employee ID` in the **Search** panel of the Library workspace and set the pattern filter to `XX-999999999 <= 100%`.

Discovery Search Criteria

Use the discovery search criteria to search for information based on criteria, such as patterns, data types, unique values, and null values. You can use the conditional operators =, >=, or <=in the search.

The following table describes the discovery search criteria that you can use for discovery search:

Option	Description
Search	Text expression that you want to search for.
Clear	Clears the search string and all the other search criteria that you previously selected.
Pattern of	Column pattern and percentage that you want to include in the search. **Note:** The option does not accept control characters in a pattern.

Option	Description
Data type of	Column data type and percentage that you want to include in the search.
Unique values	Percentage of unique values in columns that you want to include in the search.
Nulls	Percentage of null values in columns that you want to include in the search.

Searching for an Asset

You can search for an asset in the **Library** workspace. The search results include assets created in both Developer tool and Analyst tool.

1. Open the **Library** workspace.
2. Verify that you are in the **Discovery Search** section.
3. In the **Search** field, type in the search string that you want to search.
4. Configure the search filters to narrow down your search.

   The filters include patterns, data types, unique values, and null values.
5. Click the **Search** icon.

Discovery Search Results in Informatica Analyst

Discovery search finds assets in all the licensed repositories for discovery search, such as Model repository and profiling warehouse.

The discovery search results include the total number of matches and list of matches. You can expand each match to view the match properties, direct match information, indirect match information, and the total number of related assets, if any. A direct match is a match with some or all the metadata of the asset that matches the search query. An indirect match is an asset match that is linked to the asset that directly matches the search query.

The order of search results that you see depend on the following factors:

- Object property that matches the search criteria. The name of the object has a higher priority than the object description. The object description has a higher priority than other object properties.
- Object type. Data domains and data domain groups have a lower priority than other objects.
- Curation. Curated profile results have a higher priority than profile results that you did not curate.
- Number of times the search criteria matches the objects including direct and indirect matches.
- Relative frequency of the keyword.

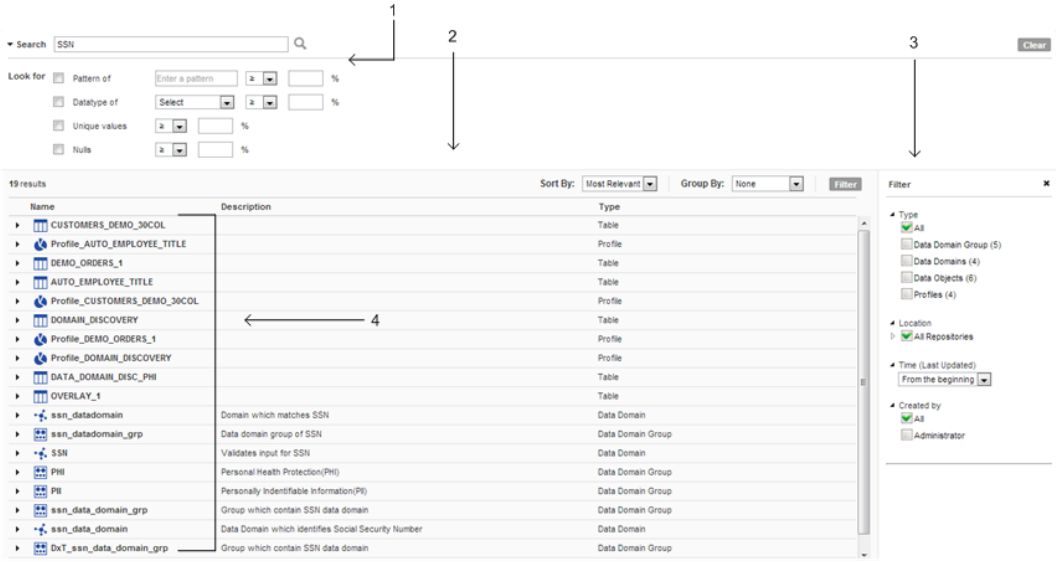
To view the search results, you need to have the appropriate permissions on projects that contain the direct matches and indirect matches.

Discovery Search Results Panel

Discovery search results include the name of the asset, asset type, and asset description. Use filters to narrow down the search results.

The search results appear in the results grid of the **Library** workspace. You can sort the results based on the relevance. You can also group the results based on the asset type, repository location, time, and user who created the assets.

The following image shows the discovery search results interface:



1. Search criteria
2. Results grid
3. Filter
4. Search results

The discovery search results panel displays the following sections by default:

Search criteria

Displays the search fields including the search filters based on profile metadata that you can set to narrow down the search. The Search fields appear at the top of the Analyst tool.

Results grid

Displays the total number of matches and list of matches based on the criteria you select in the search criteria fields. Results grid also contains object description, object type, and drop-downs to sort and group the search results.

Filter

Displays the filters that you can set to filter search results. The **Filter** section appears in the right area of the Analyst tool.

Search results

Displays the matching search results based on the search string including the objects that meet the search criteria. The search results section contains the match properties, **Direct matches** section, and **Indirect matches** section that appear when you expand a match. You can also view the total number of assets related to the match in the results grid.

Filtering Discovery Search Results

You can filter the search results based on the asset type, repository location, time, and user who created the asset. If you have Business Glossary installed, you can also use asset-specific filters for business terms, categories, and policies in Business Glossary.

1. Perform the global search or discovery search in the **Discovery Search** section in the **Library** workspace.
2. Click **Filter** in the result grid to open the **Filter** section.
3. In the **Filter** section, select the required filters and related settings.
4. The revised search results appear in the results grid based on the filter settings you selected.
5. To clear all the filter settings, click **Clear All** at the top of the results grid.

Match Types

Discovery search results include direct matches and indirect matches. A direct match is a match with some or all the metadata of the asset that matches the search query. An indirect match is an asset match that is linked to another asset that directly matches the search query.

If a search query contains multiple search criteria, search results might meet the search criteria directly, indirectly, or both. You can open some of the direct and indirect matches from the search results in read-only or edit mode.

Direct Match

A direct match is a match with some or all the metadata of the asset that matches the search query. For example, if you search for all assets with the name "Customer," the Analyst tool might list data objects and profiles with the name "Customer" as direct matches. After you perform a discovery search, the match list that you see contains links to some of the objects.

You can expand an asset in the search results to view more information about the direct match, such as asset properties.

Indirect Match

An indirect match is a match linked to the direct match. For example, a scorecard uses a rule, which contains the search keyword. Discovery search returns the rule as a direct match and scorecard as an indirect match. The scorecard is an indirect match because it references the rule.

Use the indirect match information to identify hidden relationships between objects and understand object relationships better. You can also use the indirect match results to understand why discovery search returns an object.

Viewing the Match Information

After you perform discovery search, you can view the match information including direct and indirect matches. You can also view the asset properties, such as asset type, description, and related assets. You can open some of the assets from the search results and make changes to them, if required.

1. Perform the global search or discovery search in the **Search** section of the **Library** workspace.
2. In the results grid, click the "expand" icon at the beginning of an asset name.

The asset properties and match information appear in a section under the asset name.

3. Review the direct match and indirect match information.

   You can see asset relationships and other information, such as the total number of related assets. The asset relationships include both direct matches and indirect matches.

4. If the asset information contains hyperlinks, click the hyperlinks to open the assets in another workspace.
5. Click the "expand" icon again to close the match information section.

Opening Assets from Discovery Search Results

You need to have the required project, asset, and license permissions to view assets from the discovery search results.

1. Perform the global search or discovery search in the **Search** section of the **Library** workspace.
2. In the results grid, right-click an asset name.

   A short-cut menu appears.

3. To view the asset in a read-only mode in its workspace, select **Open**.
4. To make changes to the asset in its workspace, select **Edit**.
5. To delete the asset from the search results, select **Delete**.

   When you delete an asset from the search results, the Analyst tool removes the asset from the Model repository.

6. To navigate back to the **Library** workspace, click **Library**.

Related Assets

You can view the related assets for an asset from the search results. A related asset is an asset in the Model repository or Business Glossary associated with a selected asset in the search results. The related asset shares some metadata with the asset in the search results. A data source can have profiles, inferred data domains, and mappings as related assets.

For example, a profile can be a part of the search results. You can view the related assets of the profile, such as rules and data sources for the profile. You can view the related assets in the **Related Assets** workspace. The related assets that you see differ based on the asset type. For example, when you view the related assets of a rule, you can view assets, such as the associated business term, mapping specification, and profile.

Related Assets for Each Asset Type

The related assets that you see for an asset in the Analyst tool depends on the type of the asset that you search for.

The following table describes the related assets for each asset:

Asset Type	Related Assets
Business term	Data domain, data object, mapplet, and rule.
Data domain	Business term, data domain group, data object, and profile.
Data domain group	Data domain, data object, and profile.
Data object	Business term, data domain, data domain group, mapping specification, profile, scorecard, mapping, and mapplet.
Enterprise discovery profile	Data object and profile.
Mapping **Note:** You open this object in the Developer tool.	Data object, mapping specification, mapplet, and rule.
Mapping specification	Data object, profile, scorecard, mapping, mapplet, and rule.
Mapplet **Note:** You open this object in the Developer tool.	Business term, data object, mapping specification, mapping, mapplet, and rule.
Profile	Data domain, data domain group, data object, rule, and mapping specification. **Note:** A scorecard is not included in the related assets for a profile.
Rule	Data object, rule, business term, mapping specification, profile, scorecard, and mapping.

Viewing Related Assets

You can view the total number of related assets when you view the asset match information in the search results.

1. Perform discovery search in the **Library** workspace.
2. In the results grid, click the expand icon and click the link for related assets count or right-click the asset name and select **Show Related Assets**.

   A list of all the related assets appears in the **Related Assets** workspace.
3. To view the details of an asset, click the asset name or right-click and select **Open**.
4. To view the related assets of a related asset, right-click the asset name and select **Show Related Assets**.

   The related asset information appears in the workspace.
5. To navigate between multiple related asset workspaces, from the **Related Assets** workspace, select one of the recently opened assets.

Frequently Asked Questions

Why am I unable to view some of the search results that I expect to see?

The search results might not appear due to various reasons. Verify that the search criteria meets the following guidelines:

- The assets that appear in the search results depend on project permissions.
- Discovery search results do not include value frequencies from the column profile results.
- The search results do not include the profile results that you reject when you curate the profile results.
- The search results that you view depends on the extraction interval of the search index and availability of the assets in the search index.

Can I save the discovery search results for future use or share the results with another user?

No. You cannot save or share the discovery search results.

Why do I see some of the discovery search results at the top or bottom of the search results?

The order in which the Analyst tool displays the search results depends on multiple factors. Some of the factors are object type, curated profile results, object property that primarily matches the search criteria, and internal search rank for each object.

Can I export the discovery search results?

No. You cannot export the search results.

CHAPTER 16

Business Glossary Desktop in Informatica Analyst

This chapter includes the following topics:

Business Terms

You can look up business terms in the Business Glossary Desktop. You can view business terms and perform business term tasks based on the license for Metadata Manager.

A business glossary is a set of terms that use business language to define concepts for business users. A business term provides the business definition and usage of a concept.

The Business Glossary Desktop is a client that connects to the Metadata Manager Service, which hosts the business glossary. You must have the Business Glossary Desktop open before you can look up an Analyst tool object name. You can look up the meaning of an Analyst tool object name as a business term in the Business Glossary Desktop to understand its business requirement and current implementation.

Metadata Manager hosts business glossaries. You must associate a Metadata Manager Service with the Analyst Service to browse a Metadata Manager business glossary from the Analyst tool. You can view the business terms in a business glossary or view business terms grouped by category. You can edit Metadata Manager business terms.

You can search for Metadata Manager objects in the Metadata Manager repository by a Metadata Manager business term. You can select Metadata Manager objects from the search results and import these as data objects in the Analyst tool. You cannot add a Metadata Manager business term to the Metadata Manager business glossary.

Managing Business Terms in Metadata Manager Business Glossary

You can access the Metadata Manager Business Glossary from the Analyst tool to manage Metadata Manager business terms.

1. On the Analyst tool header, click **Manage** > **Manage Terms**.

   Metadata Manager and the Metadata Manager Business Glossary open in another tab. Metadata Manager business terms appear on the **Glossary** view in Metadata Manager.
2. To choose a business glossary, select a glossary from the Show list.
3. To view business terms grouped by a category, click **Actions** > **View** > **Categories**.
4. To view all business terms in a business glossary in alphabetic order, click **Actions** > **View** > **Alphabet**.
5. To view all business terms that start with a specific letter, click the letter.
6. To edit a business term, select the business term and click **Actions** > **Edit Properties**.

Looking Up a Business Term in Business Glossary Desktop

Look up an Analyst tool object name in the Business Glossary Desktop as a business term to understand its business requirement and current implementation.

You must have the Business Glossary Desktop installed on your machine.

1. Highlight the name of an object.
2. Use the hotkey combination to look up the name of the object as a business term in the Business Glossary Desktop.

   The default hotkey combination is `SHIFT+ALT+Q`.

Part III: Data Discovery with Informatica Developer

This part contains the following chapters:

CHAPTER 17

Informatica Developer Profiles

This chapter includes the following topics:

Informatica Developer Profiles Overview

Create and run profiles in Informatica Developer to discover data quality issues in a data set and understand the column relationships in a data set.

You can create profiles for the following types of data analysis:

- Column profiling
- Column profiling on semi-structured data sources
- Primary key discovery
- Functional dependency discovery
- Foreign key discovery
- Join analysis
- Overlap discovery
- Data domain discovery
- Enterprise discovery

You create profiles in the Developer tool through a wizard. The profile creation wizard gives you **Profile**, **Multiple Profiles**, and **Enterprise Discovery Profile** options to create profiles.

Profile

Create a profile for a single data object. For a single profile, you define filters, rules, and drill-down options for column profiling. You can also choose advanced options to create a column profile, primary key profile, functional dependency profile, and for data domain discovery. The results display column profile, primary key inference, functional dependency, and data domain inference. You can create a column profile for a flat file data object, relational data object, and semi-structured data objects.

Multiple Profiles

Create a set of profiles for multiple objects. The Developer tool creates a profile for each object and runs the profiles concurrently. When you create multiple profiles at one time, you cannot analyze data across objects.

Enterprise Discovery Profile

Build a data model from multiple data objects and create a profile that analyzes data across the objects. Create an enterprise discovery profile and add physical data objects to it that you want to profile together. You can create a data object profile, foreign key profile, and join profile. For each data object in the enterprise discovery profile, you can configure general properties, columns to profile, keys, and relationships. You can discover overlapping data in a data source or mutiple data sources.

You can also run enterprise discovery that creates and runs data discovery tasks, such as column profile, data domain discovery, primary key profile, and foreign key profile. Enterprise discovery runs on a large number of data sources across multiple connections.

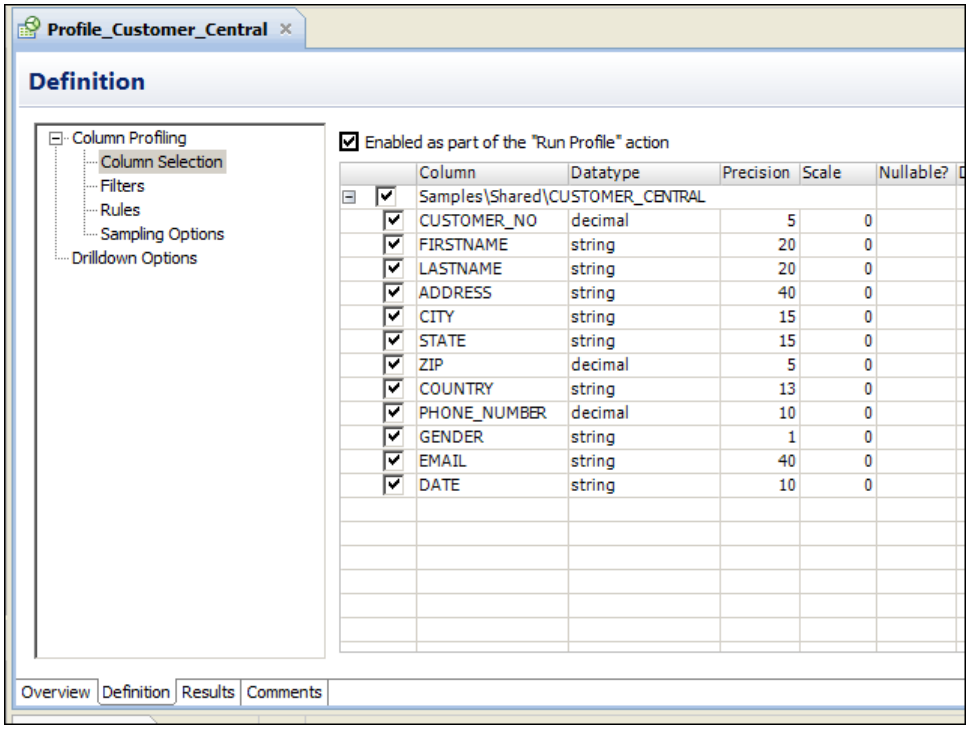
The following table lists the operations that you can perform with each profile type:

Profile Option	Profile Operations
Profile	- Run a column profile. - Find primary keys. - Find functional dependencies. - Identify data domains
Multiple Profiles	Create and run column profiles on multiple objects at the same time.
Enterprise Discovery Profile	- Run a column profile on a single data set. - Find primary keys. - Find foreign keys. - Find functional dependencies. - Perform join analysis. - Discover overlap between two columns. - Run enterprise discovery.

Informatica Developer Profile ViewsProfile Views

You can view and add information about a profile in Informatica Developer using the **Overview**, **Definition**, **Comments**, and **Results** views.

The following figure shows the profile views in the editor:



When you open a profile from the **Object Explorer** view, the editor on the right pane shows the profile information under the following views:

Overview

View and provide general information about the profile, such as name, description, and location.

Definition

View and set up the profile definition.

This information includes the list of filters and rules you assign to the profile, drilldown options, and profile functions enabled during the profile run.

This information includes the list of filters and rules you assign to the profile.

Results

Shows the results of profile run. You can export the results after you run a profile.

Comments

View and add comments to the profile.

Repository Object Locks and Team-based Development with Versioned Objects

The Model repository locks profiles to prevent users from overwriting the work of another user. If the Model repository is integrated with a version control system, it saves multiple versions of assets and assigns a version number to a version. You can check out and check in profiles, undo checkouts, and view profiles that you have checked out.

The Model repository retains object locks if the Developer tool stops unexpectedly. When you connect to the Model repository again, you can view the objects that you have locked. You can continue to edit the objects, or you can unlock the objects. You can view and unlock locked objects through the **Locked Objects** dialog box. To view the **Locked Objects** dialog box, click **View** > **Locked Objects**.

If the Model repository is integrated with a version control system, you can manage object versions in the Developer tool with the versioned object management. You can perform actions such as checking objects out and in, viewing and retrieving historical versions of objects, and undoing a checkout.

The Model repository protects objects from being overwritten by other members of the development team. If you open an object that another user has checked out, you receive a notification that identifies the user who checked it out. You can open a checked out object in read-only mode, or you can save it with a different name.

The Model repository creates new versions of objects when the object is modified.

The Model repository increments the version number after you curate a data type, primary key, foreign key, or data domain.

When you restore a version, the latest profile results appear in the Results view rather than the profile results of the restored version. This is because the version control system maintains profile definitions for all the versions in Model repository, and the profile results are extracted from the profiling warehouse. For more information about repository object locks and versioned object management, see the *Developer Tool Guide*.

CHAPTER 18

Data Object Profiles

This chapter includes the following topics:

Data Object Profiles Overview

A data object profile discovers information about the column data and metadata in a data source. You can run a profile on a single data object and multiple data objects in the Informatica Developer. A single data object profile analyzes one data source. Multiple data object profiles analyze more than one data source. When you create multiple data object profiles, you can run a column profile on them.

The following table describes the data discovery tasks that you can perform for a single data object profile:

Task	Description
Column profiling	Discovers the characteristics of data, such as frequencies, percentages, and patterns. You can add filters to determine the rows that the profile reads at run time. The profile does not process rows that do not meet the filter criteria.
Primary Key Discovery	Discovers columns with values that can uniquely identify the rows in a data source.
Functional Dependency Discovery	Discovers dependencies between pairs of columns in a data source.
Data Domain Discovery	Identifies all the data domains for a column based either on the column value or column name.

The following table describes the data discovery tasks that you can perform on multiple data objects when you create a data model using the **Enterprise Discovery Profile** option:

Task	Description
Foreign Key Discovery	Discovers columns that have values that match the primary key values in another data source.
Join Analysis	Discovers the degree of potential joins between the data in two columns in a data source or between two data sources.
Overlap Discovery	Discovers the percentage of overlapping data between pairs of columns in a data source or multiple data sources.
Enterprise Discovery	Discovers column profile statistics, data domains, primary keys, and foreign keys in a large number of data sources spread across multiple connections or schemas.

Column Profiles in Informatica Developer

Use a column profile to analyze the characteristics of columns in a data source, such as value percentages and value patterns. You can add filters to determine the rows that the profile reads at run time. The profile does not process rows that do not meet the filter criteria.

You can discover the following types of information about the columns that you run a profile on:

- The number of times a value appears in a column.
- Frequency of occurrence of each value in a column, expressed as a percentage or number of rows.
- Character patterns of the values in a column.
- Statistics, such as the maximum and minimum lengths of the values in a column, and the first and last values.
- Inferred data types, frequency, conformance criteria for data domain discovery, and data type inference status.

You can define a column profile for a data object in a mapping or mapplet or an object in the Model repository. The object in the repository can be in a single data object profile, multiple data object profile, or enterprise discovery profile.

You can choose sampling options, drill-down options, and run-time environment for a column profile. You can add rules and filters to a column profile.

Filtering Options

You can add advanced filters or SQL filters to determine the rows that a column profile uses when you run the profile. The profile does not process rows that do not meet the filter criteria.

Creating an Advanced Filter

You can create an advanced filter with expressions, such as AND, OR, and NOT to make a subset of the original data source.

1. Create or open a single data object profile.

2. Select the **Filter** view.
3. Click **Add**.

   The **Select Wizard** dialog box appears.
4. In the **Select Wizard** dialog box, click **Advanced filter**.

   The **Filter** dialog box appears.
5. Enter a name and an optional description for the advanced filter.
6. Select **Set as Active** to apply the filter to the profile. Click **Next**.
7. Select **Filter Definition** to define a filter.
8. You can create an advanced filter with the **Functions** panel or **Columns** panel.
   - In the **Functions** panel, select a function category, and click the right arrow (>>) button.
     In the dialog box, specify the parameters and click **OK**. The function along with the columns and values appears in the **Expression** panel.
   - In the **Columns** panel, select a column, and click the right arrow (>>) button. The column appears in the **Expression** panel.
     Add functions, expressions, and values to create an advanced filter.
9. To verify the advanced filter, click **Validate**.
10. After you create or edit the filter, select **Data Preview** to view the filtered data. You can set the **Maximum rows to preview** option.
11. Click **Finish**.

    The **New Profile** wizard appears with the filter in the **Filters** view.

Creating an SQL Filter

You can create an SQL filter with SQL queries. You can create an SQL filter for relational data sources.

1. Create or open a single data object profile.
2. Select the **Filter** view.
3. Click **Add**.

   The **Select Wizard** dialog box appears.
4. In the **Select Wizard** dialog box, click **Sql filter**.

   The **Filter** dialog box appears.
5. Enter a name and an optional description for the advanced filter.
6. Select **Set as Active** to apply the filter to the profile. Click **Next**.
7. Select **Filter Definition** to define a filter.
8. Use the columns in the **Columns** panel to create an SQL filter.
9. To verify the filter, click **Validate**.
10. After you create or edit the filter, select **Data Preview** to view the filtered data. You can set the **Maximum rows to preview** option.
11. Click **Finish**.

    The **New Profile** wizard appears with the filter in the **Filters** view.

Sampling Options

Configure the sampling options to determine the number of rows that the profile reads during a profiling operation.

The following table describes the sampling options:

Property	Description
All Rows	Chooses all rows in the data object.
First	The number of rows that you want to run the profile against. The Developer tool chooses the rows from the first rows in the source.
Random Sample of	The random sample algorithm chooses the rows at random in the data object to run the profile on.
Random Sample (Auto)	Random sample size is computed based on the number of rows in the data object.
Exclude approved data types and data domains from the data type and data domain inference in the subsequent profile runs	Excludes the approved data type or data domain from data type and data domain inference from the next profile run.

Property	Description
All Rows	Reads all rows from the source. Default is enabled.
First	Reads from the first row up to the row you specify.

After you choose to run the profile on a random sample of rows, the random sample algorithm chooses the rows at random in the data object to run the profile on. When you choose a random sampling option for column profiles, the Developer tool performs drilldown on the staged data. This can impact the drill-down performance. When you choose a random sampling option for data domain discovery profiles, the Developer tool performs drill down on live data.

Run-time Environment

Choose native or Hadoop as the run-time environment for a column profile. Informatica Developer sets the run-time environment in the profile definition after you choose a run-time environment.

Native Environment

When you run a profile in the native run-time environment, the Developer tool submits the profile jobs to the Profiling Service Module. The Profiling Service Module then breaks down the profile jobs into a set of mappings. The Data Integration Service runs these mappings on the same machine where the Data

Integration Service runs and writes the profile results to the profiling warehouse. By default, all profiles run in the native run-time environment.

You can use native sources to create and run profiles in the native environment. A native data source is a non-Hadoop source, such as a flat file, relational source, or mainframe source. You can also run a profile on a mapping specification or a logical data source with a Hive or HDFS data source in the native environment.

Hadoop Environment

You can choose Hive or Hadoop option to run the profiles in the Hadoop run-time environment.

If you choose the Hive option and select a Hadoop connection, the Data Integration Service pushes the profile logic to the Hive engine on the Hadoop cluster to run profiles. The Data Integration service executes only the environment SQL of the Hive connection. If the Hive sources and targets are on different clusters, the Data Integration Service does not execute the different environment SQL commands for the connections of the Hive source or target.

If you choose the Hadoop option and select a Hadoop connection, the Data Integration Service pushes the profile logic to the Blaze engine on the Hadoop cluster to run profiles.

When you run a profile in the Hadoop environment, the Developer tool submits the profile jobs to the Profiling Service Module. The Profiling Service Module then breaks down the profile jobs into a set of mappings. The Data Integration Service pushes the mappings to the Hadoop environment through the Hadoop connection. The Hive engine or Blaze engine processes the mappings and the Data Integration Service writes the profile results to the profiling warehouse.

Note: Hive engine is deprecated, and Informatica will drop support for it in a future release. You can still choose to run profiles on the Hive engine. In a future release, when Informatica drops support for the Hive engine, the Data Integration Service will ignore the Hive engine selection and run the profile on the Blaze engine.

Column Profiles for Sqoop Data Sources

You can run a column profile on data objects that use Sqoop. You can select the Hive or Hadoop run-time environment to run the column profiles.

On the Hive engine, to run a column profile on a relational data object that uses Sqoop, you must set the Sqoop argument **m** to 1 in the JDBC connection. Use the following syntax:

```
-m 1
```

When you run a column profile on a logical data object or customized data object, you can configure the num-mappers argument to achieve parallelism and optimize performance. You must also configure the split-by argument to specify the column based on which Sqoop must split the work units.

Use the following syntax:

```
--split-by <column_name>
```

If the primary key does not have an even distribution of values between the minimum and maximum range, you can configure the split-by argument to specify another column that has a balanced distribution of data to split the work units.

If you do not define the split-by column, Sqoop splits work units based on the following criteria:

- If the data object contains a single primary key, Sqoop uses the primary key as the split-by column.
- If the data object contains a composite primary key, Sqoop defaults to the behavior of handling composite primary keys without the split-by argument. See the Sqoop documentation for more information.

- If a data object contains two tables with an identical column, you must define the split-by column with a table-qualified name. For example, if the table name is CUSTOMER and the column name is FULL_NAME, define the split-by column as follows:
  `--split-by CUSTOMER.FULL_NAME`
- If the data object does not contain a primary key, the value of the m argument and num-mappers argument default to 1.

When you use Cloudera Connector Powered by Teradata or Hortonworks Connector for Teradata and the Teradata table does not contain a primary key, the split-by argument is required.

Primary Key Discovery

Primary key discovery generates primary key candidates from the columns you specify.

A primary key is a column or combination of columns that uniquely identify a row in a data source. Primary key discovery identifies the columns and combinations of columns that meet a specific confidence level. You can edit the confidence level, as well as the maximum number of columns to combine for primary key identification.

Primary key discovery can highlight potential data quality issues by identifying the non-unique rows in a primary key candidate. This is especially useful in cases where primary key discovery combines many columns, since non-conforming records are likely to contain duplicate information.

Primary Key Inference Properties

When you create a single data object profile, you can use the **Primary Key Profiling** view to configure the primary key inference properties.

The following table describes the primary key inference properties in the **Primary Key Profiling** view:

Property	Description
Override the default inference options	Allows you to configure custom settings for primary key inference.
Max Key Columns	Maximum number of columns that can make up a primary key.
Max Rows	Number of rows to profile.
Conformance Criteria	Minimum percentage or maximum number of rows of key violations that the profile allows when determining primary keys.
Exclude data objects with documented, user defined key	Excludes data objects with documented primary keys or user-defined primary keys.
Excludes data objects with approved key	Excludes data objects with approved primary keys.

Inferred Primary Key Properties

After you run a single data object profile, you can use the **Primary Key Profiling** view to view the details of the inferred primary keys in the data source.

The following table describes the inferred primary keys properties in the **Primary Key Profiling** view:

Property	Description
Column	Name of the column in the profile.
% Conforming	Percentage of unique values in the column.
% Duplicates	Percentage of duplicate values for the column.
% Null	Percentage of null values for the column.
Verified	Determines whether the column is a primary key column.
Inference Status	Inference status of the column.
Last Run Time	The date and time that the primary key profile last ran.

Key Violations Properties

After you run a single data object profile, you can use the **Primary Key Profiling** view to view the details of the primary key violations in the data source.

The following table describes the key violations properties in the **Primary Key Profiling** view:

Property	Description
Column(s)	Name of the column(s) from which the profile infers a candidate primary key.
Number of Key Violations	Number of key violations in the primary key candidate.

Functional Dependency Discovery

Functional dependency discovery provides information about dependencies between pairs of columns in a data source.

A pair of columns are functionally dependent if the values in one column can reliably predict the values in another column. For example, if a dataset contains an Employer ID column and a date of birth column, the date of birth should be the same in all rows that contain a given Employer ID.

Functional dependencies can highlight potential data quality issues by identifying records the records that do not conform to a column functional dependency. For example, if 99.8% of rows in a data source are functionally dependent, there is a high likelihood that the remaining rows may contain incorrect information.

Functional Dependency Inference Properties

The **Functional Dependency Profiling** view provides information about the functional dependencies between columns.

The following table describes the functional dependency inference properties in the **Functional Dependency Profiling** view:

Property	Description
Override the default inference options	Allows you to configure custom settings for functional dependency inference.
Max Columns in Determinant	Number of columns that the profile can combine to find a determinant.
Max Rows	Number of rows to profile.
Dependencies Returned	Number of dependencies that the profile displays. Default is **Minimum Coverage**, which displays the smallest set of dependencies where each column appears at least one time in a dependent or determinant.
Max Dependencies Returned	Maximum number of dependencies that the profile displays.
Conformance Criteria	The minimum percentage or the maximum number of rows for dependency violations that the profile allows when determining functional dependents.

Inferred Functional Dependency Properties

After you run a single data object profile, you can use the **Functional Dependency Inference** view to view the details of inferred functional dependencies between columns in the data source.

The following table describes the inferred functional dependency properties in the **Functional Dependency Inference** view:

Property	Description
Determinant Columns	Name of the column analyzed for functional dependencies.
Dependent Columns	Name of the column dependent on the determinant column.
% Null	Percentage of null values for the column.
% Conforming	Percentage of functional dependency match.
Verified	Determines whether or not the columns are functionally dependent.
Last Run Time	The date and time that the functional dependency profile last ran.

Functional Dependency Violations Properties

The view provides information about the functional dependencies between columns. After you run a single data object profile, you can use the **Functional Dependency Inference** view to view the details of the functional dependency violations in the data source.

The following table describes the functional dependency violations properties in the **Functional Dependency Inference** view:

Property	Description
Determinant column	Name of the column analyzed for functional dependencies.
Distinct Dependents	Number of unique functional dependencies.

Operating System Profiles in Informatica Developer

You can choose an operating system profile in the Developer tool. After you choose an operating system profile, the Data Integration Service creates and runs the column profiles and enterprise discovery profiles and creates scorecards based on the permission of the operating system profile user.

Selecting an Operating System Profile

You can select an operating system profile in Informatica Developer. The Data Integration Service uses the permissions of the operating system profile user to run the profiling jobs.

1. In Informatica Developer, click **Windows** > **Preferences**.

   The **Preferences** dialog box appears.
2. Click **Informatica** > **Run Configurations** > **Mapping**.

   The **Mapping** dialog box appears.
3. In the **Mapping** dialog box, clear the **Use default Data Integration Service** option.
4. Click **Browse** to select an operating system profile in the list.
5. Click **OK**.

Creating a Single Data Object Profile in Informatica Developer

You can create a single data object profile for one or more columns in a data object and store the profile object in the Model repository.

1. In the **Object Explorer** view, select the data object you want to profile.
2. Click **File** > **New** > **Profile** to open the profile wizard.
3. Select **Profile** and click **Next**.

4. Enter a name for the profile and verify the project location. If required, browse to a new location.
5. Optionally, enter a text description of the profile.
6. Verify that the name of the data object you selected appears in the **Data Objects** section.
7. Click **Next**.
8. Configure the profile operations that you want to perform. You can configure the following operations:
   - Column profiling
   - Primary key discovery
   - Functional dependency discovery
   - Data domain discovery

   Note: To enable a profile operation, select **Enabled as part of the "Run Profile" action** for that operation. Column profiling is enabled by default.
9. Review the options for your profile.

   You can edit the column selection for all profile types. Review the filter and sampling options for column profiles. You can review the inference options for primary key, functional dependency, and data domain discovery. You can also review data domain selection for data domain discovery.
10. Review the drill-down options, and edit them if necessary. By default, the **Enable Row Drilldown** option is selected. You can edit drill-down options for column profiles. The options also determine whether drill-down operations read from the data source or from staged data, and whether the profile stores result data from previous profile runs.
11. In the **Run Settings** section, choose a run-time environment. Choose **Native**, **Hive (deprecated)**, or **Hadoop** as the run-time environment. When you choose the Hive or Hadoop option, select a Hadoop connection.
12. Click **Finish**.

Creating Multiple Data Object Profiles in Informatica Developer

When you run the multiple data object profile on multiple data objects, the Developer tool uses the default column profiling options to generate column profiles for one or more data objects. Optionally, you can create an enterprise discovery profile to run a profile on multiple data objects.

1. In the **Object Explorer** view, select the data objects you want to profile.
2. Click **File** > **New** > **Profile** to open the **New Profile** wizard.
3. In the **New** wizard, select the **Multiple Profiles** option, and click **Next**.
4. In the **Multiple Profiles** window, select the location where you want to create the profiles. You can create each profile at the same location as its profiled object, or you can specify a common location for the profiles.
5. Verify that the names of the data objects you selected appear within the **Data Objects** section.

   Optionally, click **Add** to add another data object.
6. Optionally, specify the number of rows to profile, and choose whether to run the profile when the wizard completes.
7. Click **Next**.

8. In the **Validation Environment** section, choose **Native**.

   Note: Choose only the Native option to run the multiple data object profile. To run multiple data objects on the Blaze engine in the Hadoop run-time environment, you can choose the enterprise discovery profile.
9. Click **Finish**.
10. Optionally, enter prefix and suffix strings to add to the profile names.
11. Click **OK**.

Editing a Profile

You can edit a single data object profile or multiple data object profile. If version control system is enabled, then the profile is checked out by default.

1. In the **Object Explorer** view, right-click the profile, and click **Open**.

   The **Results** view appears.
2. In the **Definition** view, update the properties as needed.
3. Click **Team** > **Check In** to check in the profile.
4. Right-click the profile, and click **Run Profile** to run the profile.

   The profile results appear in the **Results** view.

Synchronize Option

When you change the metadata of an external data source, the data object metadata in the Model Repository is not updated by default. Use the Synchronize option to synchronize the data object metadata to the data source metadata.

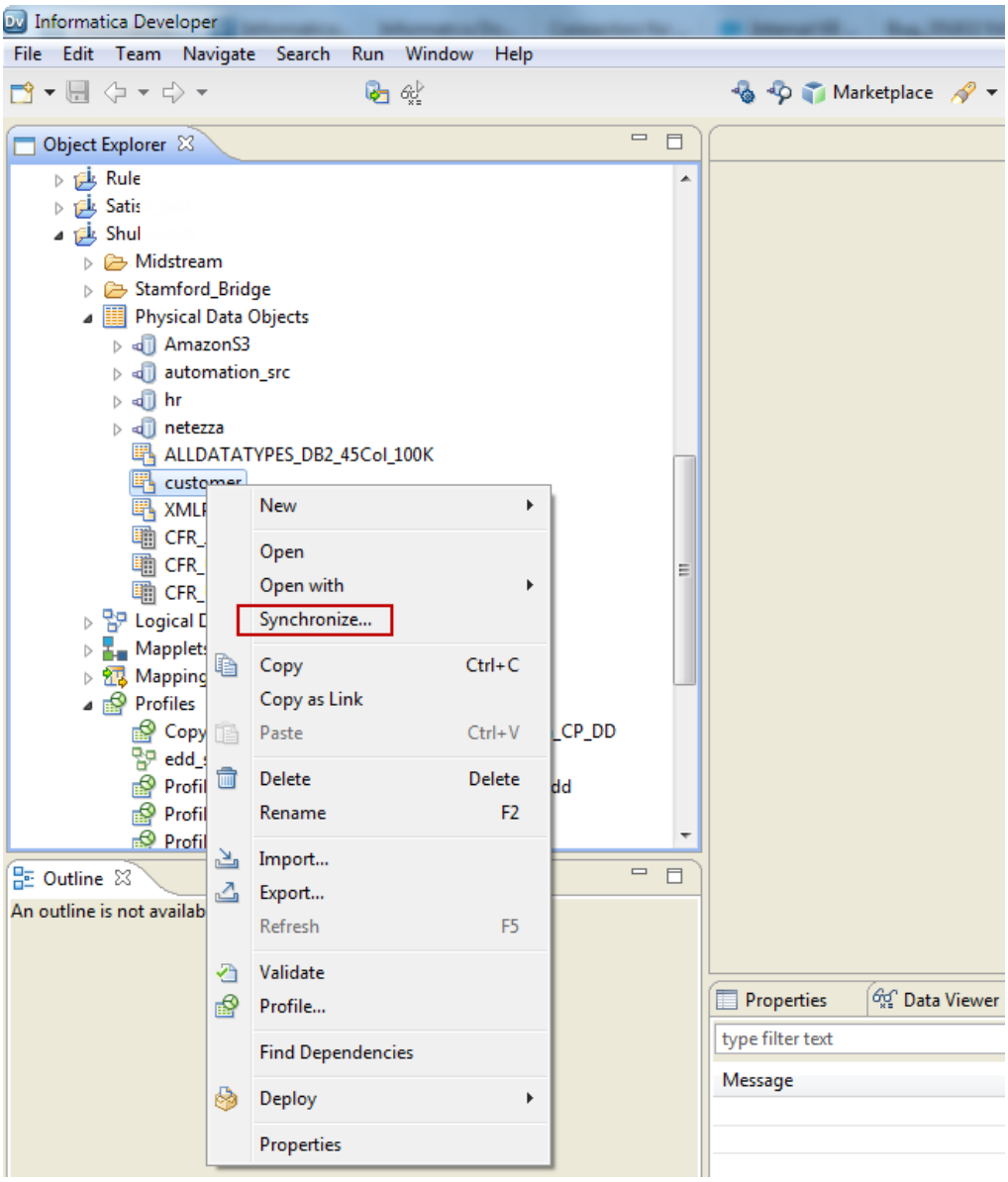
In the Developer tool, after you use the Synchronize option, when you open a profile or scorecard that uses the data object, an asterisk appears alongside the profile name or scorecard name in the editor. The asterisk indicates that the data object metadata has changed for the profile or scorecard. Open and save the profile or scorecard to update the profile definition in the Model Repository. Note that no visible changes appear in the Analyst tool when you open the profile or scorecard after you synchronize the data object for the profile or scorecard. You can use the Synchronize option for column profiles, enterprise discovery profiles, and scorecards. The external data source can be a relational data source or flat file data source.

Synchronizing a Flat File Data Object in Informatica Developer

You can synchronize the changes to an external flat file data source with its data object in Informatica Developer. Use the **Synchronize Flat File** wizard to synchronize the data objects.

1. In the **Object Explorer** view, select a flat file data object.
2. Right-click and select **Synchronize**.

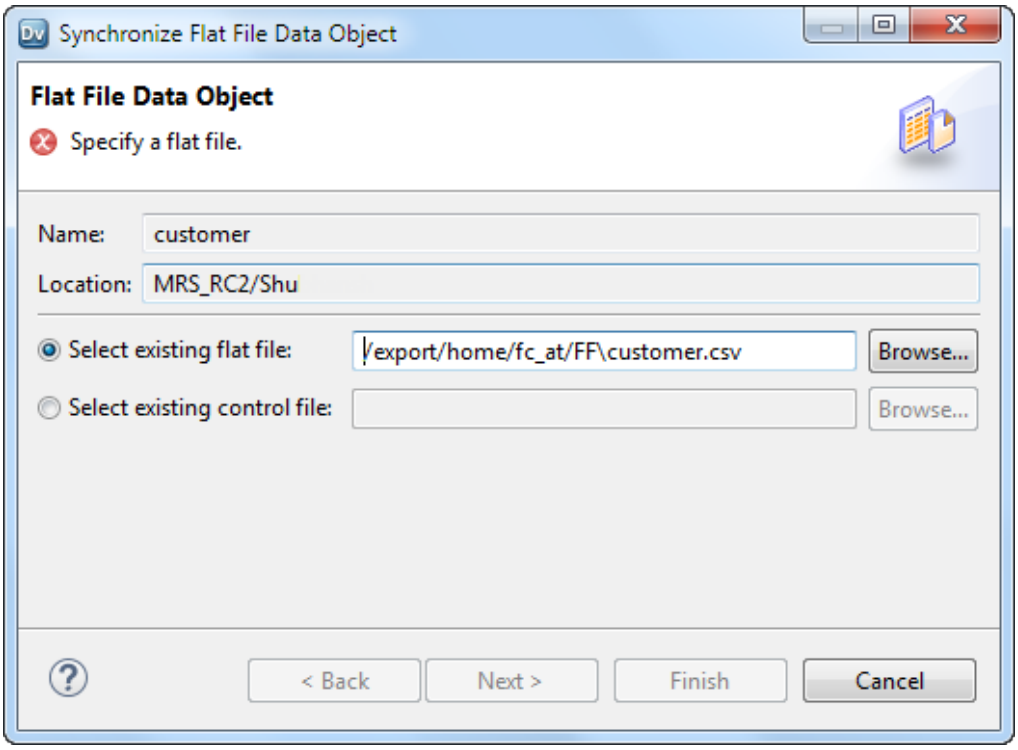
   The following image shows the Synchronize option for a data object:



The **Synchronize Flat File Data Object** wizard appears.

3. In the **Synchronize Flat File Data Object** wizard, verify the flat file path in the **Select existing flat file** field.

   The following image shows the Synchronize Flat File Data Object wizard:



4. Click **Next**.
5. Optionally, select the code page, format, delimited format properties, and column properties.
6. Click **Finish**, and then click **OK**.

Synchronizing a Relational Data Object in Informatica Developer

You can synchronize external data source changes of a relational data source with its data object in Informatica Developer. External data source changes include adding, changing, and removing columns, and changes to rules.

1. In the **Object Explorer** view, select a relational data object.
2. Right-click and select **Synchronize**.

   A message prompts you to confirm the action.
3. To complete the synchronization process, click **OK**.

   A synchronization process status message appears.
4. When you see a **Synchronization complete** message, click **OK**.

   The message displays a summary of the metadata changes made to the data object.

Comments

You can add a description as a comment to a profile. You can also add comments to the columns in the column profile results.

You can add multiple comments to columns. You can use the **Comments** view in the Developer tool to add and view the comments.

Adding Comments in Informatica Developer

Add comments to columns in column profile results. When you export the profile results, the Developer tool includes the comments.

1. In the **Object Explorer** view, open a profile.
2. Optionally, run the profile to update the profile results.
3. Select the **Comments** view.
4. Click **Add** to open the **Add Comment** dialog box.
5. From the list, select the profile name or one of the columns.

   If you previously added comments, you can view the comments in the dialog box.
6. In the **Comment** field, enter a description.
7. Click **OK**.

   The Developer tool displays the comment in the **Comments** view.
8. To delete a comment, select the comment in the **Comments** view, and click **Remove**.

CHAPTER 19

Column Profiles on Semi-structured Data Sources

This chapter includes the following topics:

Column Profiles on Semi-structured Data Sources Overview

You can create data objects from Avro, JSON, Parquet, and XML data sources and then create a column profile on the data objects.

Avro, JSON, Parquet, and XML formats are semi-structured data sources. To use the semi-structured data sources to create a column profile, you can perform the following tasks:

1. Create a physical data object on the semi-structured data source.
2. Create and run a column profile on the physical data object.

You can create flat file data objects for JSON or XML data sources. You can create complex file data objects for Avro, JSON, Parquet, and XML data sources in Hadoop Distributed File System (HDFS).

JSON and XML Data Objects

You can create a flat file data object or complex file data object from a JSON or XML data source. You can create and run a column profile on the data object.

Create a text file that contains the path of the JSON or XML data source and use the text file as the data source to create a flat file data object. You can also add the file path for multiple JSON or multiple XML data sources into the text file.

You can create a complex file data object from a JSON or XML data source with a complex file reader. The complex file reader provides input to a Data Processor transformation that parses the file and converts the source data to flat comma-separated values records.

Note: The Developer tool does not support a JSON data source with UTF-8 encoding.

Creating a Data Object from a JSON or XML Data Source

You can create a flat file data object or complex file data object from a JSON or XML data source.

1. In the **Object Explorer** view in the Developer tool, select the project where you want to create the data object and column profile.
2. Click **File** > **New** > **Data Object**.

   The **New** dialog box appears.
3. You can choose to create a flat file data object or complex file data object.
   - To create a flat file data object, perform the following tasks:
     1. Select **Physical Data Objects** > **Flat File Data Object**, and click **Next**.
        The **New Flat File Data Object** dialog box appears.
     2. Select **Create from an Existing Flat File**, and click **Browse** to choose the text file. Click **Next**.
     3. Verify that the code page is **MS Windows Latin 1 (ANSI), superset of Latin 1**, and the format is delimited. Click **Next**.
     4. Verify that the delimiter is set to **comma**. Click **Finish**.
   - To create a complex file data object, perform the following tasks:
     1. Select **Physical Data Objects** > **Complex File Data Object**, and click **Next**.
        The **New Complex File Data Object** dialog box appears.
     2. Enter a name for the data object. Select the access type as **File**.
     3. Click **Browse** to choose a JSON or XML file. Click **Finish**.
        When the Developer server is in Linux, you must update the file path of the data source to the location in the server. To update the file path, select the complex file data object, click **Read** in the **Data Object Operations** tab, and add the file path in the **Advanced** tab in the **Data Object Operation Details** pane.

   The data object appears in the project folder.

Complex File Data Objects for Semi-Structured Data Sources in HDFS

You can create and run a column profile on an Avro, JSON, Parquet, or XML file that uses HDFS. To read the JSON or XML file in HDFS, use a complex file reader to pass the JSON or XML input to the Data Processor transformation.

Complex File Data Object from a JSON or XML Data Source in HDFS

You can create a complex file data object from a JSON or XML file. You can create and run a column profile on data object.

Create a connection to HDFS before you create the data objects for JSON or XML files in HDFS.

You can use one of the following methods to create a data object from a JSON or XML file in HDFS:

- Create a complex file data object on a JSON or XML file.
- Create a complex file data object on a folder that contains multiple JSON or multiple XML files.

After you create the data object, you can create and run a column profile on the data object.

Complex File Data Object from an Avro or Parquet Data Source in HDFS

You can create a complex file data object from an Avro or Parquet data source in HDFS. You can use the data object to create and run a column profile.

You can create a complex file data object from an Avro or Parquet file or on a folder that contains multiple Avro or multiple Parquet files. You can create a complex file data object from an Avro and Parquet data source with file or connection access type and resource format as Binary, Avro, or Parquet. You have to create an HDFS connection before you create a complex file data object from the Avro and Parquet data sources.

Note: You can choose the Resource Format as **Avro** or **Parquet** only for flat structured Avro and Parquet data sources.

You can choose one of the following options when you create a data object from Avro and Parquet files in HDFS:

- Select the access type as file and resource format as Binary.
- Select the access type as file and resource format as Avro or Parquet.
- Select the access type as connection and resource format as Avro or Parquet.

Creating an HDFS Connection

Configure the HDFS connection in Informatica Developer to create a column profile on an Avro, JSON, Parquet, and XML data sources in HDFS. You can create a complex file data object after you create an HDFS connection.

1. Click **Window** > **Preferences**.
2. Select **Informatica** > **Connections**.
3. Expand the domain.
4. Select the connection type **File Systems** > **Hadoop File System**, and click **Add**.
5. Enter a connection name.
6. Optionally, enter a connection description.
7. Click **Next**.
8. Configure the connection properties.
9. Click **Test Connection** to verify the connection to HDFS.
10. Click **Finish**.

Creating a Complex File Data Object from a JSON or XML File in HDFS

You can create a complex file data object from a JSON or XML source file that uses HDFS, and create a column profile on the data object.

1. In the **Object Explorer** view in the Developer tool, select the project where you want to create the physical data object and column profile.
2. Click **File** > **New** > **Data Object**.

   The **New** dialog box appears.
3. Select **Physical Data Objects** > **Complex File Data Object**, and click **Next**.

   The **New Complex File Data Object** dialog box appears.
4. Enter a name for the data object. Select the access type as **Connection** .
5. You can create a data object from a JSON or XML file or on a folder that contains multiple JSON or multiple XML files.
   - To create a complex file data object from a JSON or XML file, perform the following steps:
     1. Click **Browse** to select a connection.
     2. In the **Add Resource** dialog box, click **Add** to choose a JSON or XML file.
     3. Click **Finish**.
        The data object appears in the project folder.
   - To create a complex file data object on a folder with multiple JSON or multiple XML files, perform the following steps:
     1. Click **Browse** to select a connection.
     2. In the **Add Resource** dialog box, click **Add** to choose a JSON or XML file in the folder.

3. Click **Finish**.
   The data object appears in the project folder.
4. Select the data object in the project folder and click **Advanced** > **Runtime: Read** > **Source file directory**.
5. Remove the file name and retain the folder name in the file path.

Creating a Complex File Data Object from an Avro or Parquet Data Source

You can create a complex file data object from an Avro or Parquet data source with **File** or **Connection** as the access type. You can create a column profile on the data object.

1. In the **Object Explorer** view, select a project.
2. Click **File** > **New** > **Data Object**.

   The **New** dialog box appears.
3. Select **Physical Data Objects** > **Complex File Data Object** and click **Next**.

   The **New Complex File Data Object** dialog box appears.
4. Enter a name for the data object.
5. You can choose the access type as **Connection** or **File**.
   - If you choose the Access Type as **Connection**, perform the following steps:
     1. Click **Browse** to choose an HDFS connection.
     2. In the **Choose Connection** dialog box, choose a data source, and click **OK**.
     3. In the **New Complex File Data Object** dialog box, click **Finish**.
        The data object appears in the project folder.
   - If you choose the Access Type as **File** and the Resource Format as **Binary**, perform the following steps:
     1. Click **Browse** to choose an Avro or Parquet file on the local machine.
     2. In the **New Complex File Data Object** dialog box, click **Finish**.
        The data object appears in the project folder.
     3. Select the data object in the project folder and click the **Data Object Operations** view.
     4. In the **Data Object Operations** view, click **Read** > **Advanced** tab.
     5. In the **Advanced** tab, enter the file path of the data source on the Linux or Windows machine in the **File path** field.
     6. Enter the File Format as **Custom Input**.
     7. Enter **com.informatica.avro.AvroToXML** in the **Input Format** field for Avro data sources, and enter **com.informatica.parquet.ParquetToXML** in the **Input Format** field for Parquet data sources. When you add the input format, the Data Processor Transformation processes and transforms the data sources in Avro or Parquet format to a data source in XML format at runtime.

- If you choose the Access Type as **File** and the Resource Format as **Avro** or **Parquet**, perform the following steps:
  1. Click **Browse** to choose an Avro or Parquet file in the local machine.
  2. In the **New Complex File Data Object** dialog box, click **Finish**.
     The data object appears in the project folder.
  3. After you create the data object, navigate to **Data Object Operations** > **Read** > **Advanced** tab, and verify whether the file path in the **File path** field corresponds to the data source in the Linux or Windows machine.

  Note: You can choose the Resource Format as **Avro** or **Parquet** only for flat-structured Avro and Parquet data sources.

  You can choose a folder with multiple Avro or multiple Parquet files to create a data object. After you create the data object, navigate to **Data Object Operations** > **Read** > **Advanced** tab, and verify whether the file path in the **File path** field points to the folder of the data sources in the Linux or Windows machine.

Creating a Column Profile on a Semi-structured Data Source

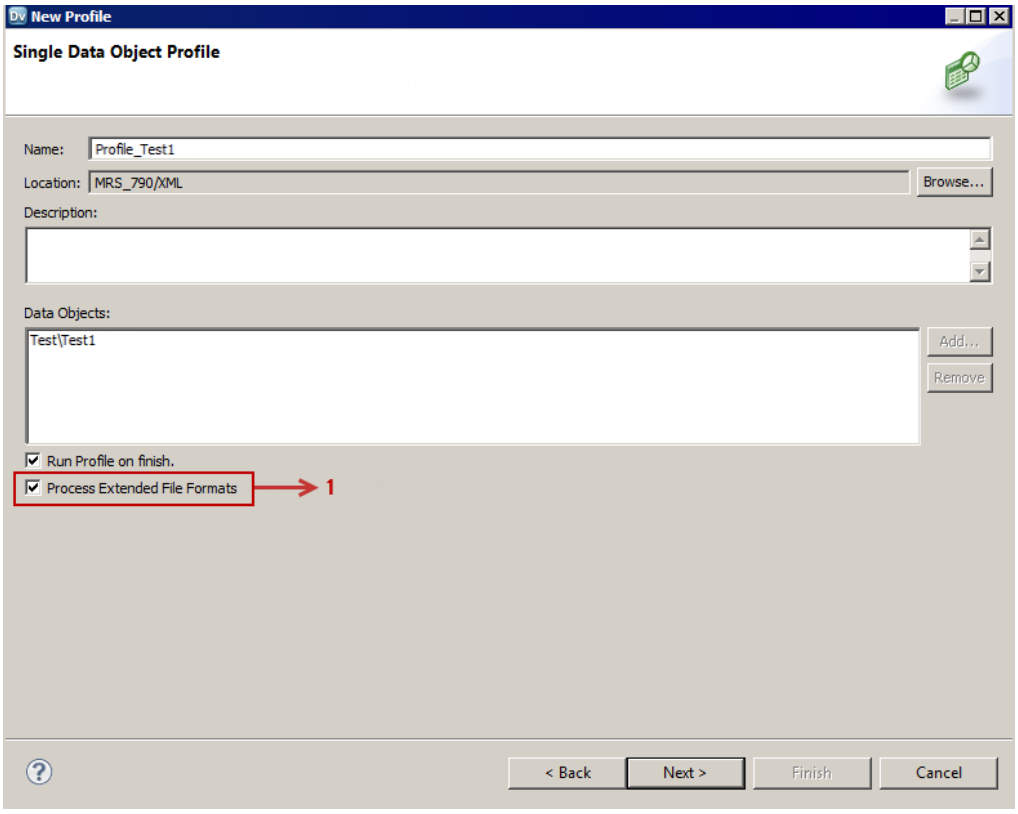
After you create a flat file data object or complex file data object from Avro, JSON, Parquet, or XML data sources, you can create and run a column profile on the data object.

1. In the **Object Explorer** view, select the data object for the Avro, JSON, Parquet, or XML file.
2. Click **File** > **New** > **Profile**.

   The **New** dialog box appears.
3. Select **Profile**. Click **Next**.

   The **New Profile** dialog box appears.
4. In the **New Profile** dialog box, add a name for the profile and an optional description.
5. Select **Process Extended File Formats** option. Click **Next**.

   The following image shows the **New Profile** wizard with the **Process Extended File Formats** option:



1. Process Extended File Formats. Select this option to process semi-structured data sources.

Note: The **Process Extended File Formats** option does not appear for Avro and Parquet data sources when you choose the Resource Format as **Avro** or **Parquet**.

6. In the **Single Data Object Profile** page, select the columns and options under **Column Selection** and **Data Domain Discovery** as required. Click **Finish**.

   Note: If the Developer tool is installed on a Linux machine and the JSON or XML physical data object is a flat file data object with a text file, then perform the following tasks:

   1. On the **Overview** tab, update the **Precision** value to include the number of characters in the file path of the data source in the server.
   2. Update the file path of the data source to the location in the server after you create a profile on the flat file data object. To update the file path, click **Runtime: Read** > **Source file directory** in the **Advanced** tab, and add the file path.

7. Right-click the profile, and select **Run Profile**.

   The profile results appear.

CHAPTER 20

Rules in Informatica Developer

This chapter includes the following topics:

Rules in Informatica Developer OverviewGuidelines for Rules

A rule is business logic that defines conditions applied to source data when you run a column profile. You can add a rule to the profile to validate data. You can use mapplets that are validated as rules, predefined rules, or reusable rules in the column profiles.

You can use the following methods to use rules in the column profiles:

- In the Developer tool, create a mapplet and validate it as a rule. The rule appears as a reusable rule in the Analyst tool. You can apply the rule to column profiles in the Analyst tool and Developer tool.
- You can use predefined rules in the column profiles. Informatica provides the predefined rules with the Developer tool and Analyst tool.
- In the Analyst tool, create a rule specification and generate a mapplet. You can apply the rule specification to column profiles in the Analyst tool. In the Developer tool, validate the mapplet as a rule. The rule appears as a reusable rule that you can use in the column profiles.

  Note: In the Developer tool, you cannot add, edit, or delete rule specifications in a column profile.

A rule must meet the following requirements:

- It must contain an Input and Output transformation. You cannot use data sources in a rule.
- It can contain Expression transformations, Lookup transformations, and passive data quality transformations. It cannot contain any other type of transformation. For example, a rule cannot contain a Match transformation as it is an active transformation.
- It does not specify cardinality between input groups.

A rule must meet the following requirements:

- It must contain an Input and Output transformation. You cannot use data sources in a rule.

- It can contain Expression transformations, Lookup transformations, and passive transformations. It cannot contain any other type of transformation. For example, a rule cannot contain a Sorter transformation as it is an active transformation.
- It does not specify cardinality between input groups.

Creating a Rule in Informatica Developer

You need to validate a mapplet as a rule to create a rule in the Developer tool.

Create a mapplet in the Developer tool.

1. Right-click the mapplet editor.
2. Select **Validate As** > **Rule**.

Applying a Rule in Informatica DeveloperApplying a Rule

You can add a rule to a saved column profile. You cannot add a rule to a profile configured for join analysis. You can add a rule to a saved column profile.

1. Browse the **Object Explorer** view and find the profile you need.
2. Right-click the profile and select **Open**.

   The profile opens in the editor.
3. Click the **Definition** tab, and select Rules.
4. Click **Add**.

   The **Apply Rule** dialog box opens.
5. Click **Browse** to find the rule you want to apply.

   Select a rule from a repository project, and click **OK**.
6. Click the **Value** column under **Input Values** to select an input port for the rule.
7. Optionally, click the **Value** column under **Output Values** to edit the name of the rule output port.

   The rule appears in the **Definition** tab.

CHAPTER 21

Mapplet and Mapping Profiling

This chapter includes the following topics:

Mapplet and Mapping Profiling OverviewMapplet and Mapping Profiles

You can define a column profile for an object in a mapplet or mapping. Run a profile on a mapplet or a mapping object when you want to verify the design of the mapping or mapplet without saving the profile results. You can also generate a mapping from a profile.

Running a Profile on a Mapplet or Mapping Object

When you run a profile on a mapplet or mapping object, the profile runs on all data columns and enables drill-down operations on the data that is staged for the data object. You can run a profile on a mapplet or mapping object with multiple output ports.When you run a profile on a mapplet or mapping object, the profile runs on all data columns and enables drill-down operations on the data. You can run a profile on a mapplet or mapping object with multiple output ports.

The profile traces the source data through the mapping to the output ports of the object you selected. The profile analyzes the data that would appear on those ports if you ran the mapping.

1. Open a mapplet or mapping.
2. Verify that the mapplet or mapping is valid.
3. Right-click a data object or transformation and select **Profile Now**.

   If the transformation has multiple output groups, the **Select Output Group** dialog box appears. If the transformation has a single output group, the profile results appear on the **Results** tab of the profile.
4. If the transformation has multiple output groups, select the output groups as necessary.

5. Click **OK**.

   The profile results appears in the **Results** tab of the profile.

Comparing Profiles for Mapping or Mapplet Objects

You can create a profile that analyzes two objects in a mapplet or mapping and you can compare the results of the column profiles for those objects.

Like column profiles of single mapping or mapplet objects, profile comparisons run on all data columns and enable drill-down operations on the data that is staged for the data objects. After you move data from a source table to a target table, you can compare profiles to verify the migration of data. You can also compare profiles on a data source that changes over time.

Like profiles of single mapping or mapplet objects, profile comparisons run on all data columns.

1. Open a mapplet or mapping.
2. Verify that the mapplet or mapping is valid.
3. Press the **CTRL** key and click two objects in the editor.
4. Right-click one of the objects and select **Compare Profiles**.
5. Optionally, configure the profile comparison to match columns from one object to the other object.
6. Optionally, match columns by clicking a column in one object and dragging it onto a column in the other object.
7. Optionally, choose whether the profile analyzes all columns or matched columns only.
8. Click **OK**.

Generating a Mapping from a Profile

You can create a mapping object from a profile. Use the mapping object you create to develop a valid mapping. The mapping you create has a data source based on the profiled object and can contain transformations based on profile rule logic. After you create the mapping, add objects to complete it.

1. In the **Object Explorer** view, find the profile on which to create the mapping.
2. Right-click the profile name and select **Generate Mapping**.

   The **Generate Mapping** dialog box displays.
3. Enter a mapping name. Optionally, enter a description for the mapping.
4. Confirm the folder location for the mapping.

   By default, the Developer tool creates the mapping in the **Mappings** folder in the same project as the profile. Click **Browse** to select a different location for the mapping.
5. Confirm the profile definition that the Developer tool uses to create the mapping. To use another profile, click **Select Profile**.
6. Click **Finish**.

   The mapping appears in the **Object Explorer**.

Add objects to the mapping to complete it.

CHAPTER 22

Column Profile Results in Informatica Developer

This chapter includes the following topics:

Column Profile Results in Informatica DeveloperColumn Profile Results

Column profile analysis provides information about data quality by highlighting value frequencies, patterns, and statistics of data.

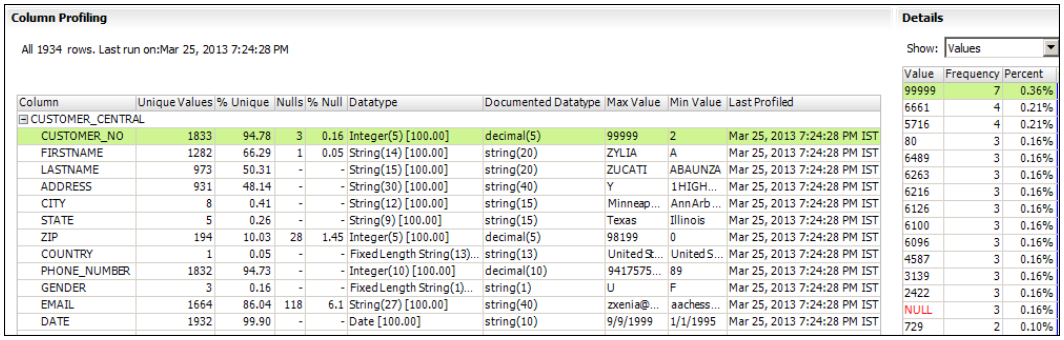
The following table describes the profile results for each type of analysis:

Column profiling analysis generates the following profile results:

Profile Type	Profile Results
Column profile	- Percentage and count statistics for unique and null values - Inferred data types - The data type that the data source declares for the data - The maximum and minimum values - The date and time of the most recent profile run - Percentage and count statistics for each unique data element in a column - Percentage and count statistics for each unique character pattern in a column
Primary key profile	- Inferred primary keys - Key violations
Functional dependency profile	- Inferred functional dependencies - Functional dependency violations

- Percentage and count statistics for unique and null values
- Inferred data types
- The data type that the data source declares for the data
- The maximum and minimum values
- The date and time of the most recent profile run
- Percentage and count statistics for each unique data element in a column
- Percentage and count statistics for each unique character pattern in a column

The following image shows the column profile results:



Column Profiling

All 1934 rows. Last run on:Mar 25, 2013 7:24:28 PM

Column	Unique Values	% Unique	Nulls	% Null	Datatype	Documented Datatype	Max Value	Min Value	Last Profiled
CUSTOMER_CENTRAL									
CUSTOMER_NO	1833	94.78	3	0.16	Integer(5) [100.00]	decimal(5)	99999	2	Mar 25, 2013 7:24:28 PM IST
FIRSTNAME	1282	66.29	1	0.05	String(14) [100.00]	string(20)	ZYLIA	A	Mar 25, 2013 7:24:28 PM IST
LASTNAME	973	50.31	-	-	String(15) [100.00]	string(20)	ZUCATI	ABAUNZA	Mar 25, 2013 7:24:28 PM IST
ADDRESS	931	48.14	-	-	String(30) [100.00]	string(40)	Y	1HIGH...	Mar 25, 2013 7:24:28 PM IST
CITY	8	0.41	-	-	String(12) [100.00]	string(15)	Minneap...	Ann Arb...	Mar 25, 2013 7:24:28 PM IST
STATE	5	0.26	-	-	String(9) [100.00]	string(15)	Texas	Illinois	Mar 25, 2013 7:24:28 PM IST
ZIP	194	10.03	28	1.45	Integer(5) [100.00]	decimal(5)	98199	0	Mar 25, 2013 7:24:28 PM IST
COUNTRY	1	0.05	-	-	Fixed Length String(13)...	string(13)	United St...	United S...	Mar 25, 2013 7:24:28 PM IST
PHONE_NUMBER	1832	94.73	-	-	Integer(10) [100.00]	decimal(10)	9417575...	89	Mar 25, 2013 7:24:28 PM IST
GENDER	3	0.16	-	-	Fixed Length String(1)...	string(1)	U	F	Mar 25, 2013 7:24:28 PM IST
EMAIL	1664	86.04	118	6.1	String(27) [100.00]	string(40)	zxenia@...	aachess...	Mar 25, 2013 7:24:28 PM IST
DATE	1932	99.90	-	-	Date [100.00]	string(10)	9/9/1999	1/1/1995	Mar 25, 2013 7:24:28 PM IST

Details

Show: Values

Value	Frequency	Percent
99999	7	0.36%
6661	4	0.21%
5716	4	0.21%
80	3	0.16%
6489	3	0.16%
6263	3	0.16%
6216	3	0.16%
6126	3	0.16%
6100	3	0.16%
6096	3	0.16%
4587	3	0.16%
3139	3	0.16%
2422	3	0.16%
NULL	3	0.16%
729	2	0.10%

Column Value Properties

Column value properties show the values in the profiled columns and the frequency with which each value appears in each column. The frequencies are shown as a number, a percentage, and a bar chart.

To view column value properties, select Values from the **Show** list. Double-click a column value to drill-down to the rows that contain the value.

The following table describes the properties for column values:

Property	Description
Values	List of all values for the column in the profile.
Frequency	Number of times a value appears in a column.
Percent	Number of times a value appears in a column, expressed as a percentage of all values in the column.
Chart	Bar chart for the percentage.

Column Pattern Properties

Column pattern properties show the patterns of data in the profiled columns and the frequency with which the patterns appear in each column. The patterns are shown as a number, a percentage, and a bar chart.

To view pattern information, select Patterns from the **Show** list. Double-click a pattern to drill-down to the rows that contain the pattern.

The following table describes the properties for column value patterns:

Property	Description
Patterns	Pattern for the selected column.
Frequency	Number of times a pattern appears in a column.
Percent	Number of times a pattern appears in a column, expressed as a percentage of all values in the column.
Chart	Bar chart for the percentage.

Column Statistics Properties

Column statistics include properties, such as maximum and minimum lengths of values and first and last values.

To view statistical information, select Statistics from the **Show** list.

The following table describes the column statistics properties:

Property	Description
Maximum Length	Length of the longest value in the column.
Minimum Length	Length of the shortest value in the column.

Property	Description
Bottom	Last five values in the column.
Top	First five values in the column.
Sum	Sum of all values in the column with a numeric data type.

Note: The profile also displays average and standard deviation statistics for columns of type Integer.

Column Data Type Properties

Column data types include all the inferred data types for each column in the profile results.

To view data type information, select **Data types** from the **Show** list. Double-click a data type to drill-down to the rows that contain the data type.

The following table describes the properties for the column data types:

Property	Description
Data type	List of all the inferred data types for the column in the profile.
Frequency	Number of times a data type appears for a column, expressed as a number.
% Conformance	Percentage that a data type appears for a column.
Status	Indicates the status of the data type. The statuses are Inferred, Approved, or Rejected. **Inferred** Indicates the data type of the column that the Developer tool inferred. **Approved** Indicates an approved data type for the column. When you approve a data type, you commit the data type to the Model repository. **Rejected** Indicates a rejected data type for the column.

Curation in Informatica DeveloperCuration in Informatica Developer

Curation is the process of validating and managing discovered metadata of a data source so that the metadata is fit for use and reporting. When you curate metadata in the Informatica Developer, you can approve, reject, and reset the inferred data types or data domains in profile results.

You can approve one data type or data domain for a column. You can hide the rejected data types or data domains for a column. After you approve or reject an inferred data type or data domain, you can reset the data type or data domain to restore the inferred status.

Approving Data typesApproving Data types in Informatica Developer

The profile results include the inferred data types, frequency, percentage of conformance, and the inference status for each column in the data source. You can choose and approve a single data type for each column.

1. In the **Object Explorer** view, select and open a profile.
2. Verify that you are in the **Results** tab.
3. In the **Column Profiling** view, select a column to view the value frequencies, patterns, data types, and statistics in the right panel.
4. Under the **Details** panel, select **Data types** from the **Show** list.

   The inferred data types for the column appear.
5. Right-click the column that you want to approve and click **Approve**.

   The status of the data type changes to **Approved**.
6. To restore the inferred status of the data type, right-click the data type and click **Reset**.

Rejecting Data TypesRejecting Data Types in Informatica Developer

Informatica Developer displays inferred data types in the profile results by default. You can reject inferred or approved data types. You can choose to show or hide the rejected data types.

1. In the **Object Explorer** view, select a profile.
2. Double-click the profile to open it.

   The profile opens in a tab.
3. In the **Column Profiling** view, select a row.
4. To reject inferred column data types, select the **Data types** view in the right panel. Select the inferred data type that you want to reject, right-click the row, and select **Reject**.

   Informatica Developer greys out the rejected data type in the list of data types.
5. To hide the rejected data types, right-click the row and select **Hide Rejected**.
6. To view the rejected data types, right-click one of the rows, and select **Show Rejected**.

Exporting Profile Results from Informatica Developer

You can export column profile results to a .csv file or Microsoft Excel file. When you export the profile results to a Microsoft Excel file, the Developer tool saves the information to an .xlsx file.

1. In the **Object Explorer** view, open a profile.
2. Optionally, run the profile to update the profile results.
3. Select the **Results** view.
4. Select a column.
5. Under **Details**, select **Values**, **Patterns**, or **Data types** and click the **Export** icon.

   The **Export data to a file** dialog box opens.
6. Accept or change the default file name.
7. Select the type of data to export. You can select **Values for the selected column**, **Patterns for the selected column**, **Data types for the selected column**, or **All (Summary, Values, Patterns, Data types, Statistics, Properties)**.
8. Click **Browse** to select a location and save the file locally in your computer.
9. If you do not want to export field names as the first row, clear the **Export field names as first row** check box.
10. Click **OK**.

CHAPTER 23

Scorecards in Informatica Developer

This chapter includes the following topics:

Scorecards in Informatica Developer Overview

A scorecard is a graphical representation of the quality measurements in a profile. You can view scorecards in the Developer tool. After you create a scorecard in the Developer tool, you can connect to the Analyst tool to open the scorecard for editing. Run the scorecard on current data in the data object or on data staged in the profiling warehouse.

You can edit a scorecard, run the scorecard, and view the scorecard lineage for a metric or metric group in the Analyst tool.

Creating a Scorecard

Create a scorecard and add columns from a profile to the scorecard. You must run a profile before you add columns to the scorecard.

1. In the **Object Explorer** view, select the project or folder where you want to create the scorecard.
2. Click **File** > **New** > **Scorecard**.

   The **New Scorecard** dialog box appears.
3. Click **Add**.

   The **Select Profile** dialog box appears. Select the profile that contains the columns you want to add.
4. Click **OK**, then click **Next**.
5. Select the columns that you want to add to the scorecard.

By default, the scorecard wizard selects the columns and rules defined in the profile. You cannot add columns that are not included in the profile.

6. Click **Finish**.

   The Developer tool creates the scorecard.

7. Optionally, click **Open with Informatica Analyst** to connect to the Analyst tool and open the scorecard in the Analyst tool.

Exporting a Resource File for Scorecard Lineage

You can export a project containing scorecards and dependent objects as a resource file for Metadata Manager. Use the exported resource file in the XML format to create and load a resource for scorecard lineage in Metadata Manager.

1. To open the **Export** wizard, click **File** > **Export**.
2. Select **Informatica** > **Resource File for Metadata Manager**.
3. Click **Next**.
4. Click **Browse** to select a project that contains the scorecard objects and lineage that you need to export.
5. Click **Next**.
6. Select the scorecard objects that you want to export.
7. Enter the export file name and file location.
8. To view the dependent objects that the **Export** wizard exports with the objects that you selected, click **Next**.

   The **Export** wizard displays the dependent objects.

9. Click **Finish**.

   The Developer tool exports the objects to the XML file.

Viewing Scorecard Lineage from Informatica Developer

To view the scorecard lineage for a metric or metric group from the Developer tool, launch the Analyst tool.

1. In the **Object Explorer** view, select the project or folder that contains the scorecard.
2. Double-click the scorecard to open it.

   The scorecard appears in a tab.

3. Click **Open with Informatica Analyst**.

   The Analyst tool opens in the browser window.

4. In the **Scorecard** view of the Analyst tool, select a metric or metric group.
5. Right-click and select **Show Lineage**.

   The scorecard lineage diagram appears in a dialog box.

CHAPTER 24

Data Domain Discovery in Informatica Developer

This chapter includes the following topics:

Data Domain Discovery in Informatica Developer Overview

Use the data domain glossary to manage data domains. To create a data domain, you can use predefined data rules and column name rules. You can also generate the data domain based on specific values or patterns in the column profile results.

You can select the source columns, data domains with which you want to match the column data and column name, sampling options, drill-down options, and run-time environment. You can choose the maximum number of rows you want to perform data domain discovery on. You can choose a conformance criteria for data domain discovery. You can exclude null values from data domain discovery. After you run a profile, you can verify, curate, and drill down on the results. You can also add the results to a data model from the editor in the Developer tool.

You can create a profile with a sampling option and filters to perform data domain discovery. When you run the profile, you apply the sampling option and filters on the data source and generate a data set. The data domain discovery process uses the data set to discover data domains.

Data Domain Glossary in Informatica Developer

You manage data domains and data domain groups in the data domain glossary. You can add, edit, and remove data domains and data domain groups. You can also search for specific data domains and data domain groups.

You can export data domains from the data domain glossary to an XML file. You can also import data domains from an XML file to the data domain glossary. You create a data domain group to organize data domains into specific groups such as Personal Health Information (PHI), Personally Identifiable Information (PII), or any other conceptual group that is relevant to the project. You can have a data domain in multiple data domain groups. For example, phone number can belong to both PII and PHI data domain groups.

Creating a Data Domain Group in Informatica Developer

You can add data domains to data domain groups for effective column data analysis.

1. Click **Window** > **Preferences**.

   The **Preferences** dialog box appears.
2. In the **Preferences** dialog box, select **Informatica** > **Data Domain Glossary**.

   The Developer tool displays a list of all data domains in the **Data Domain Glossary** panel.
3. In the **Show** field, select **Data Domain Groups**.

   The **Data Domain Glossary** panel rearranges the data domain list based on data domain groups.
4. In the **Data Domain Glossary** panel, select **Data Domain Groups**.
5. Click **Add**.

   The **Data Domain Group** dialog box appears.
6. Enter a name and description.
7. Click **Next**.
8. Click **Choose** to open the **Select Data Domains** dialog box.
9. Select the data domains you want to add to the data domain group and click **OK**.

   The Developer tool lists the selected data domains in the **Select Data Domains** panel.
10. Click **Finish**.

    The Developer tool adds the data domain group to the data domain glossary.

Creating a Data Domain in Informatica Developer

You can create data domains and add them to the data domain glossary. You can also add data domains to one or more data domain groups.

1. Click **Window** > **Preferences**.

   The **Preferences** dialog box appears.
2. In the **Preferences** dialog box, select **Informatica** > **Data Domain Glossary**.

   The Developer tool displays a list of all data domains in the **Data Domain Glossary** panel.
3. In the **Data Domain Glossary** panel, select **Data Domains**.
4. Click **Add**.

   The **Data Domain** dialog box appears.

5. Enter a name and description.
6. Click **Use data rule** to discover data domains based on column data. You can also select **Use column name rule** to discover data domains based on column names in the data source.

   The **Browse** button is enabled.
7. Click **Browse** to open the **Select Location** dialog box.
8. Select the appropriate rules and click **OK**.

   When you create a data domain, the Developer tool copies rules and other dependent objects associated with the data domain to the data domain glossary. To edit a rule associated with a data domain, you must go to the original rule and make changes to it. You can then associate the modified rule to the data domain again.

   The rules that you selected appear in the **Data rule** and **Column name rule** fields.
9. Click **Next**.
10. Click **Choose** to open the **Select Data Domain Groups** dialog box.
11. Select the data domain groups you want to include the data domain in and click **OK**.

    The Developer tool displays the selected data domain groups in the **Assign to Data Domain Groups** pane.
12. Click **Finish**.

    The Developer tool adds the data domain to the data domain glossary.

Creating a Data Domain from Profile Results in Informatica Developer

After you run a column profile, you can view the values and patterns of source data. You can then create a data domain from them.

1. Run a column profile to view its results.
2. Select the values or patterns based on which you want to create a data domain.

   The values, patterns, and statistics appear in the **Results** view.
3. Right-click the values or patterns, and then select **Send to** > **New Data Domain**.

   The **Data Domain** dialog box appears.
4. Enter the data domain name and an optional description.

   The location is set to data domain glossary by default.
5. Click **Finish**.

   The data domain gets added to the data domain glossary.

Find Data Domains in Informatica Developer

The data domain glossary displays all the data domains by default. You can search for specific data domains and data domain groups.

The **Data Domain Glossary** pane in the **Preferences** dialog box displays all the data domains and data domain groups. You can search for and view more information on data domains and data domain groups in the following ways:

Search for data domains and data domain groups.

Type in part of the data domain name or data domain group name in the field on top of the **Data Domain Glossary** panel. If you are in the **Data Domain Groups View**, the Developer tool lists the data domains that contain the search string in their names and the data domain groups associated with them. If you

are in the **Data Domain View**, the Developer tool lists all the data domains that contain the search string in their names.

View data domain groups and data domains within them.

In the **Show** field, select **Data Domain Groups**.

View all data domains.

In the **Show** field, select **Data Domains**.

View properties of a data domain.

Click a data domain name to view its properties under the **Data Domain Glossary** panel. You can view the description and associated rules. To view the domain groups a data domain belongs to, click **Show in data domain groups**.

View properties of a data domain group.

Click a data domain group name to view its description under the **Data Domain Glossary** panel.

Importing Data Domains

You can import data domains from a source XML file into the data domain glossary in the Developer tool. You must verify that the file contains information on data domains that you need to import.

1. Open the data domain glossary.
2. Verify that **Data Domains** or **Data Domain Groups** is selected.
3. Click **Import**.

   The **Import** dialog box appears.
4. In the **File Name** field, enter the XML file name from which you want to import data domains.

   Click **Browse** to choose the file.
5. Click **Next**.

   The **Select Objects to Import** pane appears where you can specify the source and target.
6. In the **Source** panel, select the data domains that you want to import.

   Note: To select multiple data domains, hold the Shift key.
7. Click **Auto Match to Target** to move the data domains to the **Target** panel.

   The Developer tool tries to match the descendents of the current source selection individually by name, type, and parent hierarchy in the target selection and adds the objects that match.
8. Click **Resolution** to specify how to handle duplicate objects.

   You can rename the imported object, replace the existing object with the imported object, or reuse the existing object. The Developer tool renames all the duplicate objects by default.
9. Click **Next**.

   The Developer tool summarizes the import settings for your review. You can specify additional import settings in the **Additional Import Settings** pane.
10. Click **Finish**.

Exporting Data Domains

You can export data domains and data domain rules from the data domain glossary in the Developer tool to an XML file.

1. Open the data domain glossary.
2. Verify that **Data Domains** or **Data Domain Groups** is selected.
3. Click **Export**.

   The **Export** dialog box appears.
4. To export data domains, select **Export Data Domains**. Select **Export Data Domain Rules** to export data domain rules.
5. Click **Next**.
6. In the **Export to file** panel, select the data domains or data domain rules you want to export.
7. To export data domains, click **Browse** to choose the export file and its location. To export data domain rules to another project in the Model Repository Service, select **Copy to project** and choose the project you want to copy the data domain rules to.
8. Click **Next**.

   The **Dependencies** pane displays a list of the dependent objects.
9. Click **Next**.

   The **Content Export Settings** pane appears. You can select associated reference tables for export.
10. Click **Finish**.

    If you associate a rule that uses reference tables with a data domain, you may not be able to export the reference tables in the same Developer tool session that you use to create the data domain. After you click **Export** in the data domain glossary, disconnect from the Model Repository Service and connect again before you can import the rule that uses reference tables.

Data Domain Discovery Options in Informatica Developer

You can select the source columns, data domains, and inference options when you create a profile to perform data domain discovery. You can also choose to omit columns from data domain discovery based on their data types and data length.

Data Domain Selection in Informatica Developer

The **Data Domain Selection** options list all the domains from the data domain glossary. You can search for specific data domains and select them before you run them as a part of data domain discovery.

The following table describes the **Data Domain Selection** options for data domain discovery:

Option	Description
Enabled as part of the "Run Profile" action	Includes the data domain discovery options when you run the profile.
Name	Data domain name.
Description	Description for the data domain.
Data Domain Group	Name of the data domain groups to which the data domain belongs.
Show data domain group in hierarchy	Lists all data domain groups with the data domains grouped under each data domain group.

Data Domain Column Selection in Informatica Developer

You use the **Column Selection** options to choose the columns you want to run as a part of data domain discovery.

The following table describes the **Column Selection** options for data domain discovery:

Option	Description
Column	Column name.
Data type	Data type of the column.
Precision	Maximum precision for the column.
Scale	Scale of the column.
Nullable	Indicates a column that can have null values.
Description	Description for the column.

Data Domain Inference Options in Informatica Developer

The inference options determine whether domain discovery must run on column data, column name, or both. You can specify whether the profile needs to process all rows in the data source. You can choose a conformance criteria for data domain match and choose to exclude nulls from data domain discovery.

The following table describes the **Inference** options for data domain discovery:

Option	Description
Override the default inference options	Enables you to change the predefined inference options.
Data	Profile runs on column data.
Column name	Profile runs on column titles.
Data and Column name	Profile runs on both column data and column titles.
Maximum rows to profile	The maximum number of rows the profile can run on. The Developer tool chooses the rows starting from the first row in the source.
Minimum percentage of rows	The minimum conformance percentage of rows in the data set required for a data domain match.
Minimum number of rows	The minimum number of rows in the data set required for a data domain match.
Exclude null values from data domain discovery	Excludes the null values from the data set for data domain discovery.

Minimum Conformance Percentage

You can choose a minimum percentage of rows in the data set as a conformance criteria for data domain discovery.

The conformance percentage is the ratio of the number of matching rows divided by the total number of rows.

Note: The Developer tool considers null values as nonmatching rows. Columns containing a high number of null values might not result in data domain inference unless you specify a low value for minimum conformance percentage.

Example

You have a data source with 10,000 rows where the Comments column has Social Security Numbers in 2,500 rows. You create a column profile and data domain discovery and set a minimum percentage of rows to 30% as the conformance criteria. When you run the profile, the profile results do not display the Social Security Numbers as an inferred data domain because the minimum conformance criteria is 30% of rows or 3,000 rows in the data source.

Minimum Conforming Rows

You can choose a minimum number of rows in the data set as a conformance criteria for data domain discovery.

Example

You have a data source with 10,000 rows where the Comments column has email address in three rows. You create a column profile and data domain discovery profile and set the minimum number of rows to 1 as the

conformance criteria. When you run the profile, the profile results display the email address as an inferred data domain with three conforming rows along with the other inferred data domains.

Exclude Null Values

You can exclude null values when you perform data domain discovery on a data source. When you select the minimum percentage of rows with the exclude null values option, the conformance percentage is the ratio of number of matching rows divided by the total number of rows minus the null values in the column.

The data domain discovery process differs when you choose the **Exclude null values from data domain discovery** option and the multiple sampling options or filters.

The following scenarios explain the data domain discovery results when you choose the exclude null values option along with a sampling option and filters:

- With **All rows** as the sampling option and no filters. Data domain discovery ignores all the null values in the column.
- With a sampling option and no filters. Data domain discovery ignores all the null values in the sampled data and runs on the rest of the sampled data.
- With **All rows** as the sampling option and with filters. Data domain discovery ignores all the null values in the filtered data and runs on the rest of the filtered data.
- With a sampling option and filters. Data domain discovery ignores the null values in the filtered data in the sample and runs on the rest of the filtered data.

Example

You have a data source with 10,000 rows where 3,000 rows have Social Security Numbers in the Comments column. You create a column profile and data domain discovery and choose the following options:

- Select the **Exclude null values from data domain discovery** option.
- Select **All rows** as the sampling option.
- Select the **Minimum percentage of rows** option and configure the option to 12%.

When you run the profile, the profile runs on the data set and ignores the null values for data domain discovery.

Creating a Profile to Perform Data Domain Discovery in Informatica Developer

You can discover data domains in a data source either as a part of single data object profile or an enterprise discovery profile. After you perform data domain discovery, you can verify, drill down, and add the results to a data model from the editor in the Developer tool.

1. In the **Object Explorer** view, select the project that contains the data object for the profile.
2. Right-click the data object and select **Profile**.

   The **New** wizard appears.
3. Select **Profile**.
4. Click **Next**.

   The Developer tool displays another pane where you can configure the general properties of the profile.
5. Change the profile name and description, if required. You can also add or remove data objects.

6. Click **Next**.
7. Select the columns you want to run data domain discovery on and the data domains you want to match the columns with.
8. Change the default inference options, as required.
9. Click **Finish** to create the profile.

Editing a Profile in Informatica Developer

You can make changes to a profile after you run it to perform data domain discovery. You can exclude columns with specific data types, change the column selection, data domain selection, and inference options.

1. In the **Object Explorer** view, select the project or folder that contains the profile that you want to edit.
2. Double-click the profile to open it.

   The profile definition appears in a tab.
3. Make changes to column selection, data domain selection, and inference options as required.
4. In the **Column Selection** section, you can click **Exclude Columns** to set up exclusion options based on data types.

   The **Exclude Columns** dialog box appears.
5. Save the changes.

Running a Profile to Perform Data Domain Discovery in Informatica Developer

You can choose to run the profile immediately after you create it. You can also run a profile manually after you create it.

1. In the **Object Explorer** view, select the project or folder that contains the profile that you want to run.

   To run a profile automatically, select **Run Profile on finish** in the **New Profile** wizard when you create the profile.
2. Double-click the profile to open it.

   The profile definition appears in a tab.
3. Right-click the profile and select **Run Profile**.

   The **Run Profile** dialog box appears that displays profile run status.

Data Domain Discovery Results in Informatica Developer

Data domain discovery results display statistics about columns that match data domains, including conformance criteria for data domain match, and whether column names match data domains.

You can drill down the results further for analysis. You can also verify the results on all the rows of the data source and add the results to a data model from the editor in the Developer tool. You can sort the results based on data domains, data domain groups, and columns. You can export data domain discovery results to a Microsoft Excel file.

The following table describes the data domain discovery results:

Column name	Description
Name	Name of the data domain, data domain group, or column based on whether you select the **Data Domain**, **Data Domain Group**, or **Columns** view.
Connection	Name of the connection.
Status	The inference status of the column.
% Data Conformance	The minimum conformance percentage of rows required for a data domain match.
Conforming row count	The minimum number of rows required for a data domain match.
% Null	The percentage of null values for the column.
Total Rows	The total number of rows.
Column Name Match	Indicates whether the column name matches a data domain name.
Data Domain Groups	The data domain group that the data domain belongs to.
Documented Data type	Data type declared for the column in the profile object.
Drilldown	If selected, drills down on rows.
Verified	Indicates the validation of the data domain match on all rows of the data source.
Last Run Time	Date and time of the last profile run.

Viewing by Data Domain Groups

You can view data domain discovery results sorted by data domain groups.

1. Run the profile to view its results.
2. Click **Results**.
3. Click **Data Domain Discovery**.

   You can view the data domain discovery results in the right panel.
4. Verify that the **Data Domain** option is selected in the **Show** field.
5. Select **Show data domain group hierarchy** to view the results sorted by the data domain groups.

Viewing by Columns

You can view data domain discovery results sorted by source columns that match data domains.

1. Run the profile to view its results.
2. Click **Results**.
3. Click **Data Domain Discovery**.

   You can view the data domain discovery results in the right panel.
4. Select **Columns** to view the results sorted by source columns that match data domains.

Verifying the Results

When you run a profile, it analyzes a sample of the data source to infer profile results. You can run the profile on all rows of the source data to verify the inference results.

1. Run the profile to view its results.
2. Click **Results**.
3. Click **Data Domain Discovery**.

   You can view the data domain discovery results in the right panel.
4. Select a column in the right panel that you want to verify.
5. Right-click the column and select **Verify** to run the profile on all rows of the data source.

   You might see a change in the **% Data Conformance** value or **Conforming row count** value after you verify the results.
6. To verify the inference results of multiple columns, select multiple columns. You can then right-click and select **Verify All**.

Approving Data Domains

If you run data domain discovery in a single data object profile, you can approve the inferred data domains for multiple columns at a time. If you run data domain discovery as a part of enterprise discovery, you can approve the data domain of one source column at a time. To approve the data domains of multiple columns after enterprise discovery, you can open the individual data object profile tasks and approve the data domains.

1. In the **Object Explorer** view, select a profile.
2. Double-click the profile to open it.

   The profile opens in a tab.
3. If you ran a single data object profile, select the **Data Domain Discovery** view and then select a row. The row contains data domain discovery results for each column.
4. Right-click the row, and select **Accept**.

   The inference status of the data domain changes to **Accepted**.
5. If you ran enterprise discovery, select the **Data Domains** view and then select a data domain.

   The columns that match the data domain appear in the right panel.
6. Right-click the column that you want to approve and select **Accept**. You can also select multiple rejected columns and approve them as required.

   The inference status of the data domain changes to **Accepted**.
7. To restore the inferred status of the data domain, right-click the row and click **Reset**.

Rejecting Data Domains

Informatica Developer displays inferred data domains in the profile results by default. You can reject inferred or approved data domains. You can choose to show or hide the rejected data domains.

1. In the **Object Explorer** view, select a profile.
2. Double-click the profile to open it.

   The profile opens in a tab.
3. In the **Data Domain Discovery** view, or **Data Domains** view, select a row.
4. To reject inferred data domains, right-click the row and select **Reject**.

   Informatica Developer greys out the rejected data domain from the data domain discovery results.
5. To hide the rejected data domains, right-click the row and select **Hide Rejected**.
6. To view the rejected data domains, right-click one of the rows, and select **Show Rejected**.

Exporting Data Domain Discovery Results from Informatica Developer

When you export the data domain discovery results to an `.xlsx` file from Informatica Developer, you can save the file either to the server or a specific location in the client machine.

1. Run a profile to perform data domain discovery.
2. Click the **Results** view.
3. Click the **Export Results to File** icon.

   The **Export data to a file** dialog box appears.
4. Enter the file name. Optionally, use the default file name.
5. Under **Save**, choose **Save on Client** and click **Browse** to select a location and save the file locally in your computer. By default, Informatica Developer writes the file to a server location set in the Data Integration Service properties of Informatica Administrator.
6. Click **OK**.

CHAPTER 25

Enterprise Discovery in Informatica Developer

This chapter includes the following topics:

Enterprise Discovery in Informatica Developer Overview

Enterprise discovery is the process of discovering column profile statistics, data domains, primary keys, and foreign keys in a large number of data sources. You can perform enterprise discovery across multiple connections or schemas.

As an enterprise data analyst, you may want to discover important data characteristics across a large number of data sources. The requirements can include identification of relational data assets, column profile runs on the discovered data assets, discovery of critical data characteristics within the enterprise, primary keys, and candidate keys. You would also want to view the foreign key relationships that exist across the data sources so that you can derive a data model based on the discovered relationships.

Enterprise discovery finds issues, patterns, trends, and critical data characteristics within information assets in your enterprise. You can choose both data sources that you imported into the Model repository and data sources from external relational connections. The data discovery process includes discovering column profile statistics, data domain analysis, data object structures including candidate keys, and data object

relationships that include foreign keys. You run enterprise discovery in the Developer tool and it performs the following tasks on each data source:

- Run a column profile.
- Discover data domains.
- Infer primary keys.

After running column profiles, data domain discovery, and primary key profiles, the Developer tool runs a foreign key profile across all the data sources. After the Developer tool completes profiling and discovery tasks, it generates a consolidated results summary in both graphical and tabular formats.

You can choose an operating system profile in Informatica Developer. After you choose an operating system profile, the Data Integration Service creates and runs the enterprise discovery profiles based on the permission of the operating system user that you define in the operating system profile.

Enterprise Discovery Process

You can run the enterprise discovery profile to perform enterprise discovery in the Developer tool. You need to configure the data discovery options for different profile types before you run the profile.

The Developer tool creates data objects for the selected data sources and profile tasks for each data object. The tool then runs the profile tasks to generate profile results.

Complete the following steps to perform enterprise discovery:

1. Create a enterprise discovery profile by choosing multiple data objects imported into the Model repository and data sources across multiple, external relational connections.
2. Define configuration settings for data domain discovery, column profile, primary key profile, and foreign key profile.
3. Run the enterprise discovery profile.
4. Refresh the Model Repository Service.

   Note: You need to perform this action as the import of metadata for external connections happens in the Model repository. You need to refresh the Model Repository Service so that the Developer tool reflects the changes to the Model repository.
5. Monitor the profile run and if required, view the statuses of profile tasks that the Developer tool runs.
6. Review the enterprise discovery results summary. The summary includes an interactive graphical user interface view and a tabular view.

Profile Options for Enterprise Discovery

Set up profile options before you run a profile to perform enterprise discovery. Profile options include data domain discovery options, column profile sampling options, and inference options for primary keys and foreign keys.

You can choose to run the enterprise discovery profile after you set up the profile options. You can also choose to create profile tasks after the setup without running the profile.

Data Domain Selection for Enterprise Discovery

Inference options determine whether data domain discovery must run on column data, column name, or both. You can specify whether the profile needs to process all the rows in the data source, and choose a conformance criteria for data domain match.

The following table describes the data domain inference options that you configure for enterprise discovery:

Option	Description
Override the default inference options	Changes the predefined inference options.
Data	The profile runs on column data.
Column name	The profile runs on column titles.
Data and column name	The profile runs on both column data and column titles.
Minimum percentage of rows	The minimum conformance percentage of rows in the data set required for a data domain match. The conformance percentage is the ratio of number of matching rows divided by the total number of rows. **Note:** The Developer tool considers null values as nonmatching rows.
Minimum number of rows	The minimum number of rows in the data set required for a data domain match.
Exclude null values from data domain discovery	Excludes the null values from the data set for data domain discovery.
All rows	The profile runs on all rows of the data source.
Sample first	The maximum number of rows the profile can run on. The Developer tool chooses rows starting from the first row in the source.
Exclude approved data types and data domains from the data type and data domain inference in the subsequent profile runs	Excludes the approved data type or data domain from data type and data domain inference from the next profile run.

Column Profile Sampling Options for Enterprise Discovery

The sampling options determine whether the Developer tool runs a column profile on all rows of the data sources or limited number of rows.

The following table describes the column profile sampling options that you configure for enterprise discovery:

Option	Description
All Rows	Chooses all rows in the data source.
First <number> Rows	The number of rows that you want to run the column profile on. The Developer tool chooses the rows starting from the first row in the data source.
Exclude data type inference for columns with an approved data type	Excludes columns with an approved data type from the data type inference of the column profile run.

Run-time Environment Option

Choose native or Hadoop run-time environment option. Informatica Developer sets the run-time environment in the profile definition after you choose the run-time environment. The run-time environment does not affect the profile results.

The following table describes the run-time environment options for an enterprise discovery profile:

Option	Description
Native	The Developer tool submits the profile jobs to the Profiling Service Module. The Profiling Service Module then breaks down the profile jobs into a set of mappings. The Data Integration Service runs these mappings and writes the profile results to the profile warehouse.
Hadoop	The Data Integration Service pushes the profile logic to the Blaze engine on the Hadoop cluster to run profiles.

Primary Key Inference Options for Enterprise Discovery

You can override the default primary key inference options for enterprise discovery. The options include the maximum number of rows you can run the profile on and minimum conformance percentage.

The following table describes the primary key inference options that you configure for enterprise discovery:

Options	Description
Override the default inference options	Allows you to configure custom settings for primary key inference.
Max Key Columns	Maximum number of columns that can make up a primary key.
Max Rows	Maximum number of rows you can run the profile on.

Options	Description
Minimum Percent	The minimum conformance percentage of the column data required for primary key match.
Maximum Violation Rows	The maximum number of rows with key violations that the profile allows when determining primary keys.

Foreign Key Inference Options for Enterprise Discovery

Set up the foreign key inference options to define the column settings for discovering foreign key relationships between data objects. The foreign key inference results depend on the primary key inference options you set up for enterprise discovery, documented primary keys, and user-defined primary keys.

You can infer the foreign keys in Informatica Developer with one of the following methods:

- Use default values.
- Configure the foriegn key inference options.
- Use foreign key configuration file to configure the auto curation parameters.

The following table describes the foreign key inference options that you configure for enterprise discovery:

Options	Description
Override the default inference options	Changes the predefined inference options.
Data types used in comparisons	The data type used in primary key and foreign key comparisons. **Note:** This option applies if you run a column profile on the data source before the foreign key inference.
Comparison case-sensitivity	Includes case-sensitivity when comparing column data.
Trim values before comparison	Determines whether the Developer tool includes leading or trailing spaces in column data while processing.
Inferred primary keys used in comparisons Use top _ ranked keys	The number of top-ranking primary keys used in foreign key inference when the Developer tool runs a foreign key profile across all the data sources. The Developer tool uses the top-ranking method along with documented primary keys and user-defined primary keys to infer the foreign key relationships. Top ranking of inferred keys is based on the descending conformance percentage rounded to a single decimal precision. For example, the Developer tool considers a conformance percentage of 99.75 as 99.8 and 99.74 as 99.7. The default value is 1. Set the value to -1 if you want the Developer tool to use all inferred keys in foreign key inference. **Note:** If the primary key data sources have approved primary keys, the Developer tool does not use inferred primary keys for foreign key inference.
Max foreign keys between data objects	The maximum number of inferred columns that the Developer tool returns after the profile run that are required for foreign key discovery.
Minimum conformance percentage	The minimum eligibility value in percentage for including columns in the foreign key results.
Regenerate signature	Reloads column signatures if the source data changes.

Auto Curation Parameters for Foreign Key Inference

You can configure the auto curation parameters to infer primary key and foreign key relationships without manual intervention. The auto curation parameters are user-defined custom attributes that you can configure to identify the data relationships based on certain conditions.

When the discovery results include a large number of primary key and foreign key relationships, you might find it difficult to identify the critical data relationships among hundreds of data relationships. You might also find it difficult to curate the relationships based on certain conditions, such as data match or data type. To resolve this issue, you can configure the auto curation parameters and run the enterprise discovery profile.

If the data sources have multiple candidate foreign keys and you want to provide rules to choose a candidate foreign key, you can perform the following actions:

- Configure the **Max Foreign Keys between data objects** and **Minimum conformance percent** options in the enterprise discovery profile wizard.
- Configure the weights and scores for the auto curation parameters in the `ForeignKeyConfig.xml` file.

An administrator can edit and save the foreign key configuration file. Configure the auto curation parameters in the foreign key configuration file. The algorithm infers the primary key and foreign key relationships between multiple data objects based on the auto curation parameters.

The foreign key configuration file, `ForeignKeyConfig.xml`, is available in the following directory:

```
<Informatica installation directory>\services\DataIntegrationService\modules\ProfilingService
```

The auto curation parameters are data overlap match, column name match, relationship type match, and data type match.

Data Overlap Match

Data overlap match is the estimated overlap of values between the primary key and foreign keys. You can set the overlap match in the enterprise discovery profile wizard with the **Minimum conformance percent** option. By default, the **Minimum conformance percent** option is set to 90.

If the data overlap match does not meet the minimum conformance percentage, the foreign key is not considered for auto curation. When the minimum conformance for the data overlap match is met, the remaining parameters are used to compute the adjusted score.

Name Match

The name match parameter is an optional parameter. It uses the Edit Distance algorithm to determine how closely the names of primary and foreign key columns match and sets the score between 0 and 1. Set the name match weight to 0 if you do not want to use this parameter to determine primary key and foreign key relationship.

Relationship Type Match

The relationship type match determines the type of relationship between the primary key and foreign key columns and assigns a fixed score between 0 and 1. The relationship type match is computed based on the column type of the foreign key column.

The following relationship type matches can be set in the `ForeignKeyConfig.xml` file:

- Primary key-foreign key relationship where the foreign key column is a non-key column. The default for this relationship type match is 1. You can find this relationship in many data sources.

- Primary key-primary key relationship where the foreign key column is a primary key column. The default for this relationship type is 0.25. You rarely find this relationship type as it represents a table that has been partitioned vertically.
- Primary key-primary key sequence relationship where the foreign key column is a primary key column and the data type of the column is a sequence data type. For example, OrderID column in a Order table has a sequence data type. The default for this relationship type is zero because the sequence keys might cause multiple false positive foreign keys, which the primary key-primary key algorithm tries to avoid. You can set the relationship type match to a higher score if the data source is known to contain a few sequence data types.

Data Type Match

The data type match compares the data types of the primary key columns to the foreign key columns and assigns a fixed conformance score which is based on how close the data types of the columns match.

The following table lists the fixed data type match scores for different combinations of primary key and foreign keys:

	Numeric Foreign Key	Date Foreign Key	String Foreign Key
Numeric Primary Key	1.0	0.5	0.0
Date Primary Key	0.5	1.0	0.5
String Primary Key	0.0	0.0	1.0

You can change the default data type match scores if required.

Creating an Enterprise Discovery Profile in Informatica Developer

You can create a profile on multiple data sources under multiple connections. The Developer tool creates individual profile tasks for each source.

1. In the **Object Explorer** view, select multiple data objects you want to run a profile on.
2. Click **File > New > Profile** to open the profile wizard.
3. Select **Enterprise Discovery Profile** and click **Next**.
4. Enter a name for the profile and verify the project location. If required, browse to a new location.
5. Verify that the name of the data objects you selected appears within the **Data Objects** section. Click **Choose** to select more data objects, if required.
6. Click **Next**.

   The **Add Resources to Profile Definition** pane appears. You can select multiple, external relational connections and data sources from this pane.
7. Click **Choose** to open the **Select Resources** dialog box.

   The **Resources** pane lists all the internal and external connections and data objects under the Informatica domain.

8. Click **OK** to close the dialog box.
9. Click **Next**.
10. Configure the profile types that you want to run. You can configure the following profile types:
    - Data domain discovery
    - Column profile
    - Primary key profile
    - Foreign key profile

    Note: Select **Enabled as part of "Run Enterprise Discovery Profile" action** for the profile types that you want to run as part of the enterprise discovery profile. Column profiling is enabled by default.
11. Review the options for the profile.

    You can edit the sampling options for column profiles. You can also edit the inference options for data domain, primary key, and foreign key profiles.
12. Select **Create profiles**.

    The Developer tool creates profiles for each individual data source.
13. Select **Run enterprise discovery profile on finish** to run the profile when you complete the profile configuration. If you enabled all the profiling operations, the Developer tool runs column, data domain, and primary key profiles on all selected data sources. Then, the Developer tool runs a foreign key profile across all the data sources.
14. Click **Finish**.

    After you run an enterprise discovery profile, you need to refresh the Model Repository Service before viewing the results. This step is required as the import of metadata for external connections happens in the Model repository. You need to refresh the Model Repository Service so that the Developer tool reflects the changes to the Model repository.

Editing a Profile

You can make changes to an enterprise discovery profile after you set it up. You can exclude columns with specific data types, change the column selection, data domain selection, and inference options.

1. In the **Object Explorer** view, select the project or folder that contains the profile that you want to edit.
2. Click **Team** > **Check Out** to check out the profile.
3. Double-click the profile to open it.
4. Click the **Properties** view.

   The Properties view is under the Default view.
5. Click **Profiles** to view the profile tasks.
6. Select a profile task in the right pane that you want to edit and click **Open**.

   The profile definition appears in a tab.
7. To make changes to the global settings of the enterprise discovery profile, select the profile at the top of **List of profiling tasks**, and click **Configure**.
8. Make the required changes to the profile definition options.
9. Save the changes.
10. Click **Team** > **Check In** to check in the profile.

Running an Enterprise Discovery Profile

You can run an enterprise discovery profile in multiple ways. You can run the profile from the **Object Explorer** view or from the **Profiles** tab in the **Properties** window. You can choose to run individual and multiple profile tasks that form a part of the enterprise discovery profile.

1. In the **Object Explorer** view, select the project or folder that contains the profile that you want to run.

   To run a profile automatically, select **Run enterprise discovery profile on finish** in the **New Enterprise Discovery** wizard when you create the profile.
2. Double-click the profile to open it.

   The profile opens in a tab.
3. In the **Object Explorer** view, right-click the profile and select **Run Enterprise Discovery Profile**.

   Alternatively, you can select **Profiles** in the **Properties** window, select the profile name under the **List of profiling tasks**, and then click **Run**.

   Note: When you run an enterprise discovery profile, you need to refresh the Model Repository Service before you can view the results. This step is required as the import of metadata for external connections happens in the Model repository. You need to refresh the Model Repository Service so that the Developer tool reflects the changes to the Model repository.
4. The **Run** dialog box appears. You can make changes to the global settings of the profile in this dialog box.

   By default, the changes you make apply to the newly added data objects in the enterprise discovery profile.
5. To apply the changes to all the data object profile tasks and the system-generated foreign key profile task in the enterprise discovery profile, select **Use global settings for current profiles**.

   The Developer tool updates all the data object profile tasks and the foreign key profile task based on the changed settings.
6. To run individual profile tasks, select a task and click **Run**.
7. To run multiple profile tasks, click **Run Multiple**.

   The **Run Multiple** dialog box appears.

   Tip: If the enterprise discovery results take a long time to load, you might want to update the statistics of the profiling warehouse database. Multiple enterprise discovery profile runs can result in significant changes to data volume and column values. When you update the statistics, the database runs an execution plan for the SQL queries based on the latest statistics and optimizes the database operations.
8. All tasks are selected by default. Clear the tasks that you do not want to run and click **OK**.

Foreign Key Discovery

A column is a foreign key if its data values match the primary key column values in another data object.

You can perform foreign key discovery on multiple data objects in the Developer tool. Create an enterprise discovery profile to select data objects and define the profile.

Before you perform foreign key discovery, you must identify the parent and child data objects in the enterprise discovery profile. The profile uses one or more keys in the parent object, including its primary key, to discover foreign keys in the child object. After you define the parent and child objects and identify the keys in the parent object, you create and run the profile.

Defining Parent and Child Object Relationships

To find foreign key relationships between two data objects, you must select a parent data object and specify the primary key in that object.

1. Open an enterprise discovery profile that contains the data objects you want to analyze.
2. Select the parent object.
3. Select the primary key in the parent object:
   - Click the **Properties** tab, and click **Keys.**
   - Click **Add**, and select the primary key column in the New Key dialog box.
   - Click **OK** in the **New Key** dialog box. Verify that the primary key is displayed in the **Selected fields** pane and that the **Primary Key** option is checked.

Create a foreign key profile to analyze the child object for foreign keys.

Discovering Foreign Key Relationships Between Data Objects

Use an enterprise discovery profile in the Developer tool to find key relationships between two data objects

The data object that contains the primary key is the parent object, and the data object that contains the foreign key is the child object.

1. Open an enterprise discovery profile that contains the data objects you want to analyze.
2. Right-click the name of a data object and select **Foreign Key Profile**.
3. Enter a name for the profile and verify the project location. If required, browse to a new location. Optionally, enter a text description of the profile.
4. Select the keys in the parent object that the profile will use to find foreign keys in the child object.
5. Save and run the profile.

Foreign Key Analysis Results

After you run a foreign key profile, click the profile name below the modeling editor to see the results of the analysis.

The results view lists the columns that meet the primary-foreign key inference criteria you defined. Click the **Options** button to edit the inference settings. Click a column name and select **Validate** to verify that an inferred key is a valid key for the data objects.

The following table describes the foreign key analysis properties:

Property	Description
Parent Primary Key	A primary key column in the parent data object that the profile uses to find foreign keys in a child object.
Child Foreign Key	A column that the profile infers as a foreign key to the parent primary key in the current row.

Property	Description
Inclusion %	The quantity of data values that match between the primary and foreign key, expresses as a percentage. **Note:** You may see a variance in the Inclusion % value for an inferred column in the foreign key results and after you validate it. For an inferred column, Inclusion % is the number of unique, foreign key column values of a child object that match the unique, primary key column values of the parent object. After you validate an inferred column, it is the number of foreign key column values of a child object that match the primary key column values of the parent object.
Relationship Type	The type of relationship defined for the primary and foreign key columns before the profile runs. If you define a relationship before the profile runs, the profile returns data for the relationship even if the inclusion percentage figure does not meet the confidence threshold set for the profile.
Verified	Indicates that a user has validated the primary-foreign key relationship.
Last Run Time	Data and time the profile last ran.
Relationship Type (In Model)	Indicates that the profile verified the relationship between the columns.

Join Analysis

Join analysis describes the degree of potential joins between two data columns. Use a join profile to analyze column joins in a data source or across multiple data sources.

A join profile displays results as a Venn diagram and as numerical and percentage values. You create and run a join profile from an enterprise discovery profile.

Creating a Join Profile

You can analyze potential joins between data objects in an enterprise discovery profile. The join profile stores the analysis in the Model repository.

1. Create or open an enterprise discovery profile.
2. Verify that the enterprise discovery profile contains the data objects that you need.

   To add a data object to the join profile, drag it from the **Object Explorer** view to the modeling editor.
3. Select the data objects to profile.
4. Right-click the objects and select **Join Profile**.

   The profile wizard opens.
5. Enter a name for the profile. Optionally, enter a text description of the profile.
6. Verify that the names of the data objects appear under **Data Objects** in the wizard.
7. Select or clear the option to **Run profile on finish**.
8. Click **Next**.
9. Select the data columns to include in the profile, and click **Next.**

   If required, scroll down the data objects to view all available columns. The profile runs on all columns by default.

10. Click **Add**.

    The **Join Condition** dialog box appears.
11. Click **New** to activate the column selection fields.
12. Select the data objects and columns to validate.

    You define a join condition between two columns. You can define multiple join conditions across one or more data objects.
13. Click **OK** to create the join condition.

    Optionally, click **Add** to define additional conditions.
14. Verify that the Left and Right join columns are prefixed with the correct data object names.
15. Click **Finish**.

Join Analysis Results

The join analysis **Results** tab provides information about the number and percentage of parent orphan rows, child orphan rows, and join rows. Join analysis results also include Venn diagrams that show the relationships between columns.

The following table describes the properties shown on the **Results** tab:

Property	Description
Left Table	Name of the left table and columns used in the join analysis.
Right Table	Name of the right table and columns used in the join analysis.
Left Only Rows	Number of rows in the left table that cannot be joined.
Right Only Rows	Number of rows in the right table that cannot be joined.
Join Rows	Number of rows included in the join.

Select a join condition to view a Venn diagram that shows the relationships between columns. The area below the Venn diagram also displays the number and percentage of orphaned, null, and joined values in columns.

Double-click a section in the Venn diagram to view the records that the section represents. These records open in the Data Viewer view.

Note: You can export the list of records from the Data Viewer view to a flat file.

Exporting Join Profile Results to File

You can export the data rows returned for a join condition to a delimited file. Export the overlapping rows between the left and right sources or the orphans rows in a source.

1. In the **Object Explorer** view, open the enterprise discovery profile that contains the join analysis.
2. Run the join profile.
3. Select the **Join Results** view.
4. On the **Data Viewer** tab, click the **Export Drilldown Results to File** icon.

   The **Export Data** dialog box appears.

5. Enter a file name, and click **Save**.

Overlap Discovery

Overlap discovery provides information about overlapping data in pairs of columns within a data source or multiple data sources. You can find overlapping data from an enterprise discovery profile. You can validate the profile results and view the results in a Venn diagram.

Overlap discovery identifies overlapping data based on either the default settings or the settings you specify. You can overrride the default settings and specify inference options, including the maximum number of top pairs the overlap discovery returns based on the percentage of overlap. You can also specify a confidence level that determines the eligibility for overlap discovery.

Overlap Discovery Results

The **Overlap Discovery** tab displays information on the participating columns and the overlapping percentage value. The overlap discovery results include Venn diagrams that represent the overlapping data in pairs of columns and the date and time when you last performed overlap discovery.

You can click a column and select **Verify** to view the results as a Venn diagram.

The following table describes the overlap discovery properties:

Property	Description
Left Column	The primary column against which the remaining columns are compared for overlap analysis.
Right Column	The column that is compared to the primary column.
% Overlap	The percentage of overlap between two columns.
Verified	Indicates that you validated the overlap results row.
Last Run Time	The date and time that the overlap discovery last ran.

Informatica Developer displays each overlapping pair two times in the overlap discovery results. Consider data sources Items and Orders. Items has columns "m" and "n." Orders has columns "p" and "q."

The following table shows the overlap discovery results for Items and Orders:

Left Column	Right Column
Items	-
m	Orders.p
m	Orders.q
n	Orders.p
n	Orders.q

Left Column	Right Column
Orders	-
p	Items.m
p	Items.n
q	Items.m
q	Items.m

Discovering Overlapping Data

You can determine overlapping data between pairs of columns in an enterprise discovery profile. The overlap analysis is based on unique values in the columns and does not consider null values.

1. Create or open an enterprise discovery profile that contains the data objects.
2. Select the data objects on which you want to find overlap data.

   You can select a single data object to find overlap data within pairs of columns or multiple data objects.
3. Right-click the objects and select **Overlap Discovery**.

   The **New Overlap Discovery** dialog box appears.
4. Enter a name.
5. Optionally, enter a text description for the overlap analysis.
6. Verify that the names of the data objects appear under **Data Objects** in the wizard.
7. Optionally, select **Run Profile on finish** to run the profile when you complete configuring the settings.
8. Click **Next**.
9. Select the columns for overlap discovery.
10. Click **Next**.

    The default inference options appear in the dialog box.
11. Optionally, specify the inference options for overlap discovery to override the default settings.
12. Click **Finish**.

DDL Script Files

The Data Definition Language (DDL) script files contain the `Create`, `Alter`, and `Drop` SQL statements.

You can specify a file name, location, and target database type when you generate the script files. The Developer tool appends the script file names with the "_create" and "_drop" labels. Virtual columns are not part of the DDL script files.

Creating DDL Scripts from an Enterprise Discovery Profile

When you generate DDL script files from an enterprise discovery profile, you can choose the location where you want to save the script files. You can also choose the database type that you want to run the scripts against. Make sure that you verify and commit all the necessary changes in the enterprise discovery profile before you generate DDL scripts.

1. In the **Object Explorer** view, select an enterprise discovery profile.
2. Right-click the profile, and select **Generate DDL**.

   The **Generate DDL** dialog box appears.
3. Click **Browse** to open the **Save As** dialog box.

   The default file extension is `.sql`.
4. Choose a file location, and enter a file name.
5. Select the type of target database.
6. Click **OK**.

   The Developer tool generates the DDL script files in the location you specified.

Synchronize an Enterprise Discovery Profile

You can synchronize an enterprise discovery profile in the Developer tool.

After you upgrade from version 9.5 or earlier to version 9.6 or later, you can migrate the profiles from the previous version to the upgraded version. For enterprise discovery profiles, if you have added any user-defined keys, documented keys, or relationships in the previous version, then the keys and relationship information persists only in the Model Repository and not in the profiling warehouse. In the upgraded version, when you open the enterprise discovery profile in the Developer tool, the documented or user-defined keys and relationships do not appear in the curated results for the profile.

To synchronize the user-defined keys, documented keys, and relationships in the Model repository to the profiling warehouse, use the Synchronize Enterprise Discovery Profile option in the Developer tool. After you synchronize the enterprise discovery profile, the user-defined and documented keys and relationships are set to Approve, and you can view the curated results in the Developer tool.

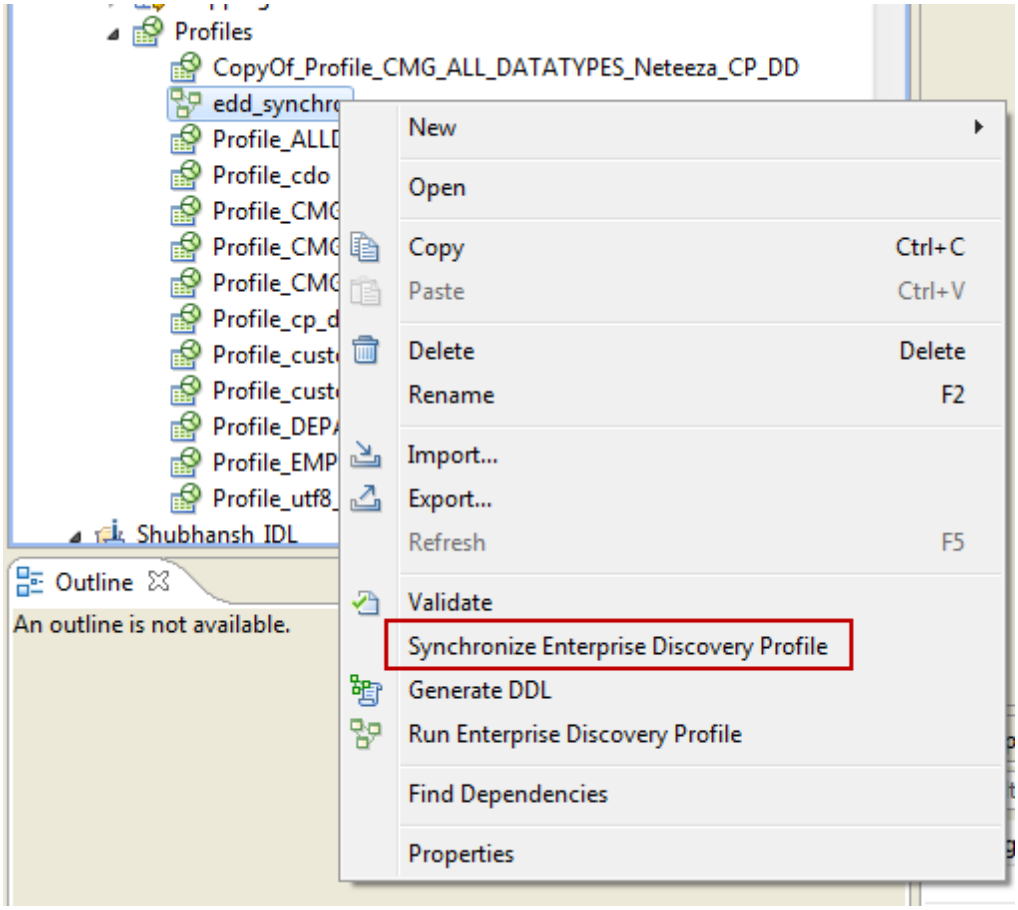
Synchronizing an Enterprise Discovery Profile

In Informatica Developer, you can synchronize the curated results for an enterprise discovery profile after you upgrade from version 9.5 or earlier to version 9.6 or later.

1. In the **Object Explorer** view, select an enterprise discovery profile.

2. Right-click the profile, and select **Synchronize Enterprise Profile**.

   The following image shows the Synchronize Enterprise Discovery Profile option in the Developer tool:



   The curated results for the profile is synchronized.

CHAPTER 26

Enterprise Discovery Results

This chapter includes the following topics:

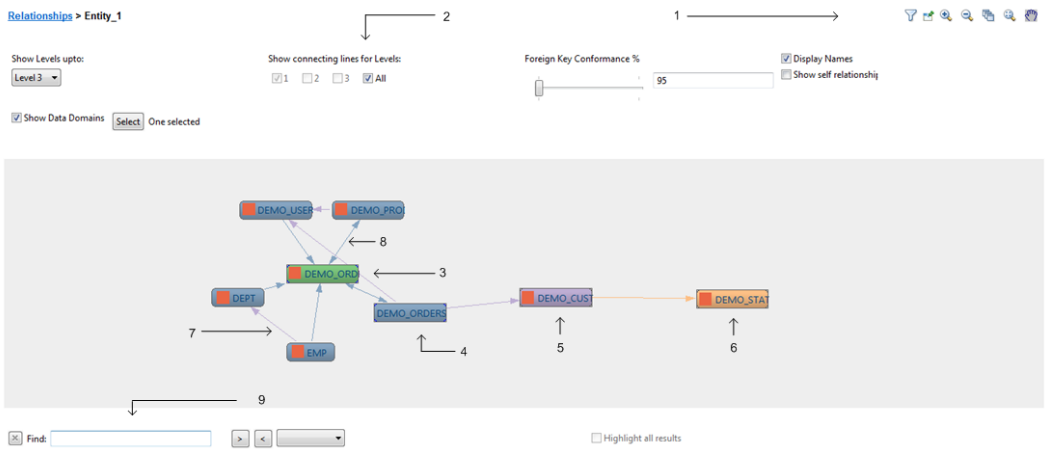
Enterprise Discovery Results Overview

You can view the results of enterprise discovery in multiple views. The views are **Relationships**, **Data Domains**, **Column Profile**, **Join Profile**, and **Overlap Discovery**.

The **Relationships** view displays groups of data objects as circles. You can launch the foreign key profile results from this view. You can view the foreign key profile results in both graphical and tabular views. The **Data Domains** view displays data domain discovery results. The **Column Profile** view displays column profile results for each data object. The **Join Profile** view displays the number of parent orphan rows, child orphan rows, and rows included in the join. The **Overlap Discovery** view provides information about the participating columns and the overlapping percentage value.

Data objects can have multiple relationships between them. The graphical view of the foreign key results displays data object relationships that have the highest conformance percentage.

The following image shows the graphical view of some sample data objects from the enterprise discovery results:



1. Toolbar icons that include filter, pin data object, zoom in, zoom out, arrange all, fit to window, and pan icons.
2. Filter options, such as display different levels of data object relationships, show connecting lines, and show self-related data objects.
3. The selected data object based on which the visual editor displays the rest of the relationships between data objects.
4. First level of data object relationships.
5. Second level of data object relationships.
6. Third level of data object relationships.
7. Connector between data objects. A single arrow head indicates a primary key to foreign key relationship between two data objects. The arrow head points to the data object with the primary key.
8. Connector between data objects. A double arrow head connector indicates a primary key to primary key relationship between two data objects. Mouse over the connector to view the join columns with maximum conformance for the inferred relationship.
9. Press CTRL+F to display the Find field and use an asterisk (*) as a wildcard character to find data objects in the graphical view.

Relationships View

You can view a summary of the enterprise discovery results in the **Relationships** view including entities. Entities are groups of data objects represented as circles. Entities include related, self-related, and unrelated data objects from the multiple connections and schemas of the source databases.

A self-related data object has columns within the data object that have relationships. An unrelated data object neither has relationships with other data objects in the source database nor has a relationship between columns within the data object. The entity relationship diagram of data objects in enterprise discovery results is based on inferred relationships and not the documented relationships in the data sources.

Searching for a Data Object

You can search for a data object in the **Relationships** view or the **Foreign Key Profiling** view. You can use an asterisk (*) as a wild character to find data objects.

1. Verify that you are in either the **Relationships** or **Foreign Key Profiling** view.

2. Type in a part of the data object name you are searching for and add the * wild card character to the beginning or end of the search string based on the search requirement. For example, to search for all data objects starting with the string "CA", type in "CA*" and press the **Enter** key. To search for all data objects that have the string "ZIP" in their names, type in "*ZIP*".

   The search is case-sensitive.

Navigating to the Foreign Key Profiling View

The **Foreign Key Profiling** view displays a consolidated view of foreign key relationships across multiple data objects that you ran the profile on. The circles in the view represent entities, self-related objects, and unreferenced objects.

1. Verify that you are on the **Relationships** tab.

   You can view the foreign key profile link in the right pane.

2. Click **Foreign Key Profile** to open the view.

   The view appears in a new tab. The view displays data objects in different sets of circles based on their relationship type. You can also see the total number of data objects that are a part of the consolidated foreign key view.

3. Optionally, you can click the **Relationships** link to go back to the **Relationships** view.

Foreign key Profiling View

You can view a summary of the enterprise discovery results in a graphical format in the **Foreign Key Profiling** view. You can open the profile results and column-level relationships for a data object in a tabular format from the view.

The profile results for a data object include the column profile, primary key inference, functional dependency inference, and data domain discovery results. After you open the column-level relationships for a data object, you can verify and curate the data relationships. When you verify a data relationship, the Developer tool runs the profile on all rows of the source data to verify the inference results. You can approve, reject, and reset data relationships in the **Foreign Key Profiling** view.

Note: The **Foreign Key Profiling** view does not display any data objects when you use Hive data sources to create an enterprise discovery profile.

Viewing Data Object Relationships

You can view the relationships between data objects in a graphical format. Double-click an entity circle to view the tables and their relationships in it.

1. Verify that you are in the **Foreign Key Profiling** view.
2. To include data domains in the consolidated graphical view, select **Show Data Domains**.

   The **Select** button is enabled.

3. Click **Select** to choose the data domains that you want to include in the graphical view.

   The **Select Data Domains** dialog box appears.

4. Select the required data domains and click **OK**.

   The Developer tool highlights the entity circles that include the data domains that you selected.

5. Double-click an entity circle to view the visual representation of table relationships within the entity. The Developer tool displays the tables in a graphical format that represents the relationship each data object has with other data objects within the entity.

   The data object with the maximum number of relationships with other data objects or the data object from where you start navigating is highlighted in Green color. If you included data domains, the Developer tool highlights the data domain selection to the left of the visual representation of each data object.
6. Verify the direct relationship information and data domain information in the right pane of the **Foreign Key Profiling** view.
7. Optionally, you can click the **Relationships** link to go back to the **Relationships** view.

Zooming In and Out of the View

You can zoom into the graphical representation of data object relationships in the **Foreign Key Profiling** view for better visual clarity. When you zoom in, the Developer tool increases the magnification level of the image. Zoom out to reduce the magnification level.

1. Verify that you are in the **Foreign Key Profiling** view.
2. Right-click the view and select **Zoom In** to increase the magnification level of the image.
3. To reduce the magnification level of the graphical layout, right-click the view and select **Zoom Out**.

Finding a Data Object

You can search for and find a data object in the graphical view of foreign key results. Use an asterisk (*) as a wildcard character to find data objects.

1. Verify that you are in the **Foreign Key Profiling** view.

   Make sure you open the graphical view of foreign key results in the view.
2. Press Ctrl+F to diaply the **Find** field.
3. In the **Find** field, enter a part of the data object name you are searching for and add the ***** wildcard character to the beginning or end of the search string based on the search requirement. For example, to search for all data objects starting with the string "EMP", type in "EMP*" and press the **Enter** key. To search for all data objects that have the string "ZIP" in their names, type in "*ZIP*".
4. Click the **Next Match** button to move to the next data object match.

   Use the **Previous Match** button to move to the previous data object match.
5. Select **Highlight all results** to highlight all the data object matches.
6. To clear the search string in the **Find** field, click the **Clear** button next to the field.

Viewing Column Relationships

You can view the relationship of each column in a data object with columns in related data objects. You can also verify and commit the data object relationship to the data model.

1. Verify that you are in the **Foreign Key Profiling** view.
2. Right-click a data object and select **View Column Relationships**.

   The column relationships appear in a tabular view. The view displays relationship information, such as source data object, related data object, and columns in the related data object.
3. Verify the inference status, verification status, and curation status.

4. Select **All data objects in the group** to view all the data objects in the parent entity and their column relationship information.

   By default, the view displays relationship information for the data object you selected in the view.
5. Optionally, you can click the **entity** link on top of the view to go back to the graphical representation of the data objects.

Saving the Entity Relationship Diagram as an Image

You can save the entity relationship diagram of data objects from the enterprise discovery results as a ".png" file.

1. Run a profile to perform enterprise discovery.
2. Verify that you are in the **Foreign Key Profiling** view.
3. Switch to the graphical format of the data object relationships from the view.
4. Right-click and select **Save as Image**.

   The **Save As** dialog box appears. You save the image as a ".png" file by default.
5. Choose the file location, and enter a file name.
6. Click **Save**.

Viewing Data Object Profile Results From the Foreign Key Profiling View

You can view the column profile, primary key, and data domain discovery results of a selected data object from the **Foreign Key Profiling** view. Make sure you choose the data object by pinning it as the selected table in the canvas.

1. Verify that you are in the **Foreign Key Profiling** view.
2. Right-click a data object and select **Pin Data Object as Focus** to select the table.

   Alternately, use the **Pin** icon to select a data object.
3. Right-click anywhere in the canvas and select **View Data Object Profile**.

   The data object profile results appear in a tab.

Tabular View

When you open the **Foreign Key Profiling** view, The Developer tool displays graphical view of the results by default. Switch to a tabular view to see the tables and their relationship details in a tabular format.

You can view the number of data objects in the entity, names of the related tables, their connection information, and the number of relationships between the two data objects. You can also verify the column relationships and add them to a data model.

Table Details Pane

You can view data object details in both graphical and tabular views of the enterprise discovery results. In the graphical view, the table details pane displays the number of data objects that have a direct relationship with the selected data object and the data object names.

The following table describes the columns of the table details pane in the tabular view:

Column Name	Description
Table Name	Name of the data object that has a direct relationship with the selected data object in the left pane.
Connection	Name of the connection for the related data object.
Relationships	The number of relationships between the selected data object in the left pane and related data object in the table details pane.

Verifying the Enterprise Discovery Results

When you verify the results of enterprise discovery, the Developer tool runs the profile on all rows of the data source. The conformation percentage value may vary after verification depending on the column values across all rows in the data source.

1. Open a profile after you run it.
2. Verify that you are in the **Foreign Key Profiling** view.
3. Click the **Tabular View** icon on top of the view.

   The tabular view displays entities in the left pane.
4. Right-click a data object in the left pane and select **View Column Relationships**.

   You can view the relationships of columns in the selected data object with columns in the other data objects. Scroll to the right to view details such as relationship type, conformation percentage, verification status, and commit status.
5. Right-click a row and select **Verify**.

   The **Run Profile** dialog box appears. After the verification is complete, select the row to view the overlap of primary key and foreign key relationships in a Venn diagram.

Curating Column Relationships

You can approve, reject, and reset data relationships in the **Foreign Key Profiling** view.

1. Open a profile after you run it.
2. Verify that you are in the **Foreign Key Profiling** view.
3. In the graphical view to reject an inferred column relationship, select the data object, right-click a data object, and select one of the following options:
   - **Reject All Relationships with** > **Inferred Primary Keys**. Choose this option to reject all the relationships between the columns with inferred primary keys in the data object to the columns with inferred foreign keys in the other connected data objects.
   - **Reject All Relationships with** > **Inferred Foreign Keys**. Choose this option to reject all the relationships between the columns with inferred foreign keys in the data object to the columns with inferred primary keys in the other connected data objects.

- **Reject All Relationships with > Inferred Primary and Foreign Keys**. Choose this option to reject all the relationships between the columns with inferred primary keys in the data object to the columns with inferred foreign keys in the other connected data objects, and all the relationships between the columns with inferred foreign keys in the data object to the columns with inferred primary keys in the other connected data objects.

4. In the graphical view, right-click a data object and select **View Column Relationships**.
5. Select a data object relationship that you want to curate.
6. To approve the column relationship, right-click and click **Approve**.

   The status of the row changes to **Approved**.
7. To restore the inferred status of the column relationship, right-click and click **Reset**.
8. To view the rejected column relationships, right-click one of the rows, and then select **Show Rejected**.
9. To hide the rejected data types, right-click one of the rows and select **Hide Rejected**.

Committing the Results to the Model Repository

After you run a profile, you can save the column relationships between data objects to the Model repository. You can commit the relationships to the Model repository from the tabular view of the **Foreign Key Profiling** view.

1. Open a profile after you run it.
2. Verify that you are in the **Foreign Key Profiling** view.
3. Right-click a data object in the left pane and select **View Column Relationships**.

   You can view the relationships of columns in the selected data object with columns in the other data objects.
4. Right-click a row and select **Approve**.

Data Domains View

The **Data Domains** view lists the data domains and matching column statistics that the Developer tool discovers as part of enterprise discovery. You can verify columns, drill down on rows, and view data object profile results from the **Data Domains** view.

Viewing Data Domain Discovery Results

You can view the data domain discovery results on the **Data Domains** tab. You can search for data domains and view data domains sorted by data domain groups.

1. Run a profile to perform enteprise discovery.
2. Open the profile.
3. Click the **Data Domains** tab to view the data domain discovery results.

   The data object profile results appear in the right pane.
4. Type in a part of a data domain name in the search field to find a specific data domain.

   Select **Show data domain group hierarchy** to sort the list of data domains by the data domain groups.

Verifying Data Domain Discovery Results

When you run a profile, it analyzes a sample of the data source to infer profile results. You can run the profile on all rows of the source data to verify the inference results.

1. Open a profile after you run it.
2. Click the **Data Domains** tab to view the results.

   You can view the data domain discovery results in the right panel.
3. Select a column in the right panel that you want to verify.
4. Right-click the column and select **Verify** to run the profile on all rows of the data source.

   You may see a change in the **% Data Conformance** value or **Conforming row count** value after you verify the results.

Drilling Down on Rows

You can drill down on data domain discovery results for further data analysis.

1. Open a profile after you run it.
2. Click the **Data Domains** tab to view the results.

   You can view the data domain discovery results in the right panel.
3. Select a row in the right panel that you want to drill down.
4. Right-click the column and select **Drill Down** to drill down to the source rows.

Viewing Data Object Profile Results from the Data Domains View

You can view data object profile results of a selected data object from the **Data Domains** view.

1. Verify that you are in the **Data Domains** view.
2. Select a data domain in the **Domains Profiled** pane.
3. In the **Columns** pane to the right, select a column.
4. Right-click the column and select **Open Data Object Profile**.

   The data object profile results appear in a tab.

Column Profile View

The **Column Profile** view displays a summary of the column profile results for single data object profiles that the Developer tool runs as part of enterprise discovery. You can view the column statistics such as unique values, null values, data types and maximum and minimum values for each column in the data objects.

Viewing Data Object Profile Results

Enterprise discovery includes running a data object profile to discover column data statistics, primary keys, candidate keys, and data domains. You can view data object profile results of a selected data object from the **Column Profile** view.

1. Verify that you are in the **Column Profile** view.

2. Select a data object in the **Data Objects Profiled** pane.
3. In the **Columns** pane to the right, select a column.
4. Right-click the column and select **View Data Object Profile**.

   The data object profile results appear in a tab. Column profile results appear by default.
5. Click **Primary Key Inference** to view the primary key profile results.
6. Click **Functional Dependency Inference** to view the functional dependency discovery results.
7. Click **Data Domain Discovery** to view the data domain discovery results.

Viewing Column Profile Results During Enterprise Discovery Run

The time taken to complete enterprise discovery depends on the number of profile tasks, data source size, and profile type. When the Developer tool continues to run the data discovery tasks, you can view the results of column profiles that it completes in the initial stages of data discovery.

1. After you run the profile, click **Profiles** in the **Properties** window.
2. Select the column profile that you want to view the results for. Make sure the the status of the profile run is **Success** in the **Properties** window.
3. Click **Open** to view the results in another tab.
4. On the **Results** section, select **Column Profiling** to view the results in the right pane.

Viewing Data Domain Discovery Results During Enterprise Discovery Run

When the Developer tool continues to run data discovery tasks included in enterprise discovery, you can view the results of data domain discovery that the Developer tool completes in the initial stages of enterprise discovery.

1. After you start running the profile, click **Profiles** in the **Properties** window.
2. Select the profile that you want to view the data domain results for. Make sure the the status of the profile run is **Success** in the **Properties** pane.
3. Click **Open** to view the results in another tab.
4. On the **Results** section, select **Data Domain Discovery Profiling** to view the results in the right pane.

Viewing the Run-time Status of Enterprise Discovery

The **Progress** view of the Developer tool displays the progress of operations such as a profile run. You can view the run-time status of enterprise discovery tasks from the **Progress** view.

1. After you run a profile to perform enterprise discovery on the data sources, click the **Progress View** button at the bottom, right corner of the Developer tool.

   The **Progress** pane appears, if it is not open already.
2. Click the **Enterprise Discovery Running : View Tasks Status** link to open the subtasks dialog box.

   The dialog box lists multiple profile tasks that are part of the enterprise discovery. You can view the profile name, type, and its status.
3. Click the column header to sort the profile tasks. For example, to sort the profile tasks by its status, click the **Status** column header.
4. If you need to cancel a specific profile task, select the task and click **Cancel**.

   The status of the canceled task changes to **Terminated**.

Enterprise Discovery Export Files

After you run an enterprise discovery profile, you can export information including all the data object relationships, data domains, and individual foreign key task results. You can save the graphical image of the data object relationships as an `.jpg` file.

When you export the profile results, the Developer tool saves all the enterprise discovery results in multiple Microsoft Excel files. You can view data object relationships, column profile results, data domain discovery results, entities, and individual foreign key task results in separate files.

Exporting Enterprise Discovery Results

You can export the list of entities and composition of each entity, all data objects and column-level data object relationships of an entity, data domains, and column profiling results.

1. Run a profile to perform enterprise discovery.
2. From the **Relationships**, **Data Domains**, or **Column Profile** view, click the **Export** icon on the top right area of the window.

   The **Export data to a file** dialog box appears.
3. Enter the file name. Optionally, use the default file name.
4. Under **Save**, choose **Save on Client** and click **Browse** to select a location and save the file locally in your computer. By default, Informatica Developer writes the file to a location set in the Data Integration Service properties of Informatica Administrator.
5. Click **OK**.

CHAPTER 27

Business Glossary Desktop in Informatica Developer

This chapter includes the following topics:

Business Glossary Search

Look up the meaning of a Developer tool object name as a business term in the Business Glossary Desktop to understand its business requirement and current implementation.

A business glossary is a set of terms that use business language to define concepts for business users. A business term provides the business definition and usage of a concept. The Business Glossary Desktop is a client that connects to the Metadata Manager Service, which hosts the business glossary. Use the Business Glossary Desktop to look up business terms in a business glossary.

If Business Glossary Desktop is installed on your machine, you can select an object in the Developer tool and use hotkeys or the Search menu to look up the name of the object in the business glossary. You can look up names of objects in Developer tool views, such as the **Object Explorer** view, or names of columns, profiles, and transformation ports in the editor.

For example, a developer wants to find a business term in a business glossary that corresponds to the Sales_Audit data object in the Developer tool. The developer wants to view the business term details to understand the business requirements and the current implementation of the Sales_Audit object in the Developer tool. This can help the developer understand what the data object means and what changes may need to be implemented on the object.

Looking Up a Business Term

Look up a Developer tool object name in the Business Glossary Desktop as a business term to understand its business requirement and current implementation.

You must have the Business Glossary Desktop installed on your machine.

1. Select an object.
2. Choose to use hotkeys or the Search menu to open the Business Glossary Desktop.
   - To use hotkeys, use the following hotkey combination:

     ```
     CTRL+Shift+F
     ```
   - To use the Search menu, click **Search** > **Business Glossary**.

   The **Business Glossary Desktop** appears and displays business terms that match the object name.

Customizing Hotkeys to Look Up a Business Term

Customize hotkeys to change the combination of keys that open the Business Glossary Desktop.

1. From the Developer tool menu, click **Window** > **Preferences** > **General** > **Keys**.
2. To find or search **Search Business Glossary** in the list of commands, select one of the following choices:
   - To search for the keys, enter Search Business Glossary in the search box.
   - To scroll for the keys, scroll to find the **Search Business Glossary** command under the **Command** column.
3. Click the **Search Business Glossary Command**.
4. Click **Unbind Command**.
5. In the **Binding** field, enter a key combination.
6. Click **Apply** and then click **OK**.

INDEX

B

C

D

E

F

I

J

M

O

P

R

S

T