



Informatica®
10.5

Profile Guide

Informatica Profile Guide
10.5
December 2019

© Copyright Informatica LLC 2010, 2021

This software and documentation are provided only under a separate license agreement containing restrictions on use and disclosure. No part of this document may be reproduced or transmitted in any form, by any means (electronic, photocopying, recording or otherwise) without prior consent of Informatica LLC.

Informatica and the Informatica logo are trademarks or registered trademarks of Informatica LLC in the United States and many jurisdictions throughout the world. A current list of Informatica trademarks is available on the web at <https://www.informatica.com/trademarks.html>. Other company and product names may be trade names or trademarks of their respective owners.

U.S. GOVERNMENT RIGHTS Programs, software, databases, and related documentation and technical data delivered to U.S. Government customers are "commercial computer software" or "commercial technical data" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, the use, duplication, disclosure, modification, and adaptation is subject to the restrictions and license terms set forth in the applicable Government contract, and, to the extent applicable by the terms of the Government contract, the additional rights set forth in FAR 52.227-19, Commercial Computer Software License.

Portions of this software and/or documentation are subject to copyright held by third parties. Required third party notices are included with the product.

The information in this documentation is subject to change without notice. If you find any problems in this documentation, report them to us at infa_documentation@informatica.com.

Informatica products are warranted according to the terms and conditions of the agreements under which they are provided. INFORMATICA PROVIDES THE INFORMATION IN THIS DOCUMENT "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

Publication Date: 2021-04-13

Table of Contents

Preface	8
Informatica Resources.	8
Informatica Network.	8
Informatica Knowledge Base.	8
Informatica Documentation.	8
Informatica Product Availability Matrixes.	9
Informatica Velocity.	9
Informatica Marketplace.	9
Informatica Global Customer Support.	9
Part I: Introduction to Profiles	10
Chapter 1: Introduction to Profiles	11
Introduction to Profiles Overview.	11
Profiling Process.	12
Profiling Tools.	12
Profile Components.	12
Chapter 2: Column Profile Concepts	14
Column Profile Concepts Overview.	14
Column Profile Options.	15
Repository Profile Locks and Versioned Profile Management.	15
Scorecards.	16
Chapter 3: Curation Concepts	17
Curation Concepts Overview.	17
Curation Tasks.	17
Part II: Profiling with Informatica Analyst	19
Chapter 4: Column Profiles in Informatica Analyst	20
Column Profiles in Informatica Analyst Overview.	20
Column Profiling Process.	21
Profile Options.	21
Sampling Options.	22
Drilldown Options.	22
Run-time Environment.	22
Native Environment.	23
Hadoop Environment.	23
Databricks Environment.	24

Operating System Profiles in Informatica Analyst Overview.	24
Selecting an Operating System Profile.	24
Repository Asset Locks and Team-based Development Overview.	24
Creating a Column Profile in Informatica Analyst.	25
Editing a Column Profile.	26
Running a Profile.	27
Running a Profile on the Spark Engine.	27
Running a Profile on the Databricks Cluster.	27
Synchronize Option.	27
Synchronizing a Flat File Data Object in Informatica Analyst.	28
Synchronizing a Relational Data Object in Informatica Analyst.	30
Chapter 5: Rules in Informatica Analyst.	31
Rules in Informatica Analyst Overview.	31
Predefined Rules.	31
Predefined Rules Process.	32
Applying a Predefined Rule.	32
Expression Rules.	33
Creating an Expression Rule.	33
Creating an Expression Rule Using Rule Specification.	34
Chapter 6: Filters in Informatica Analyst.	36
Filters in Informatica Analyst Overview.	36
Creating a Filter.	36
Creating a Simple Filter.	37
Creating an Advanced Filter.	38
Creating an SQL Filter.	39
Managing Filters.	39
Chapter 7: Column Profile Results in Informatica Analyst.	41
Column Profile Results in Informatica Analyst Overview.	41
Summary View.	42
Summary View Properties.	43
Default Filters in Summary View.	43
Detailed View.	44
Detailed View Panes.	45
Statistics.	46
Data Preview.	47
Data Types.	47
Outliers.	48
Patterns.	49
Values.	50
Types of Profile Run.	53

Latest Profile Run.	53
Historical Profile Run.	53
Consolidated Profile Run.	53
Selecting a Profile Run.	53
Compare Multiple Profile Results Overview.	54
Comparing Multiple Profile Results.	54
Summary View of Compare Profile Results	55
Detailed View of Compare Profiles Results	58
Column Profile Drilldown.	59
Drilling Down on Row Data.	59
Applying Filters to Drilldown Data.	59
Curation in the Analyst tool.	60
Approving Data types and Data Domains.	60
Rejecting Data types and Data Domains.	60
Column Profile Export Files in Informatica Analyst.	60
Profile Export Results in a CSV File.	61
Profile Export Results in Microsoft Excel.	61
Exporting Profile Results from Informatica Analyst.	61
Chapter 8: Scorecards in Informatica Analyst.	63
Scorecards in Informatica Analyst Overview.	63
Informatica Analyst Scorecard Process.	64
Creating a Scorecard in Informatica Analyst.	65
Adding Columns to an Existing Scorecard.	66
Running a Scorecard.	67
Viewing a Scorecard.	67
Editing a Scorecard.	67
Metrics.	68
Metric Weights.	68
Value of Data Quality.	68
Defining Thresholds.	69
Metric Groups.	70
Creating a Metric Group.	70
Moving Scores to a Metric Group.	70
Editing a Metric Group.	71
Deleting a Metric Group.	71
Drilling Down on Columns.	71
Trend Charts.	72
Score Trend Chart.	72
Cost Trend Chart.	72
Viewing Trend Charts.	73
Exporting Trend Charts.	74
Scorecard Export Files in Informatica Analyst.	75

Scorecard Export Results in Microsoft Excel.	75
Exporting Scorecard Results from Informatica Analyst.	75
Scorecard Notifications.	75
Notification Email Message Template.	76
Setting Up Scorecard Notifications.	77
Configuring Global Settings for Scorecard Notifications.	77
Scorecard Lineage.	78
Viewing Scorecard Lineage in Informatica Analyst.	78
Part III: Profiling with Informatica Developer.	79
Chapter 9: Data Object Profiles.	80
Column Profiles in Informatica Developer.	80
Filtering Options.	80
Sampling Options.	81
Creating a Single Data Object Profile in Informatica Developer.	81
Creating Multiple Data Object Profiles in Informatica Developer.	82
Synchronizing a Flat File Data Object in Informatica Developer.	83
Synchronizing a Relational Data Object in Informatica Developer.	84
Chapter 10: Column Profiles on Semi-structured Data Sources.	85
Column Profiles on Semi-structured Data Sources Overview.	85
JSON and XML Data Objects.	86
Creating a Data Object from a JSON or XML Data Source.	86
Complex File Data Objects for Semi-Structured Data Sources in HDFS.	87
Complex File Data Object from a JSON or XML Data Source in HDFS.	87
Complex File Data Object from an Avro or Parquet Data Source in HDFS.	87
Creating an HDFS Connection.	88
Creating a Complex File Data Object from a JSON or XML File in HDFS.	88
Creating a Complex File Data Object from an Avro or Parquet Data Source.	89
Creating a Column Profile on a Semi-structured Data Source.	90
Chapter 11: Rules in Informatica Developer.	92
Rules in Informatica Developer Overview.	92
Creating a Rule in Informatica Developer.	93
Applying a Rule in Informatica Developer.	93
Chapter 12: Mapplet and Mapping Profiling.	94
Mapplet and Mapping Profiling Overview.	94
Running a Profile on a Mapplet or Mapping Object.	94
Comparing Profiles for Mapping or Mapplet Objects.	95
Generating a Mapping from a Profile.	95

Chapter 13: Column Profile Results in Informatica Developer.....	96
Column Profile Results in Informatica Developer.....	96
Column Value Properties.....	97
Column Pattern Properties.....	97
Column Statistics Properties.....	97
Column Data Type Properties.....	98
Curation in Informatica Developer.....	99
Approving Datatypes.....	99
Rejecting Data Types.....	99
Exporting Profile Results from Informatica Developer.....	100
Chapter 14: Scorecards in Informatica Developer.....	101
Scorecards in Informatica Developer Overview.....	101
Creating a Scorecard.....	101
Exporting a Resource File for Scorecard Lineage.....	102
Viewing Scorecard Lineage from Informatica Developer.....	102
Index.....	103

Preface

Use the *Informatica Profile Guide* to learn how you can use profile to analyze the content and structure of data sources. You can determine the characteristics of source data in columns, such as value frequency, percentages, and patterns. The guide is written for data analysts and developers.

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-matrixes>.

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 Profiles

This part contains the following chapters:

- [Introduction to Profiles, 11](#)
- [Column Profile Concepts, 14](#)
- [Curation Concepts, 17](#)

CHAPTER 1

Introduction to Profiles

This chapter includes the following topics:

- [Introduction to Profiles Overview, 11](#)
- [Profiling Process, 12](#)
- [Profiling Tools, 12](#)
- [Profile Components, 12](#)

Introduction to Profiles Overview

Create and run a profile 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 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.
- Add rules to column profiles.
- Curate the inferred data types in the profile results.
- Use scorecards to monitor data quality.
- Generate a mapping from a profile.

Profiling 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. You uncover and document data quality issues. Complete the following tasks to perform profiling:

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. Review profiling results.
3. Create reference data.
4. Drill down on profile results.
5. Document data issues.
6. Create and run rules.
7. Create scorecards to monitor data quality.

Profiling Tools

You use multiple Informatica tools to manage the profiling process.

You can use the following tools to manage the profiling 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. You create profiles using a wizard.

Informatica Analyst

You can run a column profile in the Analyst tool. After you run a profile, you can drill down on data rows in a data source.

Profile 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.

Comment

Description about the profile. Use comments to share information about profiles with other Analyst and Developer tool users.

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.

CHAPTER 2

Column Profile Concepts

This chapter includes the following topics:

- [Column Profile Concepts Overview, 14](#)
- [Column Profile Options, 15](#)
- [Repository Profile Locks and Versioned Profile Management, 15](#)
- [Scorecards, 16](#)

Column Profile Concepts 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 3

Curation Concepts

This chapter includes the following topics:

- [Curation Concepts Overview, 17](#)
- [Curation Tasks, 17](#)

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 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: Profiling with Informatica Analyst

This part contains the following chapters:

- [Column Profiles in Informatica Analyst, 20](#)
- [Rules in Informatica Analyst, 31](#)
- [Filters in Informatica Analyst, 36](#)
- [Column Profile Results in Informatica Analyst, 41](#)
- [Scorecards in Informatica Analyst, 63](#)

CHAPTER 4

Column Profiles in Informatica Analyst

This chapter includes the following topics:

- [Column Profiles in Informatica Analyst Overview, 20](#)
- [Column Profiling Process, 21](#)
- [Profile Options, 21](#)
- [Run-time Environment, 22](#)
- [Operating System Profiles in Informatica Analyst Overview, 24](#)
- [Repository Asset Locks and Team-based Development Overview, 24](#)
- [Creating a Column Profile in Informatica Analyst, 25](#)
- [Editing a Column Profile, 26](#)
- [Running a Profile, 27](#)
- [Running a Profile on the Spark Engine, 27](#)
- [Running a Profile on the Databricks Cluster, 27](#)
- [Synchronize Option, 27](#)

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. You can choose a maximum of 2,147,483,647 rows.
Random sample <number> rows	The random sample algorithm chooses the rows at random in the data object to run the profile on. You can choose a maximum of 2,147,483,647 rows.
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 or Hadoop as the run-time environment for a column profile. You can choose a Blaze or Spark engine in the Hadoop run-time environment. 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 the Hadoop option to run the profiles in the Hadoop run-time environment.

After you choose the Hadoop option, you can 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 Blaze engine processes the mappings and the Data Integration Service writes the profile results to the profiling warehouse.

Column Profiles for Sqoop Data Sources

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

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.

Databricks Environment

You can choose the Databricks Spark option to run the profiles in the Databricks run-time environment.

After you choose the Databricks Spark option, you can select a Databricks connection. The Data Integration Service pushes the profile logic to the Spark engine on the Databricks cluster to run profiles.

When you run a profile in the Databricks 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 pushes the mappings to the Spark engine through the Hadoop connection. The Spark engine processes the mappings and the Data Integration Service writes the profile results to the profiling warehouse.

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 domain 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** or **Hadoop** option as the run-time environment. If you choose the 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.

You can view the profile summary and mapping log files to get more information about the tasks performed by the Analyst tool.

Note: If you ran the profile using an operating system profile, the summary log appears in the log directory configured for the Data Integration Service and the mapping log appears in the log directory configured for the operating system profile.

Running a Profile on the Spark Engine

When you run a profile with the JDBC connection on the Spark engine, the profile run fails.

Before you run the profile on the Spark engine, perform the following steps:

1. Create a JDBC warehouse connection.
2. Obtain the Data Direct JAR files of the database that you use to extract data.
3. Copy the files to the following locations: `<INFA_HOME>/externaljdbcjars`

Running a Profile on the Databricks Cluster

When you run a profile with the JDBC connection on the Spark engine using the Databricks cluster, the profile run fails.

Before you run the profile on the Spark engine, perform the following steps:

1. Create a JDBC warehouse connection.
2. Obtain the Data Direct JAR files of the database that you use to extract data.
3. Copy the files to the following locations: `<INFA_HOME>/externaljdbcjars`

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

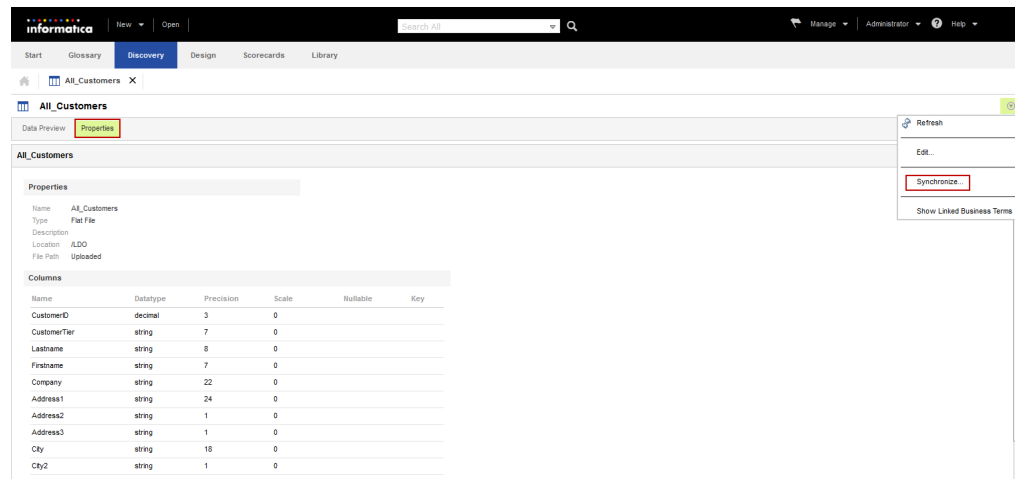
source metadata. You can use the Synchronize option for column profiles, enterprise discovery profiles, and scorecards. The external data source can be a relational data source or flat file data source.

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:

Synchronize Flat File: Step 1 of 5

Specify a location to import the flat file from and specify how to import the flat file.

Browse and Upload:

No file selected.

Enter a Network Path:

Hadoop File System

Description

Upload files from a local machine. Recommended for smaller files up to 10 MB. The Analyst tool uploads a copy of the file to the node on which the Analyst Service runs. Upload the file again if you modify the file.



Back

Next

Finish

Cancel

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 5

Rules in Informatica Analyst

This chapter includes the following topics:

- [Rules in Informatica Analyst Overview, 31](#)
- [Predefined Rules, 31](#)
- [Expression Rules, 33](#)

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 maplet 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.

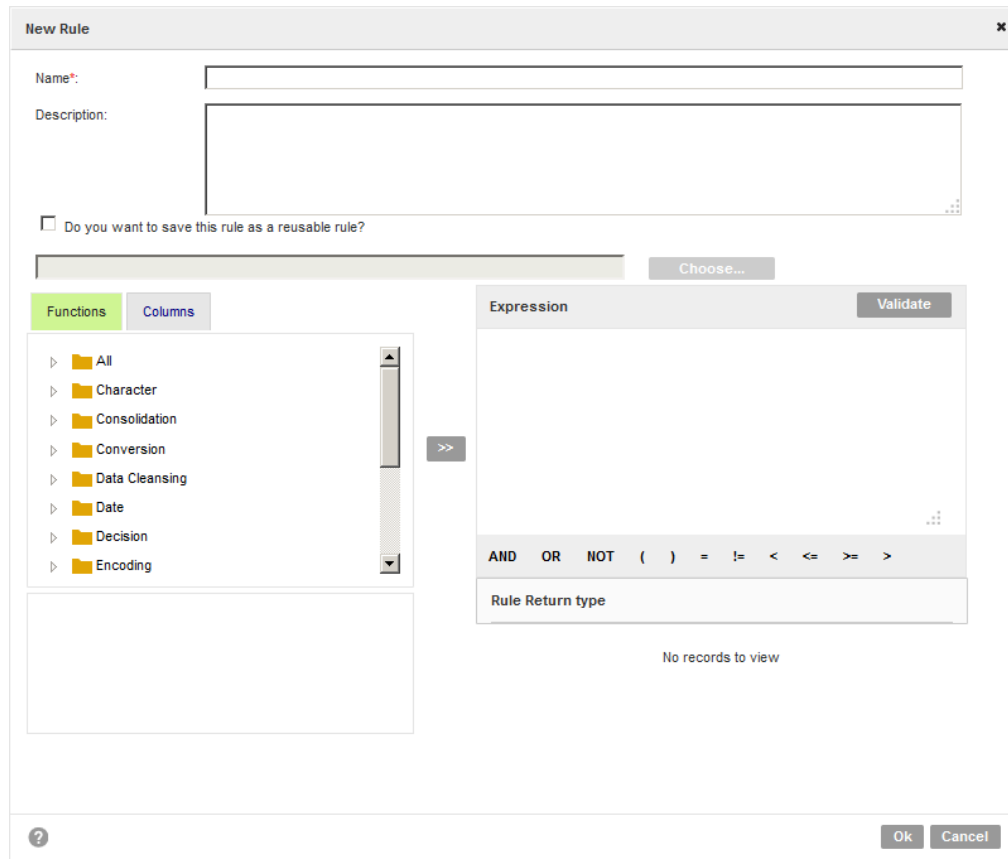
The output of an expression rule is a virtual column that uses the name of the rule as the column name. The Analyst tool runs a column profile on the virtual column. For example, you use an expression rule to validate a ZIP code. The rule returns 1 if the ZIP code is valid and 0 if the ZIP code is not 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 6

Filters in Informatica Analyst

This chapter includes the following topics:

- [Filters in Informatica Analyst Overview, 36](#)
- [Creating a Filter, 36](#)
- [Managing Filters, 39](#)

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

New Filter [Close]

Create a filter. The filter is used to create a subset of the data rows before profiling.

Name*:

Description:

Choose the filter type*: Simple Advanced SQL

Columns	Operator	Values(s)
-Select-	-Select-	+

Filter Preview [Refresh]

[Help] [Ok] [Cancel]

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**.

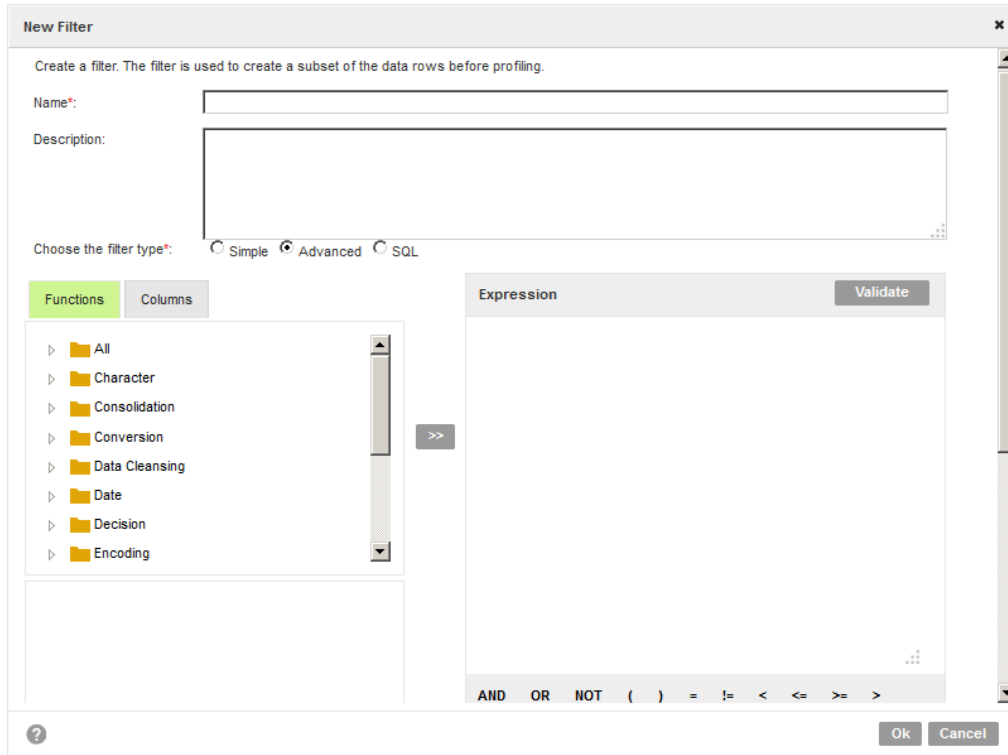
The **Specify Rules and Filters** page appears with the filter in the Filters pane.

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.
4. To verify the advanced filter, click **Validate**.
5. Click **OK**.

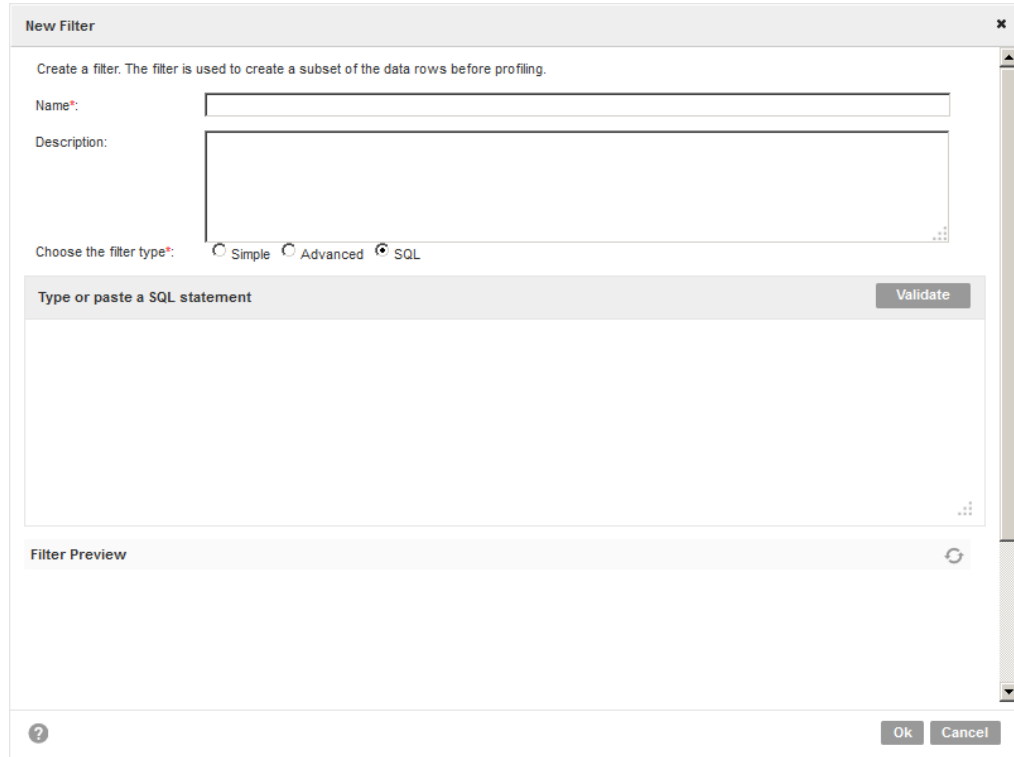
The **Specify Rules and Filters** screen appears with the filter in the Filters pane.

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:



The screenshot shows the 'New Filter' dialog box. At the top, it says 'Create a filter. The filter is used to create a subset of the data rows before profiling.' Below this are fields for 'Name*' and 'Description:'. Underneath, there are radio buttons for 'Simple', 'Advanced', and 'SQL', with 'SQL' selected. A section titled 'Type or paste a SQL statement' contains a large text area and a 'Validate' button. At the bottom, there is a 'Filter Preview' section with a refresh icon, and 'Ok' and 'Cancel' buttons.

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 7

Column Profile Results in Informatica Analyst

This chapter includes the following topics:

- [Column Profile Results in Informatica Analyst Overview, 41](#)
- [Summary View, 42](#)
- [Detailed View, 44](#)
- [Statistics, 46](#)
- [Types of Profile Run, 53](#)
- [Compare Multiple Profile Results Overview, 54](#)
- [Column Profile Drilldown, 59](#)
- [Curation in the Analyst tool, 60](#)
- [Column Profile Export Files in Informatica Analyst, 60](#)

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.

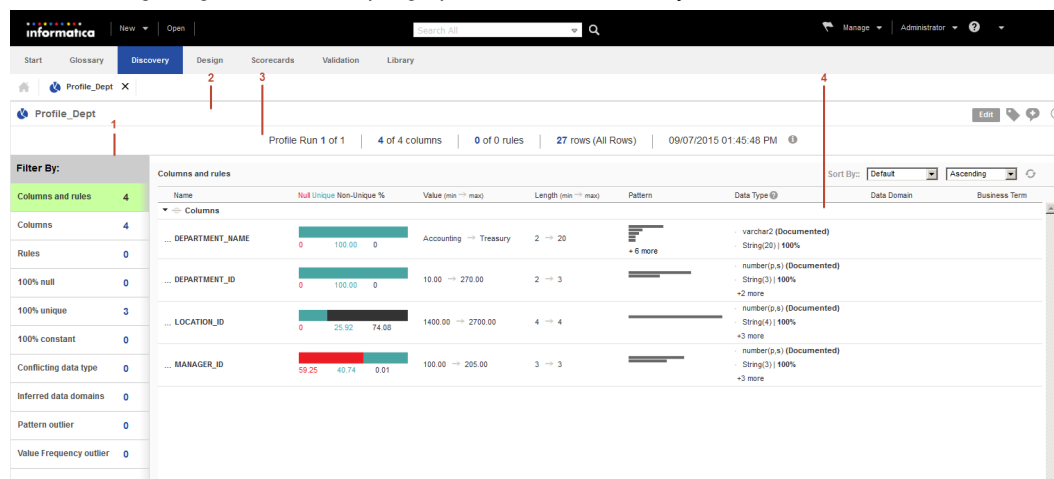
Note: You can view and run a profile on Avro, JSON, Parquet, and XML data sources. You can view profile results for the latest run, historical run, and consolidated run and compare profile results for two profile runs.

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	<p>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:</p> <ul style="list-style-type: none">- String- Varchar- Decimal- Integer- Date <p>You can also view the percentage of conformance based on the inferred data types.</p> <p>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.</p>
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:

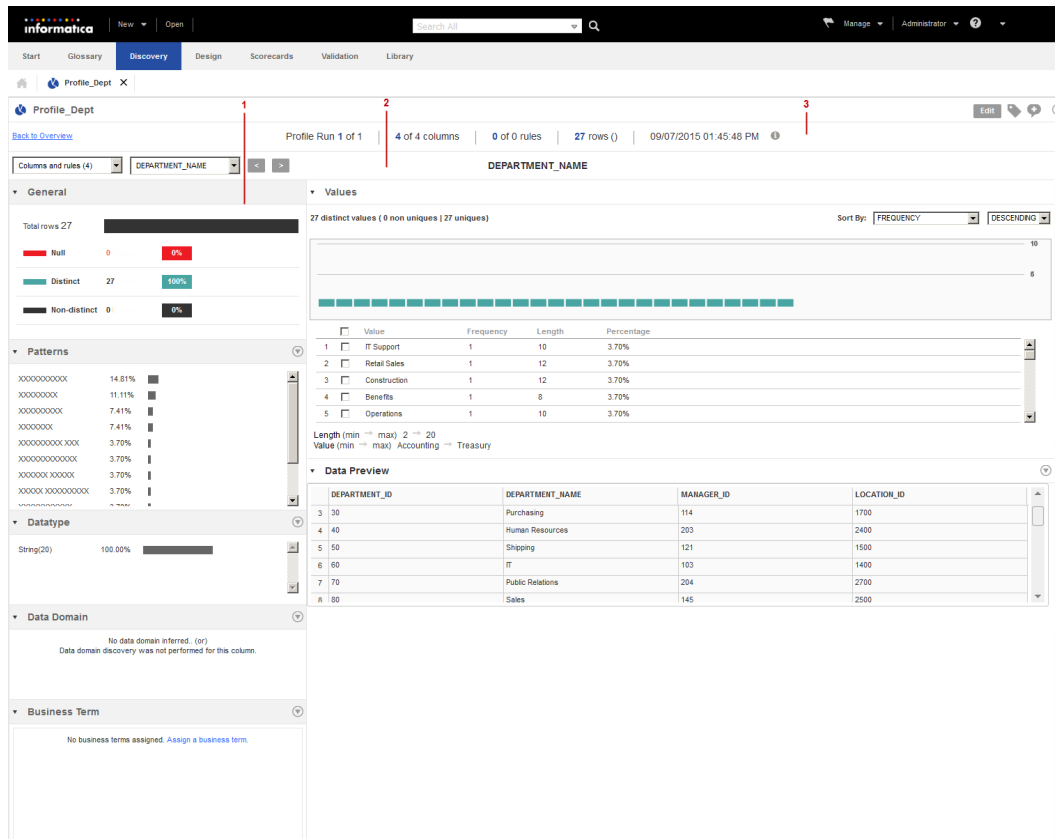
Default Filter Option	Description
Columns and rules	Displays the profile results for the source columns, and rule columns. You can expand and collapse the source columns and rule columns to view the results.
Columns	Displays the profile results for the source columns.
Rules	Displays the profile results 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 rules 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.

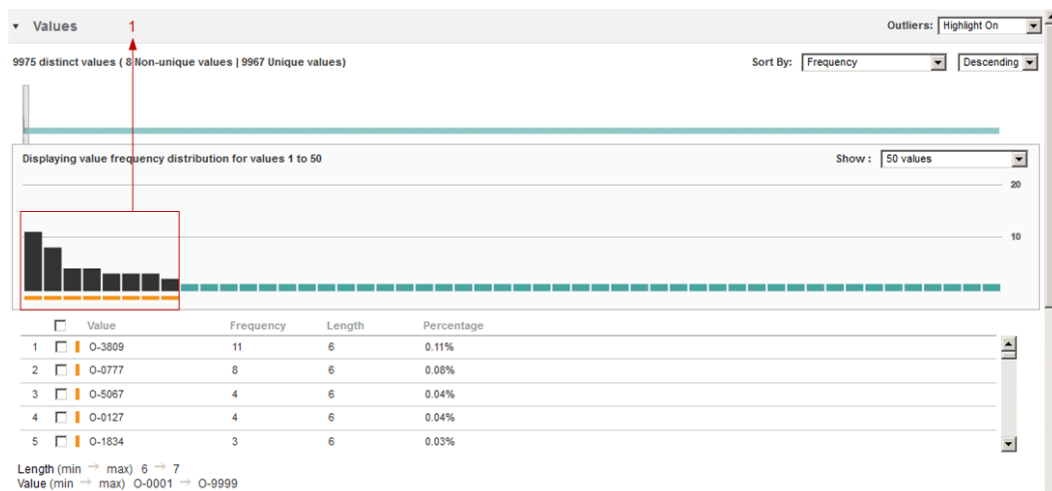
Outliers

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
'B' or 'b' or ' '	Represents a blank space.
'C' or 'c'	Represents any character.

Character	Description
'L' or 'l'	Represents any lowercase alphabetic character.
'T' or 't'	Represents a tab.
'U' or 'u'	Represents any uppercase alphabetic character.
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' or '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' or 'p'	Represents "(", the opening parenthesis.
'Q' or '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

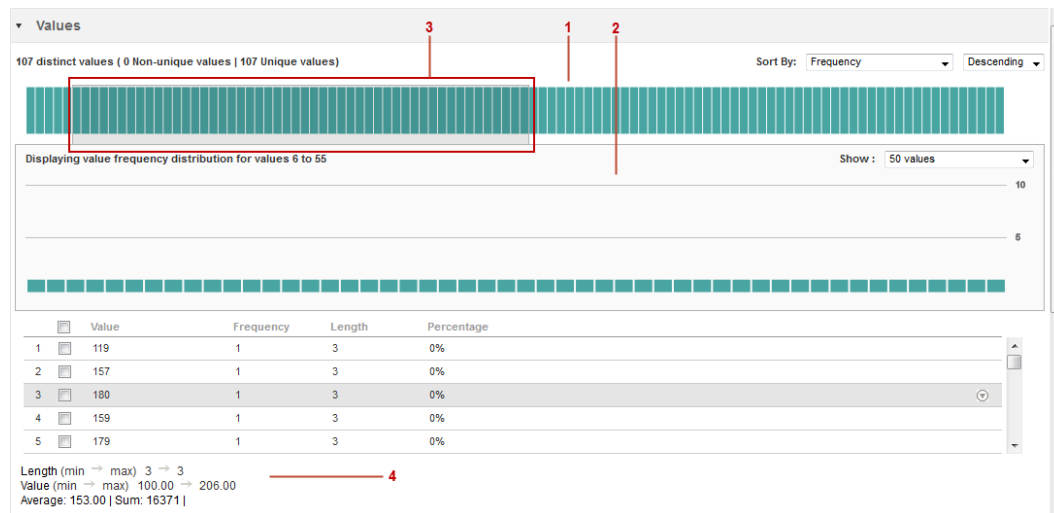
The column values in detailed view include values for a column and the frequency in which the values appear in the column.

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.

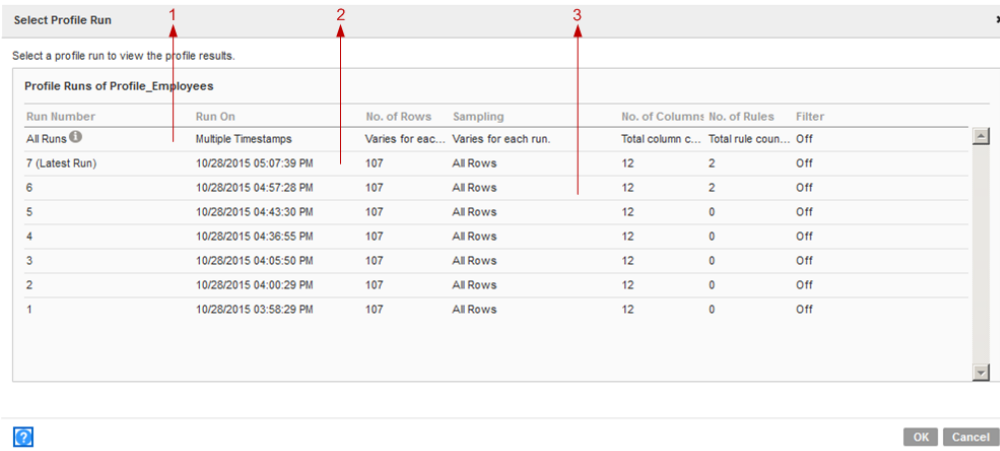
1. In the **Library** workspace, select the project or folder that contains the profile, or select the profile in the **Assets** pane.
2. Click **Actions > Open** to open the profile.

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

- 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.



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

- 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.

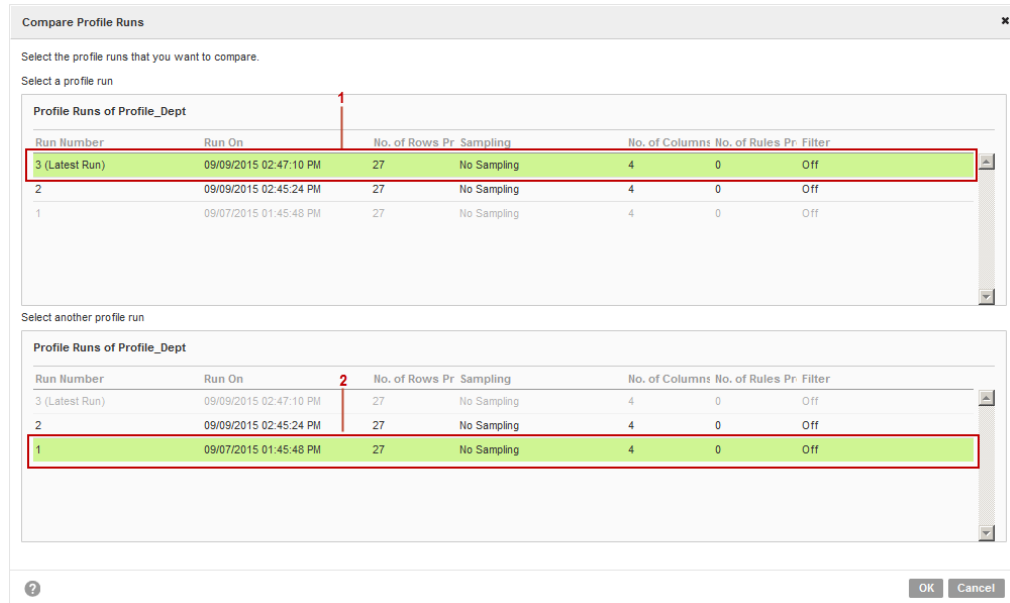
In the summary view, you can view the compare results for all the columns in both the profile runs.

Comparing Multiple Profile Results

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

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

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

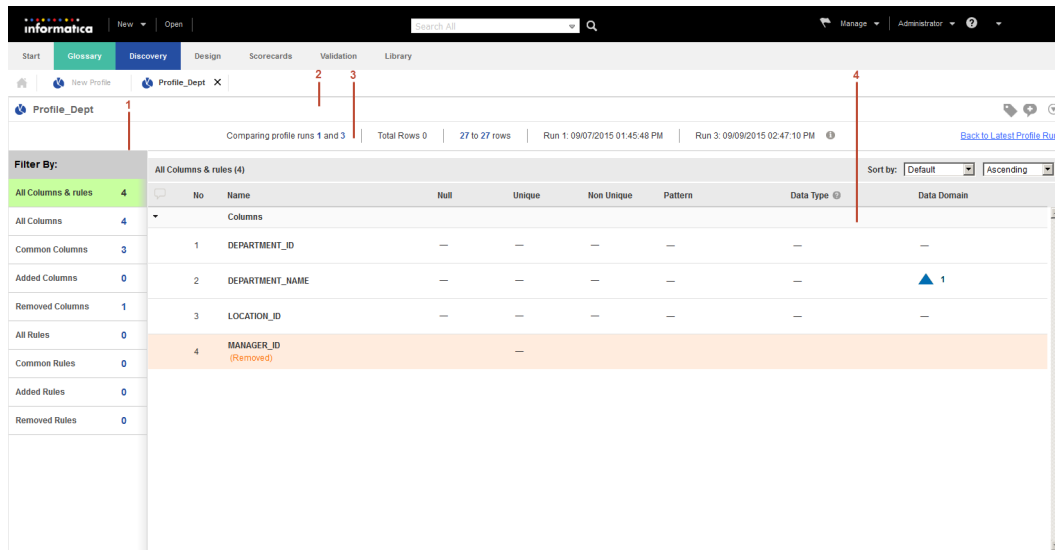


1. Run A. Choose a profile run as Run A.
 2. Run B. Choose a profile run as Run B.
- The **Compare Profile Runs** dialog box appears.
2. Select a profile from the **Run A** pane, and select another profile from the **Run B** pane.
 3. Click **OK**.
- The summary view displays a consolidated view of the 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.

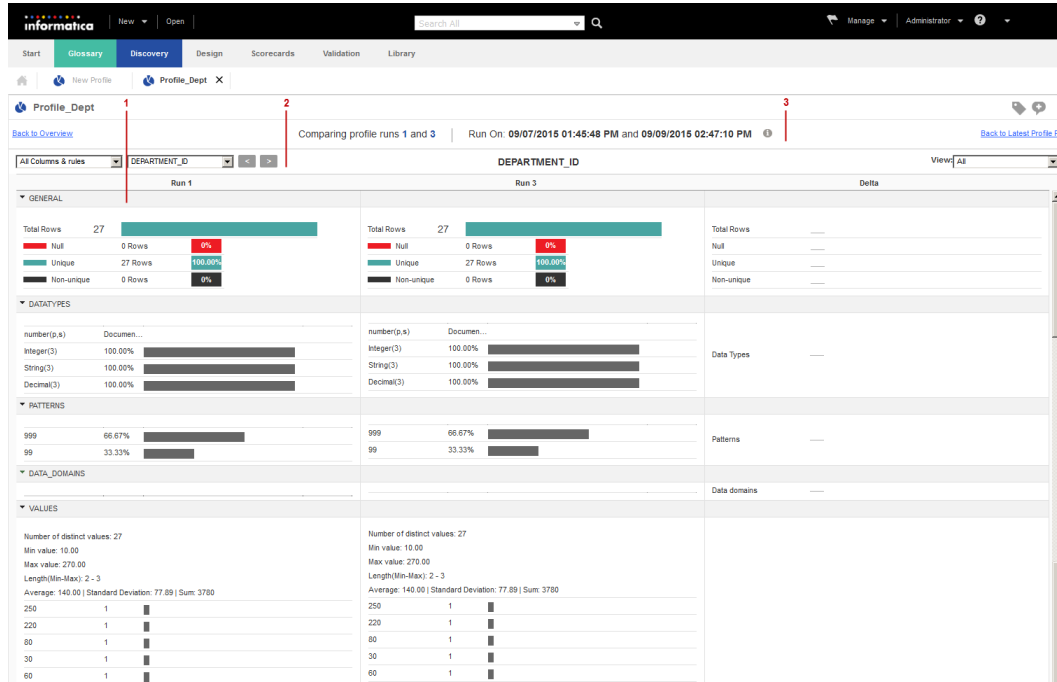
The summary view displays the profile results for all source columns and virtual columns by default.

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 domain 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 8

Scorecards in Informatica Analyst

This chapter includes the following topics:

- [Scorecards in Informatica Analyst Overview, 63](#)
- [Informatica Analyst Scorecard Process, 64](#)
- [Creating a Scorecard in Informatica Analyst, 65](#)
- [Adding Columns to an Existing Scorecard, 66](#)
- [Running a Scorecard, 67](#)
- [Viewing a Scorecard, 67](#)
- [Editing a Scorecard, 67](#)
- [Metrics, 68](#)
- [Metric Groups, 70](#)
- [Drilling Down on Columns, 71](#)
- [Trend Charts, 72](#)
- [Scorecard Export Files in Informatica Analyst, 75](#)
- [Scorecard Notifications, 75](#)
- [Scorecard Lineage, 78](#)

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.

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:

$$(M1 \times W1 + M2 \times W2) / (W1 + W2)$$

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.

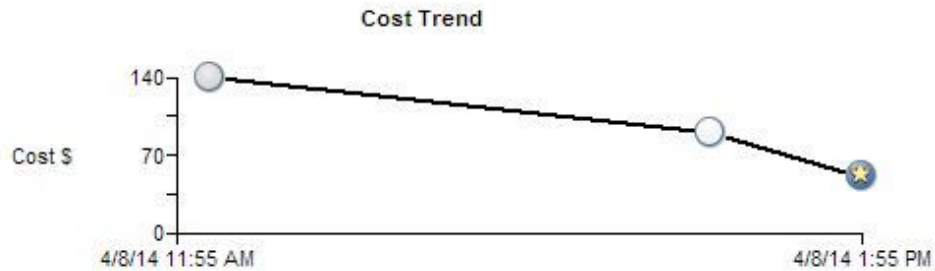
Cost Trend Chart

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

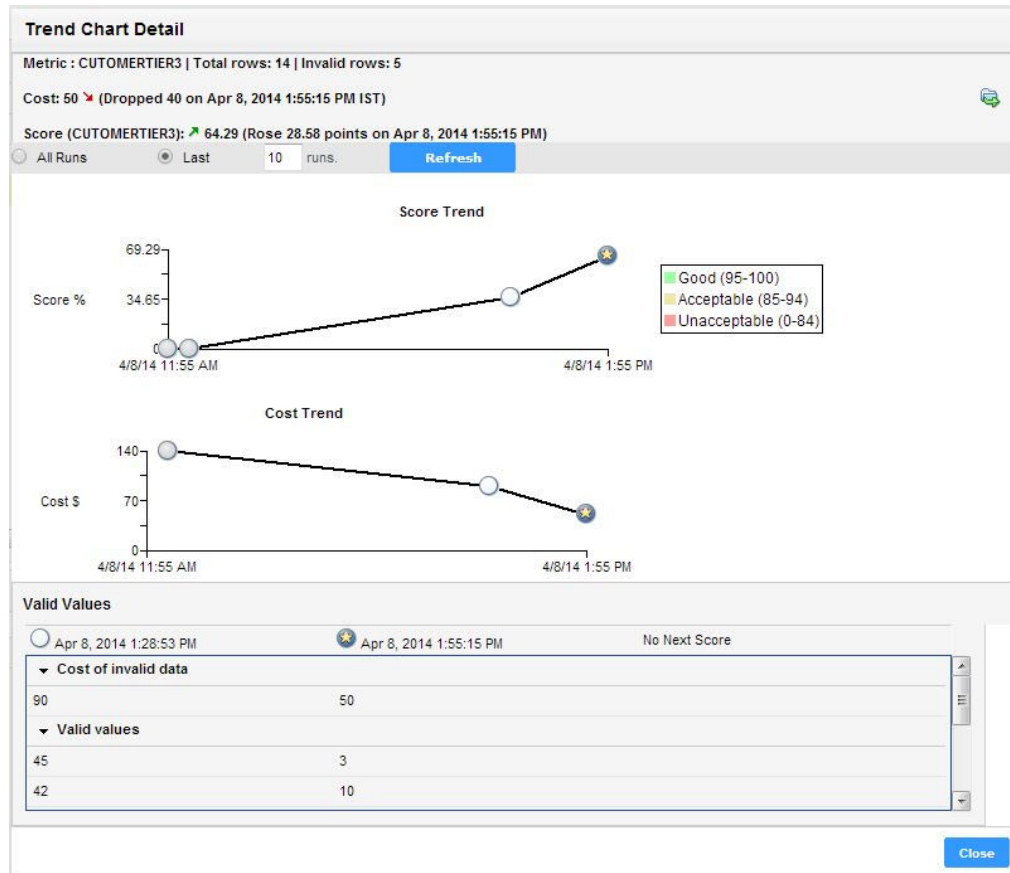
You can view trend charts for each metric to monitor how the score or cost of invalid data changes over time.

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.

- 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.

- Open a scorecard.
- Select a metric, and click **Actions > Show Trend Chart**.
The **Trend Chart Details** dialog box appears.
- Click the **Export Data** icon.
The **Export data to a file** dialog box appears.
- Enter a file name. Optionally, use the default file name.
The default file format is Microsoft Excel.
- Select the code page of the file.
- Click **OK**.

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.
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.

Tag	Description
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 tools 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 102](#).

Part III: Profiling with Informatica Developer

This part contains the following chapters:

- [Data Object Profiles, 80](#)
- [Column Profiles on Semi-structured Data Sources, 85](#)
- [Rules in Informatica Developer, 92](#)
- [Mapplet and Mapping Profiling, 94](#)
- [Column Profile Results in Informatica Developer, 96](#)
- [Scorecards in Informatica Developer, 101](#)

CHAPTER 9

Data Object Profiles

This chapter includes the following topics:

- [Column Profiles in Informatica Developer, 80](#)
- [Creating a Single Data Object Profile in Informatica Developer, 81](#)
- [Creating Multiple Data Object Profiles in Informatica Developer, 82](#)
- [Synchronizing a Flat File Data Object in Informatica Developer, 83](#)
- [Synchronizing a Relational Data Object in Informatica Developer, 84](#)

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 maplet 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.

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. You can choose a maximum of 2,147,483,647 rows.
Random Sample of	The random sample algorithm chooses the rows at random in the data object to run the profile on. You can choose a maximum of 2,147,483,647 rows.
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 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.

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** or **Hadoop** as the run-time environment. After you choose the Hadoop option, you can 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**.

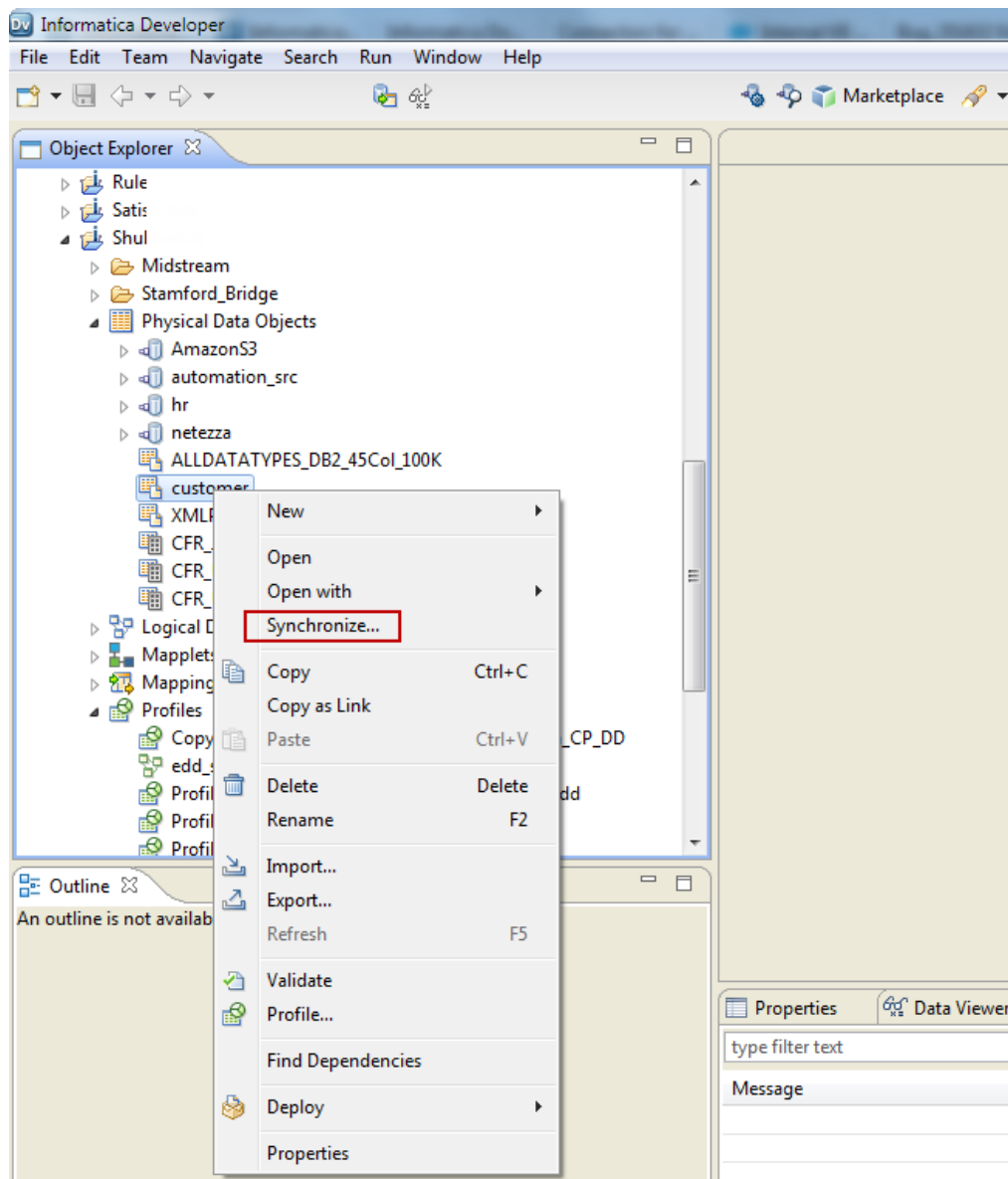
Note: Choose only the Native option to run the multiple data object profile. To run multiple data objects on the Blaze engine in the Hadoop run-time environment, you can choose the enterprise discovery profile.
9. Click **Finish**.
10. Optionally, enter prefix and suffix strings to add to the profile names.
11. Click **OK**.

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**.

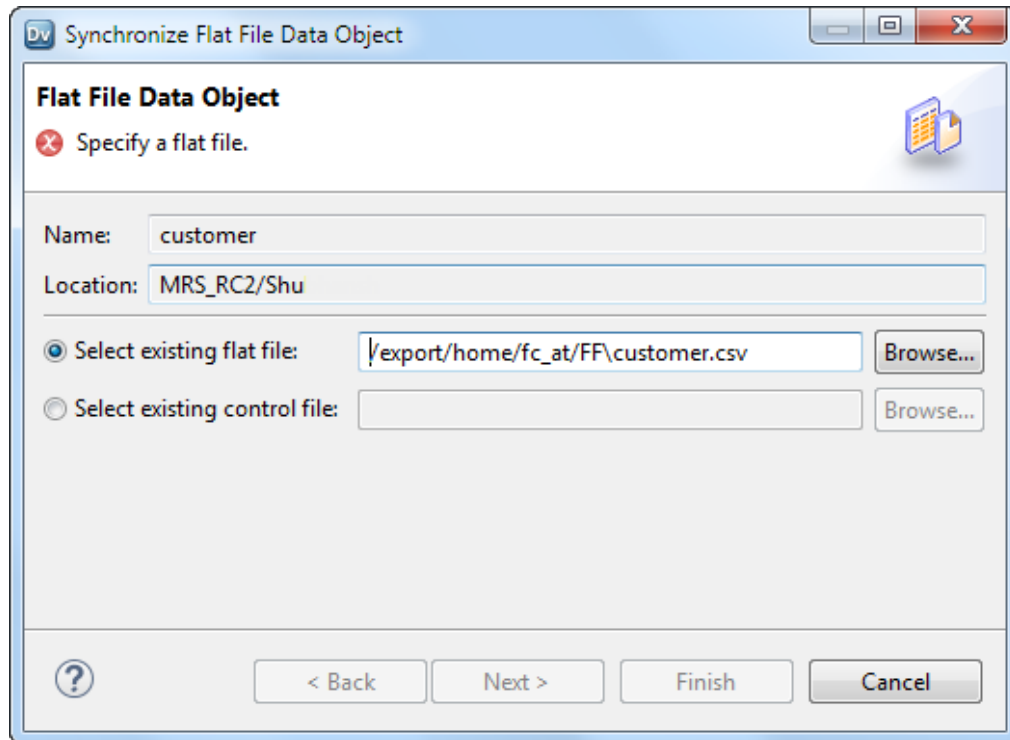
The following image shows the Synchronize option for a data object:



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.

CHAPTER 10

Column Profiles on Semi-structured Data Sources

This chapter includes the following topics:

- [Column Profiles on Semi-structured Data Sources Overview, 85](#)
- [JSON and XML Data Objects, 86](#)
- [Complex File Data Objects for Semi-Structured Data Sources in HDFS, 87](#)
- [Creating an HDFS Connection, 88](#)
- [Creating a Complex File Data Object from a JSON or XML File in HDFS, 88](#)
- [Creating a Complex File Data Object from an Avro or Parquet Data Source, 89](#)
- [Creating a Column Profile on a Semi-structured Data Source, 90](#)

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.

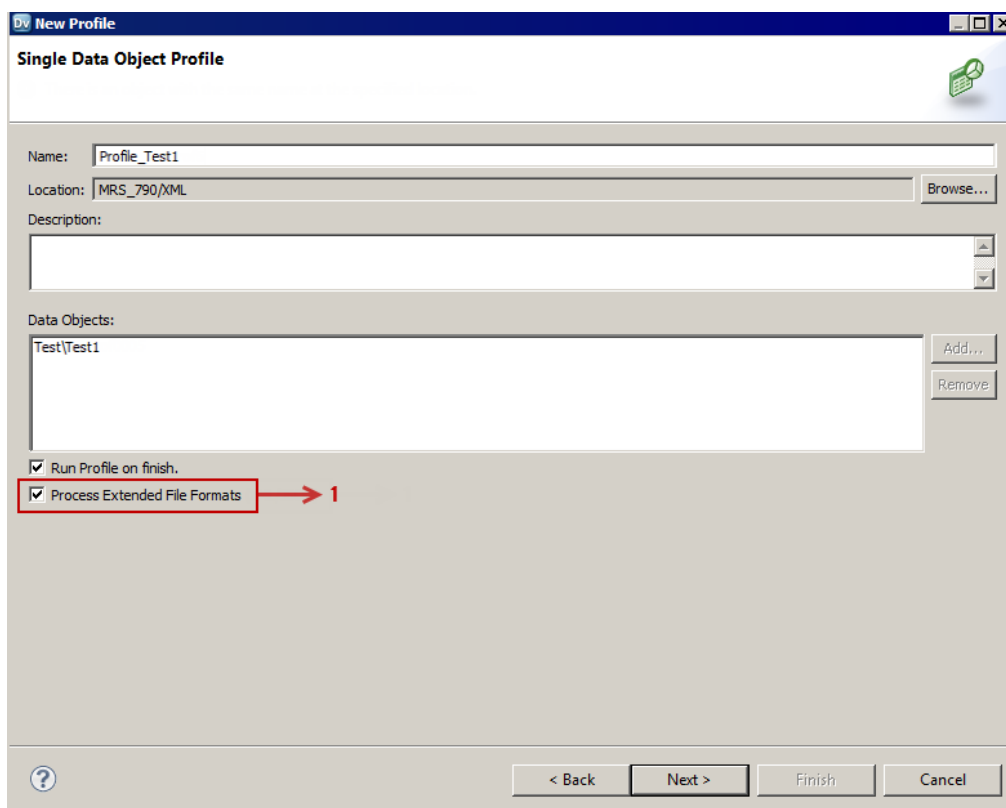
You can choose a folder with multiple Avro or multiple Parquet files to create a data object. After you create the data object, navigate to **Data Object Operations > Read > Advanced** tab, and verify whether the file path in the **File path** field points to the folder of the data sources in the Linux 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 11

Rules in Informatica Developer

This chapter includes the following topics:

- [Rules in Informatica Developer Overview, 92](#)
- [Creating a Rule in Informatica Developer, 93](#)
- [Applying a Rule in Informatica Developer, 93](#)

Rules in Informatica Developer 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 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.

Creating a Rule in Informatica Developer

You need to validate a maplet as a rule to create a rule in the Developer tool.

Create a maplet in the Developer tool.

1. Right-click the maplet editor.
2. Select **Validate As > Rule**.

Applying a Rule in Informatica Developer

You can add a rule to a saved column profile. You cannot add a rule to a profile configured for join analysis.

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 12

Mapplet and Mapping Profiling

This chapter includes the following topics:

- [Mapplet and Mapping Profiling Overview, 94](#)
- [Running a Profile on a Mapplet or Mapping Object, 94](#)
- [Comparing Profiles for Mapping or Mapplet Objects, 95](#)
- [Generating a Mapping from a Profile, 95](#)

Mapplet and Mapping Profiling Overview

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.

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.
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. You can view the profile summary and mapping log files to get more information about the tasks performed by the Developer tool.

Note: If you ran the profile using an operating system profile, the summary log appears in the log directory configured for the Data Integration Service and the mapping log appears in the log directory configured for the operating system 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.

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 13

Column Profile Results in Informatica Developer

This chapter includes the following topics:

- [Column Profile Results in Informatica Developer, 96](#)
- [Column Value Properties, 97](#)
- [Column Pattern Properties, 97](#)
- [Column Statistics Properties, 97](#)
- [Column Data Type Properties, 98](#)
- [Curation in Informatica Developer, 99](#)
- [Exporting Profile Results from Informatica Developer, 100](#)

Column Profile Results in Informatica Developer

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:

Profile Type	Profile Results
Column profile	<ul style="list-style-type: none">- 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	<ul style="list-style-type: none">- Inferred primary keys- Key violations
Functional dependency profile	<ul style="list-style-type: none">- Inferred functional dependencies- Functional dependency violations

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.
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	<p>Indicates the status of the data type. The statuses are Inferred, Approved, or Rejected.</p> <p>Inferred</p> <p>Indicates the data type of the column that the Developer tool inferred.</p> <p>Approved</p> <p>Indicates an approved data type for the column. When you approve a data type, you commit the data type to the Model repository.</p> <p>Rejected</p> <p>Indicates a rejected data type for the column.</p>

Curation 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 Datatypes

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 Types

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 14

Scorecards in Informatica Developer

This chapter includes the following topics:

- [Scorecards in Informatica Developer Overview, 101](#)
- [Creating a Scorecard, 101](#)
- [Exporting a Resource File for Scorecard Lineage, 102](#)
- [Viewing Scorecard Lineage from Informatica Developer, 102](#)

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.

INDEX

C

- column profile
 - drilldown [59](#)
 - Informatica Developer [80](#)
 - operating system profile [24](#)
 - options [15](#)
 - overview [14](#)
 - process [21](#)
- column profile results
 - Informatica Developer [96](#)
- column profile results in Analyst tool
 - column details [44](#), [58](#)
 - interface [43](#), [55](#), [57](#)
 - summary [42](#)
- creating a column profile
 - profiles [25](#)
- creating an expression rule
 - rules [33](#)
- curation
 - concepts [17](#)
 - Informatica Analyst [60](#)
 - Informatica Developer [99](#)
 - tasks [17](#)

D

- data object profiles
 - creating a single profile [81](#)
 - creating multiple profiles [82](#)

E

- export
 - scorecard lineage to XML [102](#)

F

- filters
 - overview [36](#)
- flat file data object
 - synchronizing [28](#)

I

- Informatica Analyst
 - column profile results [41](#), [53](#)
 - column profiles overview [20](#), [54](#)
 - lock and version management [24](#)
 - rules [31](#)
- Informatica Developer
 - rules [92](#)

M

- mapping object
 - running a profile [94](#)
- Mapplet and Mapping Profiling
 - Overview [94](#)

O

- outlier
 - detecting [49](#)

P

- predefined rules
 - process [32](#)
- profile
 - Avro or Parquets formats [87](#)
 - components [12](#)
 - XML and JSON formats [86](#), [87](#)
- profile results
 - approving data types [60](#)
 - approving data types in Informatica Developer [99](#)
 - column data types [47](#), [98](#)
 - column patterns [49](#)
 - column values [50](#)
 - detailed view [45](#)
 - drilling down [59](#)
 - Excel [61](#)
 - exporting [60](#)
 - exporting from Informatica Analyst [61](#)
 - exporting in Informatica Developer [100](#)
 - rejecting data types [60](#)
 - rejecting data types in the Developer tool [99](#)
 - summary [56](#), [58](#)
 - summary view [43](#)
- profiles
 - creating a column profile [25](#)
 - creating a filter [36](#)
 - editing a column profile [26](#)
 - editing a filter [39](#)
 - introduction [11](#)
 - running [27](#), [53](#), [54](#)
- profiling
 - lock and version management [15](#)
 - process [12](#)
 - tools [12](#)

R

- rules
 - applying a predefined rule [32](#)
 - applying in Informatica Developer [93](#)

rules (*continued*)
 creating an expression rule [33](#)
 creating an expression rule using rule specification [34](#)
 creating in Informatica Developer [93](#)
 expression [33](#)
 predefined [31](#)
run-time environment
 Analyst Tool [23](#)
 Hadoop [23, 24](#)

S

scorecard
 configuring global notification settings [77](#)
 configuring notifications [77](#)
scorecard lineage
 viewing from Informatica Developer [102](#)
 viewing in Informatica Analyst [78](#)
scorecard results
 export to Excel [75](#)
 exporting [75](#)
 exporting from Informatica Analyst [75](#)
scorecards
 adding columns to a scorecard [66](#)
 cost of invalid data [69](#)
 creating a metric group [70](#)
 defining thresholds [69](#)
 deleting a metric group [71](#)
 drilling down [71](#)
 editing [67](#)

scorecards (*continued*)
 editing a metric group [71](#)
 fixed cost [69](#)
 Informatica Analyst [63](#)
 Informatica Analyst process [64](#)
 Informatica Developer [101](#)
 metric groups [70](#)
 metric weights [68](#)
 metrics [68](#)
 moving scores [70](#)
 notifications [76](#)
 overview [16](#)
 running [67](#)
 trend chart [72](#)
 variable cost [69](#)
 viewing [67](#)
Sqoop configuration
 profiling [23](#)

T

table data object
 synchronizing [30](#)
trend charts
 cost [73](#)
 exporting from Informatica Analyst [74](#)
 score [72](#)
 viewing [73](#)