Informatica® Cloud Data Profiling
December 2022

Data Profiling

Informatica Cloud Data Profiling Data Profiling
December 2022
December 2022

Publication Date: 2023-01-10

Table of Contents

Preface

Use*Data Profiling* to learn how to create and run data profiling tasks, and view profile results. Learn how to compare columns and profile runs, export profile results, tune the performance of data profiling tasks, and troubleshoot errors in Data Profiling.

Informatica Resources

Informatica provides you with a range of product resources through the Informatica Network and other online portals. Use the resources to get the most from your Informatica products and solutions and to learn from other Informatica users and subject matter experts.

Informatica Network

The Informatica Network is the gateway to many resources, including the Informatica Knowledge Base and Informatica Global Customer Support. To enter the Informatica Network, visit https://network.informatica.com.

As an Informatica Network member, you have the following options:

- Search the Knowledge Base for product resources.
- View product availability information.
- Create and 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 find product resources such as how-to articles, best practices, video tutorials, and answers to frequently asked questions.

To search the Knowledge Base, visit https://search.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

Use the Informatica Documentation Portal to explore an extensive library of documentation for current and recent product releases. To explore the Documentation Portal, visit https://docs.informatica.com.

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

Informatica Product Availability Matrices

Product Availability Matrices (PAMs) indicate the versions of the operating systems, databases, and types of data sources and targets that a product release supports. You can browse the Informatica PAMs at https://network.informatica.com/community/informatica-network/product-availability-matrices.

Informatica Velocity

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

You can find 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 extend and enhance your Informatica implementations. Leverage any of the hundreds of solutions from Informatica developers and partners on the Marketplace to improve your productivity and speed up time to implementation on your projects. You can find the Informatica Marketplace at https://marketplace.informatica.com.

Informatica Global Customer Support

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

To find your local Informatica Global Customer Support telephone number, visit the Informatica website at the following link:
https://www.informatica.com/services-and-training/customer-success-services/contact-us.html.

To find online support resources on the Informatica Network, visit https://network.informatica.com and select the eSupport option.

CHAPTER 1

Data Profiling

You can use Informatica Intelligent Cloud Services Data Profiling to analyze data schemas, determine the quality of data across sources, and understand the completeness, conformity, and consistency of data in the data sources.

Use Data Profiling to create and run data profiling tasks and view profiling results. You can view other assets, such as rule specifications and dictionaries that you created using other services within the same organization. You can import and export assets within the organization.

Data profiling tasks

You can create data profiling tasks in Data Profiling. Create and run a data profiling task to determine the characteristics of columns in a source object, such as value frequency, patterns, and data types. Data profiling tasks are also called profiles.

You can create a profile for a source object after you create a connection to the data source. After you create the profile, run it to view the profile results.

You can add a filter to run the profile on filtered results. For example, to view county-specific sales, you can filter the Sales table based on a county and then run the profile. You can add a schedule to run the profile at regular intervals. You can add rule specification, cleanse, and verifier assets as rules to a profile. You have to create these assets in Data Quality.

When you run a profile on a source object, the results include the following column statistics:

- Number of distinct, non-distinct, and null values
- Percentage of distinct, non-distinct, null, zero, and blank values
- Documented and inferred data types
- Number of patterns
- Percentage of top pattern
- Maximum and minimum length of values
- Maximum and minimum values
- Average, sum, and standard deviation for numeric data types
- Value frequencies
- Outliers

After you run a profile, you can perform the following actions:

- Drill down on value, data type, and pattern to view drilldown results.

- View historical and latest profile results
- Create and run queries to view source rows that have data quality issues.
- Compare multiple columns in a profile run
- Compare two profile runs to analyze the statistics
- Export profile results to a Microsoft Excel file
- Monitor profile jobs in Data Profiling, Monitor, or Operational Insights.

Prerequisites

Before you can use Data Profiling to define and run data profiling tasks, make sure that you complete the following prerequisites.

Create users and user groups

Create users and user groups in Administrator. A user is an individual Informatica Intelligent Cloud Services account that allows secure access to an organization. A user can perform tasks and access assets based on the roles that are assigned to the user. You can assign roles directly to the user or to a group that the user is a member of. Administrators can create and configure user accounts for the organization.

For more information about creating user and user groups, see *Administrator* in the Administrator help.

Assign permissions and privileges

Permissions determine the access rights that a user has for a Secure Agent, Secure Agent group, connection, schedule, or asset. To configure permissions on an object, you need privileges that you can assign to users or groups.

You need one or more of the following permissions to access objects or assets:

- Create
- Read
- Update
- Delete
- Run

For more information about permissions, see *Asset Management*.

A role is a collection of privileges that you can assign to users and groups. To ensure that every user can access assets and perform tasks in your organization, assign at least one role to each user or user group. A role defines the privileges for different types of assets and service features. For example, users with the Designer role can create, read, update, delete, and set permissions on data profiling assets. However, they have no access to certain Administrator service features, such as sub-organizations and audit logs.

Administrators can configure and assign roles for the organization. If the user has a system-defined role, you do not have to set privileges or asset permissions because the system-defined roles include necessary privileges and permissions.

Data Profiling users can be assigned the Admin, Data Integration Data Previewer Designer, Monitor, and Operator roles.

For more information about creating user and user groups, see *Administrator* in the Administrator help.

The following table lists the roles, asset permissions, and features that you require for Data Profiling:

Role	Create	Read	Update	Delete	Run	Set Permissions	Features
Admin	X	X	X	X	X	X	Compare Columns Compare Data Profiling Runs Data Profiling results - view Drill down Export Data Profiling Results Manage Rules Query - Create Query - Submit Operational Insights - view Sensitive Data - view
Data Integration Data Previewer*		X					Data Profiling - Data Preview
Designer	X	X	X	X	X	X	Compare Columns Compare Data Profiling Runs Data Profiling results - view Drill down Export Data Profiling Results Manage Rules Query - Create Query - Submit Sensitive Data - view
Monitor		X					Compare Columns Compare Data Profiling Runs Data Profiling results - view
Operator		X					Compare Columns Compare Data Profiling Runs Data Profiling results - view Operational Insights - view
*In addition to the Data Integration Data Previewer role, you also need the Admin or Designer role to view the Data Preview tab.							

The following table lists the privileges that are available for Data Profiling:

Privilege	Description
Data Profiling	Create, read, update, delete, run, and set permissions for a data profiling task.
Data Profiling - Compare Columns	Compare columns in a profile run.
Data Profiling - Compare Data Profiling Runs	Compare multiple profile runs.

Privilege	Description
Data Profiling - Data Profiling Results - View	- View the profiling results for a data profiling task for any user including the user who created the data profiling task. - View the valid and invalid rows in the Data Governance and Catalog scorecard using the **Preview of Successful Rows** and **Preview of Unsuccessful Rows**.
Data Profiling - Drill down	View and select the drill-down option when you create a data profiling task.
Data Profiling - Export Data Profiling Results	Export the profiling results to a Microsoft Excel file.
Data Profiling - Manage Rules	Add or delete rules for a data profiling task.
Data Profiling - Query - Create	Create a query.
Data Profiling - Query - Submit	Run a query and view query results.
Data Profiling - Data Preview	View source object data in the **Data Preview** area.
Data Profiling Sensitive Data - view	Hide sensitive information for a particular user role. When the Sensitive Data - view privilege is configured, you cannot view the minimum value, maximum value, and most frequent values information in the **compare column** tab.

Create connections

When you create a data profiling task, you need a connection to the source object. You can create connections in Administrator.

The following table lists the connections and the source objects that Data Profiling supports:

Connections	Supported source object
Flat file	Flat file
Google BigQuery V2	Google BigQuery
Microsoft SQL Server	Microsoft SQL Server Azure SQL Database
Microsoft Azure Synapse SQL	Azure Synapse SQL
Oracle	Oracle
Salesforce	Salesforce Sales Cloud Salesforce Service Cloud Applications on Force.com Salesforce Verticals which include Salesforce Health Cloud and Salesforce Financial Cloud.
Amazon S3 v2	Amazon S3
Azure Data Lake Store Gen2	Azure Data Lake Store

Connections	Supported source object
Amazon Redshift V2	Amazon Redshift
Snowflake Cloud Data Warehouse V2	Snowflake Cloud Data Warehouse
ODBC	Applicable for the data sources that are not supported with native driver and have a compliant ODBC driver.
Oracle Cloud Object Storage	Oracle Cloud Object Storage

You can run a profile on Databricks Delta tables using an ODBC driver. You can download the Databricks ODBC driver here: https://databricks.com/spark/odbc-drivers-download. For more information about how to run a profile on Databricks Delta tables using Azure Databricks with an ODBC connection, see the Informatica How-To-Library article: *How to run a profile on Databricks Delta tables using Azure Databricks with ODBC connection*.

For more information about creating connections, see Getting Started and Connections.

Flat file connection

To access flat files, you need to create a flat file connection to the source object.

Configure the following flat file connection properties to create and run a data profiling task on a flat file source object:

Property	Value
Runtime Environment	Choose the active Secure Agent.
Directory	Enter the full directory or click **Browse** to locate and select the directory.
Date Format	Enter the date format for date fields in the flat file. Default date format is: MM/dd/yyyy HH:mm:ss.
Code Page	Choose **UTF-8**.

For more information about the flat file connection properties, see the help for the flat file connector.

Google BigQuery V2 connection

To access a Google BigQuery source object, you need to create a Google BigQuery V2 connection to the source object. Make sure that you have the Google BigQuery V2(Connector) license to access the source object.

Configure the following Google BigQuery V2 connection properties to create and run a data profiling task on a Google BigQuery source object:

Property	Value
Runtime Environment	Choose the active Secure Agent.
Service Account ID	Enter the client_email value present in the JSON file that you download after you create a service account.
Service Account Key	Enter the private_key value present in the JSON file that you download after you create a service account.
Project ID	Enter the id of the project in the Google service account that contains the dataset that you want to connect to.
Connection mode	Choose the mode that you want to use to read data from or write data to Google BigQuery.

For more information about the Google BigQuery V2 connection properties, see the *Google BigQuery V2 connection properties* in the *Google BigQuery Connectors* guide.

Microsoft SQL Server connection

To access a Microsoft SQL Server source object and a Microsoft SQL Server source object on Azure, you need to create a Microsoft SQL Server connection to the source object.

Microsoft SQL Server

Configure the following Microsoft SQL Server connection properties to create and run a data profiling task on a Microsoft SQL Server source object:

Property	Value
Runtime Environment	Choose the active Secure Agent.
SQL Server Version	Enter the Microsoft SQL Server database version.
Authentication Mode	Choose **SQL Server**.
User Name	Enter the user name for the database login.
Password	Enter the password for the database login.
Host	Enter the name of the machine that hosts the database server.
Port	Enter the network port number used to connect to the database server. Default is 1433.
Database Name	Enter the database name for the Microsoft SQL Server target.

Property	Value
Schema	Enter the schema used for the target connection.
Code Page	Choose **UTF-8**.

Azure SQL Database

Configure the following Microsoft SQL Server connection properties to create and run a data profiling task on a Azure SQL Database source object:

Property	Value
Runtime Environment	Choose the active Secure Agent.
SQL Server Version	Enter the Azure SQL Server database version.
Authentication Mode	Choose **SQL Server**.
User Name	Enter the user name for the database login.
Password	Enter the password for the database login.
Host	Enter the name of the machine that hosts the database server.
Port	Enter the network port number used to connect to the database server. Default is 1433.
Database Name	Enter the database name for the Microsoft SQL Server target.
Schema	Enter the schema used for the target connection.
Code Page	Choose **UTF-8**.
Encryption Method	Enter **SSL**.
Crypto Protocol Version	Enter **TLSv1.2**.

For more information about the Microsoft SQL Server connection properties, see the help for the Microsoft SQL Server connector.

Oracle connection

To access an Oracle source object, you need to create an Oracle connection to the source object. Make sure that you have the Oracle license to access the source object.

Configure the following Oracle connection properties to create and run a data profiling task on an Oracle source object:

Property	Value
Runtime Environment	Choose the active Secure Agent.
User Name	Enter the user name of the database login.
Password	Enter the password of the database login.
Host	Enter the name of the machine that hosts the database server.
Port	Enter the network port number used to connect to the database server. Default is 1521.
Service Name	Enter the service name or System ID (SID) that uniquely identifies the Oracle database.
Schema	Enter the schema name. If you do not specify a schema, Data Profiling uses the default schema.
Code Page	Choose **UTF-8**.

For more information about the Oracle connection properties, see the help for the Oracle connector.

Oracle Cloud Object Storage

To access an Oracle Cloud Object Storage source object, you must create an Oracle Cloud Object Storage connection to the source object.

Configure the following Oracle Cloud Object Storage connection properties to create and run a data profiling task on an Oracle Cloud Object Storage source object:

Property	Value
Runtime Environment	Name of the runtime environment where you want to run the tasks.
Authentication Type	Authentication type to connect to Oracle Cloud Object Storage to stage the files. Select one of the following options: - Simple Authentication. API key-based authentication. - ConfigFile Authentication. Identity credential details are provided through a configuration file.
User	The Oracle Cloud Identifier (OCID) of the user for whom the key pair is added.
Finger Print	Fingerprint of the public key.
Tenancy	Oracle Cloud Identifier (OCID) of the tenancy, that is the globally unique name of the OCI account.

Property	Value
Config File Location	Location of configuration file on the Secure Agent machine. Enter the absolute path. If you do not enter any value, `<system_default_location>/.oci/config` is used to retrieve the configuration file.
Private Key File Location	Location of the private key file in .PEM format on the Secure Agent machine.
Profile Name	Required if you use the `ConfigFile` for authentication. Name of the profile in the configuration file that you want to use. Default is `DEFAULT`.
Bucket Name	The Oracle Cloud Storage bucket name. This bucket contains the objects and files.
Folder Path	The path to the folder under the specified Oracle Cloud Storage bucket. For example, `bucket/Dir_1/Dir_2/FileName.txt`. Here, `Dir_1/Dir_2` is the folder path.
Region	Oracle Cloud Object Storage region where the bucket exists. Select the Oracle Cloud Object Storage region from the list.

For more information about the Oracle Cloud Object Storage connection properties, see the help for the Oracle Cloud Object Storage connector.

Salesforce connection

To access a Salesforce source object, you need to create a Salesforce connection to the source object. Make sure that you have the Salesforce(connector) license to access the source object. When you set up a Salesforce connection, you can select Standard or OAuth connection type.

Data Profiling supports the following Salesforce source objects:

- Salesforce Sales Cloud
- Salesforce Service Cloud
- Salesforce Verticals:
  - Salesforce Financial Cloud
  - Salesforce Health Cloud
- Applications on Force.com

Standard connection type

Configure the following Salesforce connection properties to create and run a data profiling task on a Salesforce source object:

Property	Value
Runtime Environment	Choose the active Secure Agent.
User Name	Enter the user name for the Salesforce account.

Property	Value
Password	Enter the password for the Salesforce account.
Security Token	Enter the security token generated from the Salesforce application. To generate a security token in the Salesforce application, click **Reset My Security Token** in the **Setup** > **Personal Setup** > **My Personal Information** section. You do not need to generate the security token every time if you add the Informatica Cloud IP address ranges: 209.34.91.0-255, 206.80.52.0-255, 206.80.61.0-255, and 209.34.80.0-255 to the **Trusted IP Ranges** field. To add the Informatica Cloud IP address ranges, navigate to the **Setup** > **Security Controls** > **Network Access** section in your Salesforce application.
Service URL	Enter the URL of the Salesforce service. Maximum length is 100 characters.

OAuth connection type

Configure the following Salesforce connection properties to create and run a data profiling task on a Salesforce source object:

Property	Value
Runtime Environment	Choose the active Secure Agent.
OAuth Consumer Key	Enter the consumer key that you get from Salesforce, which is required to generate a valid refresh token.
OAuth Consumer Secret	Enter the consumer secret that you get from Salesforce, which is required to generate a valid refresh token.
OAuth Refresh Token	Enter the refresh token generated in Salesforce using the consumer key and consumer secret.
Service URL	Enter the URL of the Salesforce service endpoint. Maximum length is 100 characters.

For more information about the Salesforce connection properties, see the help for the Salesforce connector.

Amazon S3

To access an Amazon S3 source object, you need to create a Amazon S3 v2 connection to the source object.

Configure the following Amazon S3 v2 connection properties to create and run a data profiling task on a Amazon S3 source object:

Property	Value
Runtime Environment	Name of the runtime environment where you want to run the tasks. Specify a Secure Agent, Hosted Agent, or serverless runtime environment.
Folder Path	Bucket name or complete folder path to the Amazon S3 objects. Do not use a slash at the end of the folder path. For example, `<bucket name>/<my folder name>`.

For more information about the Amazon S3 v2 connection properties, see the help for the Amazon S3 V2 connector.

Azure Data Lake Store Gen2

To access an Azure Data Lake Store Gen2 source object, you need to create a Azure Data Lake Store connection to the source object.

Configure the following Azure Data Lake Store Gen2 connection properties to create and run a data profiling task on a Azure Data Lake Store Gen2 source object:

Property	Value
Runtime Environment	The name of the runtime environment where you want to run the tasks. Specify a Secure Agent, Hosted Agent, or serverless runtime environment.
AccountName	Microsoft Azure Data Lake Storage Gen2 account name or the service name.
Client ID	The ID of your application to complete the OAuth Authentication in the Azure Active Directory (AD).
Client Secret	The client secret key to complete the OAuth Authentication in the Azure AD.
Tenant ID	The Directory ID of the Azure AD.
File System Name	The name of an existing file system in the Microsoft Azure Data Lake Storage Gen2 account.

For more information about the Azure Data Lake Store Gen2 connection properties, see the help for the Azure Data Lake Store Gen2 connector.

Amazon Redshift V2

To access an Amazon Redshift source object, you must create an Amazon Redshift V2 connection to the source object. You must use the Amazon Redshift V2 native driver connection instead of the ODBC driver connection.

Configure the following Amazon Redshift V2 connection properties to create and run a data profiling task on an Amazon Redshift source object:

Property	Value
Runtime Environment	Name of the runtime environment where you want to run the tasks. Specify a Secure Agent, Hosted Agent, or serverless runtime environment. **Note:** You cannot run a database ingestion task on a Hosted Agent or serverless runtime environment.
Username	Enter the user name for the Amazon Redshift account.
Password	Enter the password for the Amazon Redshift account.
Access Key ID	Access key to access the Amazon S3 bucket. Provide the access key value based on the following authentication methods: - Basic authentication: Provide the actual access key value. - [1] IAM authentication: Do not provide the access key value. - [1] Temporary security credentials via assume role: Provide access key of an IAM user who has no permissions to access the Amazon S3 bucket. - [1] Assume role for EC2: Do not provide the access key value. If you want to use the connection for a database ingestion task, you must use the basic authentication method to provide the access key value.

Property	Value
Secret Access Key	Secret access key to access the Amazon S3 bucket. The secret key is associated with the access key and uniquely identifies the account. Provide the access key value based on the following authentication methods: - Basic authentication: provide the actual access secret value. - ¹ IAM authentication: do not provide the access secret value. - Temporary security credentials via assume role: provide access secret of an IAM user who has no permissions to access the Amazon S3 bucket. - ¹ Assume role for EC2: do not provide the access secret value. If you want to use the connection for a database ingestion task, you must provide the actual access secret value.
¹ IAM Role ARN	The Amazon Resource Number (ARN) of the IAM role assumed by the user to use the dynamically generated temporary security credentials. Set the value of this property if you want to use the temporary security credentials to access the AWS resources.You cannot use the temporary security credentials in streaming ingestion tasks. For more information about how to obtain the ARN of the IAM role, see the AWS documentation.
¹ External Id	Optional. Specify the external ID for a more secure access to the Amazon S3 bucket when the Amazon S3 bucket is in a different AWS account.
¹ Use EC2 Role to Assume Role	Optional. Select the check box to enable the EC2 role to assume another IAM role specified in the IAM Role ARN option. **Note:** The EC2 role must have a policy attached with a permission to assume an IAM role from the same or different account. By default, the Use EC2 Role to Assume Role check box is not selected.
¹ Master Symmetric Key	Optional. Provide a 256-bit AES encryption key in the Base64 format when you enable client-side encryption. You can generate a key using a third-party tool.
JDBC URL	The URL of the Amazon Redshift V2 connection. Enter the JDBC URL in the following format: `jdbc:redshift://<amazon_redshift_host>:<port_number>/<database_name>`

Property	Value
[1] Cluster Region	Optional. The AWS cluster region in which the bucket you want to access resides. Select a cluster region if you choose to provide a custom JDBC URL that does not contain a cluster region name in the JDBC URL connection property.If you specify a cluster region in both Cluster Region and JDBC URL connection properties, the Secure Agent ignores the cluster region that you specify in the JDBC URL connection property.To use the cluster region name that you specify in the JDBC URL connection property, select None as the cluster region in this property.Select one of the following cluster regions: - None - Asia Pacific(Mumbai) - Asia Pacific(Seoul) - Asia Pacific(Singapore) - Asia Pacific(Sydney) - Asia Pacific(Tokyo) - Asia Pacific(Hong Kong) - AWS GovCloud (US) - AWS GovCloud (US-East) - Canada(Central) - China(Bejing) - China(Ningxia) - EU(Ireland) - EU(Frankfurt) - EU(Paris) - EU(Stockholm) - South America(Sao Paulo) - Middle East(Bahrain) - US East(N. Virginia) - US East(Ohio) - US West(N. California) - US West(Oregon) Default is None . You can only read data from or write data to the cluster regions supported by AWS SDK used by the connector.
[1] Customer Master Key ID	Optional. Specify the customer master key ID generated by AWS Key Management Service (AWS KMS) or the ARN of your custom key for cross-account access. **Note:** Cross-account access is not applicable to an advanced cluster. You must generate the customer master key ID for the same region where your Amazon S3 bucket resides. You can either specify the customer-generated customer master key ID or the default customer master key ID.
[1]Does not apply to version 2021.07.M	

For more information about the Amazon Redshift V2 connection properties, see the help for the Amazon Redshift V2 connector.

Snowflake Cloud Data Warehouse V2

To access a Snowflake source object, you must create a Snowflake Cloud Data Warehouse V2 connection to the source object. You must use the Snowflake Cloud Data Warehouse V2 native driver connection instead of the ODBC driver connection.

Note: Before you run a profile on a Snowflake source object, you must perform the steps listed in the Increase the Java heap size for the Snowflake Cloud Data Warehouse V2 connection on page 106 section of the *Troubleshooting* chapter.

Configure the following Snowflake connection properties to create and run a data profiling task on a Snowflake source object:

Property	Value
Runtime Environment	Choose an active Secure Agent with a package.
Authentication	Select the authentication method that the connector must use to log in to Snowflake. Default is Standard.
Username	Enter the user name for the Snowflake account.
Password	Enter the password for the Snowflake account.
Account	Name of the Snowflake account. In the Snowflake URL, your account name is the first segment in the domain. For example, if `123abc` is your account name, the URL must be in the following format: `https://123abc.snowflakecomputing.com`
Warehouse	Name of the Snowflake warehouse.
Role	Enter the Snowflake user role name.
Additional JDBC URL Parameters:	Optional. The additional JDBC connection parameters. Enter one or more JDBC connection parameters in the following format: `<param1>=<value>&<param2>=<value>&<param3>=<value>....` For example, `user=jon&warehouse=mywh&db=mydb&schema=public` To override the database and schema name used to create temporary tables in Snowflake, enter the database and schema name in the following format: `ProcessConnDB=<DB name>&ProcessConnSchema=<schema_name>` To view only the specified database and schema while importing a Snowflake table, specify the database and schema name in the following format: `db=<database_name>&schema=<schema_name>` To access Snowflake through Okta SSO authentication, enter the web-based IdP implementing SAML 2.0 protocol in the following format: `authenticator=https://<Your_Okta_Account_Name>.okta.com` **Note:** Microsoft Active Directory Federation Services is not supported. For more information about configuring Okta authentication, see the following website:https://docs.snowflake.com/en/user-guide/admin-security-fed-auth-configure-snowflake.html To load data from Google Cloud Storage to Snowflake for pushdown optimization, enter the Cloud Storage Integration name created for the Google Cloud Storage bucket in Snowflake in the following format: `storage_integration=<Storage Integration name>` For example, if the storage integration name you created in Snowflake for the Google Cloud Storage bucket is abc_int_ef, you must specify the integration name in uppercase. For example, storage_integration=ABS_INT_EF. **Note:** Verify that there is no space before and after the equal sign (=) when you add the parameters.

For more information about the Snowflake Cloud Data Warehouse V2 connection properties, see the help for the Snowflake Cloud Data Warehouse V2 connector.

Microsoft Azure Synapse SQL

To access a Microsoft Azure Synapse SQL source object, you must create a Microsoft Azure Synapse SQL connection to the source object. You must use the Microsoft Azure Synapse SQL native driver connection instead of the ODBC driver connection.

Configure the following Microsoft Azure Synapse SQL connection properties to create and run a data profiling task on a Microsoft Azure Synapse SQL source object:

Property	Value
Runtime Environment	The name of the runtime environment where you want to run the tasks. Specify a Secure Agent, Hosted Agent, or serverless runtime environment.

Azure DW JDBC URL	Microsoft Azure Synapse SQL JDBC connection string.
	Example for Microsoft SQL Server authentication:
	`jdbc:sqlserver://<Server>.database.windows.net:1433;database=<Database>`
	Example for Azure Active Directory (AAD) authentication:
	`jdbc:sqlserver://<Server>.database.windows.net:1433;database=<Database>;encrypt=true;trustServerCertificate=false;hostNameInCertificate=*.database.windows.net;loginTimeout=30;Authentication=ActiveDirectoryPassword;`

Azure DW JDBC Username	User name to connect to the Microsoft Azure Synapse SQL account. Provide AAD user name for AAD authentication.
Azure DW JDBC Password	Password to connect to the Microsoft Azure Synapse SQL account.
Azure DW Schema Name	Name of the schema in Microsoft Azure Synapse SQL.
Azure Storage Type	Type of Azure storage to stage the files. You can select any of the following storage type: - Azure Blob. Default. To use Microsoft Azure Blob Storage to stage the files. - ADLS Gen2. To use Microsoft Azure Data Lake Storage Gen2 as storage to stage the files.
Authentication Type	Authentication type to connect to Azure storage to stage the files.Select one of the following options: - Shared Key Authentication . Select to use the account name and account key to connect to Microsoft Azure Blob Storage or Microsoft Azure Data Lake Storage Gen2. - Service Principal Authentication . Applicable to Microsoft Azure Data Lake Storage Gen2. To use Service Principal authentication, you must register an application in the Azure Active Directory, generate a client secret, and then assign the Storage Blob Contributor role to the application.
Azure Blob Account Name	Applicable to Shared Key Authentication for Microsoft Azure Blob Storage.Name of the Microsoft Azure Blob Storage account to stage the files.
Azure Blob Account Key	Applicable to Shared Key Authentication for Microsoft Azure Blob Storage.Microsoft Azure Blob Storage access key to stage the files.

Property	Value
ADLS Gen2 Storage Account Name	Applicable to Shared Key Authentication and Service Principal Authentication for Microsoft Azure Data Lake Storage Gen2.Name of the Microsoft Azure Data Lake Storage Gen2 account to stage the files.
ADLS Gen2 Account Key	Applicable to Shared Key Authentication for Microsoft Azure Data Lake Storage Gen2.Microsoft Azure Data Lake Storage Gen2 access key to stage the files.
Client ID	Applicable to Service Principal Authentication for Microsoft Azure Data Lake Storage Gen2.The application ID or client ID for your application registered in the Azure Active Directory.
Client Secret	Applicable to Service Principal Authentication for Microsoft Azure Data Lake Storage Gen2.The client secret for your application.
Tenant ID	Applicable to Service Principal Authentication for Microsoft Azure Data Lake Storage Gen2.The directory ID or tenant ID for your application.
Blob End-point	Type of Microsoft Azure endpoints. You can select any of the following endpoints: - `core.windows.net`. Default. - `core.usgovcloudapi.net` . To select the Azure Government endpoints.
VNet Rule	Enable to connect to a Microsoft Azure Synapse SQL endpoint residing in a virtual network (VNet).When you use a serverless runtime environment, you cannot connect to a Microsoft Azure Synapse SQL endpoint residing in a virtual network.

For more information about the Microsoft Azure Synapse SQL connection properties, see the help for the Microsoft Azure Synapse SQL connector.

Create projects and folders

You can manage projects, and the assets and folders within them, on the **Explore** page. You can create projects and folders to organize the data.

For more information about projects and folders, see *Asset Management*.

Data profiling REST API

You can use the Data Profiling REST API to interact with the Data Profiling Service through API calls. You can use the REST API to perform tasks and get details for your organization. For example, you can perform tasks such as create, delete, and update queries and profiles.

To use the Data Profiling REST API, you need a valid Data Profiling Service login and an understanding of REST API guidelines.
Informatica Intelligent Cloud Services supports the platform REST API version 2 and version 3 resources, and service-specific resources.

For more information about Data Profiling REST API, see the Getting Started with Cloud Data Profiling REST API guide.

CHAPTER 2

Profiles

You can create a profile for a source object. You can add rules, filters, and schedules to the profile. You can also configure advanced options for the profile. Make sure that the prerequisites are met before you create and run a profile.

Profile definition

On the **Definition** tab of a profile, you can configure asset and source details. You can also select the columns and choose a filter for the profile run.

Asset Details

You can enter a name and choose a location for the profile.

The following table lists the options that you can configure in the **Asset Details** area:

Option	Description
Name	Enter a name for the profile. The profile name must be unique in the folder where you save it. The profile name cannot contain special characters.
Description	Optionally, enter a description for the profile.
Location	Choose a location to save the profile. If you do not choose a location, the profile is saved in the Default project.

Source Details

You can choose a source object after you create a connection to the data source in Administrator.

The following table lists the options that you can configure in the **Source Details** area:

Option	Description
Connection	Choose an existing connection. You can create a connection in Administrator.
Source object	Select a source object to run the profile on. When you click **Select** to browse for a source object, the **Select a Source Object** dialog box shows a maximum of 200 source objects. Use the **Find** field to search for a source object in the list. Optionally, you can use the copy  icon to copy the directory path for directory override in the **Advanced Options** for Azure Data Lake Store Gen2 and Amazon S3 V2 connections.
Formatting Options	Optional. Define the file format options. Data Profiling supports CSV and TXT files that have UTF-8 encoding enabled. Appears when you select a file-based connection.
Advanced Options	Mandatory. Configure the advanced options for the source objects.

Formatting Options

You can optionally configure the formatting options if you choose a file as a source object.

Flat File

You can run a profile on delimited flat files with multi-byte characters. The following table lists the options that you can configure for a flat file:

Option	Description
Delimiter	Indicates the boundary between two columns of data. Choose one of the following options: - Comma - Tab - Colon - Semicolon - Non Printable. When you choose this option, the **Non-printable character** drop-down list appears. Select a non-printable character to use as the delimiter. - Other. Select this option and specify the character to use as the delimiter. **Note:** - If you specify a comma, colon, or semicolon, the corresponding options are selected. - If the character specified here matches with any of the values in the **Non-printable character** drop-down list, the value appears in the **Non-printable character** drop-down list. If you use an escape character or a quote character as the delimiter, or if you use the same character as consecutive delimiter and qualifier, you might receive unexpected results. Default is comma.
Text Qualifier	Character that defines the boundaries of text strings. If you select a quote character, Data Profiling ignores delimiters within quotes. Default is double quote (").
Escape Character	Character that immediately precedes a column delimiter character embedded in an unquoted string, or immediately precedes the quote character in a quoted string. When you specify an escape character, Data Profiling reads the delimiter character as a regular character.
Field Labels	Choose one of the following options to display the column names in profile results: - Auto-generate. Data Profiling auto-generates the column names. - Import from Row <*row_number*>. Imports the column name from the specified row number.
First Data Row <*row_number*>	Row number from which Data Profiling starts to read when it imports the file. For example, if you enter **2**, Data Profiling skips the first row. **Note:** Data Profiling sets the **First Data Row** automatically when you set the **Import from Row** option. For example, if you set the **Import from Row** option to 10, Data Profiling sets the **First Data Row** to 11.

Amazon S3 v2

The following table lists the options for the delimited format type:

Option	Description
Schema Source	You must specify the schema of the source file. You can select one of the following options to specify a schema: - Read from data file. Amazon S3 V2 Connector imports the schema from the file in Amazon S3. - Import from schema file. Imports schema from a schema definition file in your local machine. Default is Read from data file.
Delimiter	Character used to separate columns of data. You can configure parameters such as comma, tab, colon, semicolon, or others. To set a tab as a delimiter, you must type the tab character in any text editor. Then, copy and paste the tab character in the Delimiter field.If you specify a multibyte character as a delimiter in the source object, the mapping fails. Default is comma (,).
Escape Character	Character immediately preceding a column delimiter character embedded in an unquoted string, or immediately preceding the quote character in a quoted string. Default is backslash (\).
Text Qualifier	Character that defines the boundaries of text strings. If you select a quote character, Data Profiling ignores delimiters within quotes. Default is double quote (").
Qualifier Mode	Specify the qualifier behavior for the target object. You can select one of the following options: - Minimal. Default mode. Applies qualifier to data that have a delimiter value or a special character present in the data. Otherwise, the Secure Agent does not apply the qualifier when writing data to the target. - All. Applies qualifier to all data. Default is Minimal.
Code Page	UTF-8. Select for Unicode and non-Unicode data. Select the code page that the Secure Agent must use to read data.
Header Line Number	Specify the line number that you want to use as the header when you read data from Amazon S3. You can also read data from a file that does not have a header. Default is 1. To read data from a file with no header, specify the value of the Header Line Number field as 0.To read data from a file with a header, set the value of the Header Line Number field to a value that is greater or equal to one. This property is applicable during runtime and data preview to read a file. This property is applicable during data preview to write a file.
First Data Row	Specify the line number from where you want the Secure Agent to read data. You must enter a value that is greater or equal to one. To read data from the header, the value of the Header Line Number and the First Data Row fields should be the same. Default is 2. This property is applicable during runtime and data preview to read a file. This property is applicable during data preview to write a file.
Row Delimiter	Character used to separate rows of data. You can set values as`\r\n`,`\n`, and `\r`.

The following table lists the options for the avro and parquet format type:

Option	Description
Schema Source	The schema of the source or target file. You can select one of the following options to specify a schema: - Read from data file. Default. Amazon S3 V2 Connector reads the schema from the source file that you select. - Import from Schema File. Imports schema from a schema definition file in your local machine.
Schema File	Upload a schema definition file. You cannot upload a schema file when you create a target at runtime.

The following table lists the options for the JSON format type:

Option	Description
Schema Source	The schema of the source or target file. You can select one of the following options to specify a schema: - Read from data file. Default. Amazon S3 V2 Connector reads the schema from the source file that you select. - Import from Schema File. Imports schema from a schema definition file in your local machine.
Schema File	Upload a schema definition file. You cannot upload a schema file when you create a target at runtime.
Sample Size	Specify the number of rows to read to find the best match to populate the metadata.
Memory Limit	The memory that the parser uses to read the JSON sample schema and process it. The default value is 2 MB.If the file size is more than 2 MB, you might encounter an error. Set the value to the file size that you want to read.

Azure Data Lake Store Gen2

The following table lists the options for the delimited format type:

Option	Description
Schema Source	You must specify the schema of the source file. You can select one of the following options to specify a schema: - Read from data file. Azure Data Lake Store Gen2 Connector imports the schema from the file in Azure Data Lake Store. - Import from schema file. Imports schema from a schema definition file in your local machine. Default is Read from data file.
Delimiter	Character used to separate columns of data. You can configure parameters such as comma, tab, colon, semicolon, or others. **Note:** You cannot set a tab as a delimiter directly in the **Delimiter** field. To set a tab as a delimiter, you must type the tab character in any text editor. Then, copy and paste the tab character in the **Delimiter** field. Default is comma (,).
Escape Character	Character immediately preceding a column delimiter character embedded in an unquoted string, or immediately preceding the quote character in a quoted string. Default is backslash (\).

Option	Description
Text Qualifier	Character that defines the boundaries of text strings. If you select a quote character, Data Profiling ignores delimiters within quotes. Default is double quote (").
Qualifier Mode	Specify the qualifier behavior for the target object. You can select one of the following options: - Minimal. Default mode. Applies qualifier to data that have a delimiter value or a special character present in the data. Otherwise, the Secure Agent does not apply the qualifier when writing data to the target. - All. Applies qualifier to all data. Default is Minimal.
Code Page	Select the code page that the Secure Agent must use to read data. Microsoft Azure Data Lake Storage Gen2 Connector supports only UTF-8. Ignore rest of the code pages.
Header Line Number	Specify the line number that you want to use as the header when you read data from Microsoft Azure Data Lake Storage Gen2. You can also read a data from a file that does not have a header. To read data from a file with no header, specify the value of the **Header Line Number** field as 0. **Note:** This property is applicable when you perform data preview. Default is 1.
First Data Row	Specify the line number from where you want the Secure Agent to read data. You must enter a value that is greater or equal to one. To read data from the header, the value of the **Header Line Number** and the **First Data Row** fields should be the same. Default is 2. **Note:** This property is applicable when you perform data preview.
Row Delimiter	Character used to separate rows of data. You can set values as `\r\n`, `\n`, and `\r`.

The following table lists the options for the avro and parquet format type:

Option	Description
Schema Source	The schema of the source or target file. You can select one of the following options to specify a schema: - Read from data file. Default. Azure Data Lake Store Gen2 Connector reads the schema from the source file that you select. - Import from Schema File. Imports schema from a schema definition file in your local machine.
Schema File	Upload a schema definition file. You cannot upload a schema file when you create a target at runtime.

The following table lists the options for the JSON format type:

Option	Description
Schema Source	The schema of the source or target file. You can select one of the following options to specify a schema: - Read from data file. Default. Azure Data Lake Store Gen2 Connector reads the schema from the source file that you select. - Import from Schema File. Imports schema from a schema definition file in your local machine.
Schema File	Upload a schema definition file. You cannot upload a schema file when you create a target at runtime.
Sample Size	Specify the number of rows to read to find the best match to populate the metadata.
Memory Limit	The memory that the parser uses to read the JSON sample schema and process it. The default value is 2 MB.If the file size is more than 2 MB, you might encounter an error. Set the value to the file size that you want to read.

Oracle Cloud Object Storage

The following table lists the options for the delimited format type:

Option	Description
Schema Source	You must specify the schema of the source file. You can select one of the following options to specify a schema: - Read from data file. Oracle Cloud Object Storage Connector imports the schema from the file in Oracle Cloud Object Storage. - Import from schema file. Imports schema from a schema definition file in your local machine. Default is Read from data file.
Delimiter	Character used to separate columns of data. You can configure parameters such as comma, tab, colon, semicolon, or others. **Note:** You cannot set a tab as a delimiter directly in the **Delimiter** field. To set a tab as a delimiter, you must type the tab character in any text editor. Then, copy and paste the tab character in the **Delimiter** field. Default is comma (,).
Escape Character	Character immediately preceding a column delimiter character embedded in an unquoted string, or immediately preceding the quote character in a quoted string. Default is backslash (\).
Text Qualifier	Character that defines the boundaries of text strings. If you select a quote character, Data Profiling ignores delimiters within quotes. Default is double quote (").
Qualifier Mode	Specify the qualifier behavior for the target object. You can select one of the following options: - Minimal. Default mode. Applies qualifier to data that have a delimiter value or a special character present in the data. Otherwise, the Secure Agent does not apply the qualifier when writing data to the target. - All. Applies qualifier to all data. Default is Minimal.
Code Page	Select the code page that the Secure Agent must use to read data. Oracle Cloud Object Storage Connector supports only UTF-8. Ignore rest of the code pages.

Option	Description
Header Line Number	Specify the line number that you want to use as the header when you read data from Oracle Cloud Object Storage. You can also read a data from a file that does not have a header. To read data from a file with no header, specify the value of the **Header Line Number** field as 0. **Note:** This property is applicable when you perform data preview. Default is 1.
First Data Row	Specify the line number from where you want the Secure Agent to read data. You must enter a value that is greater or equal to one. To read data from the header, the value of the **Header Line Number** and the **First Data Row** fields should be the same. Default is 2. **Note:** This property is applicable when you perform data preview.
Row Delimiter	Character used to separate rows of data. You can set values as`\r\n`,`\n`, and `\r`.

Advanced Options

If you choose source object such as Amazon S3, Azure Data Lake Store, Snowflake Cloud Data Warehouse V2, Microsoft Azure Synapse SQL, or Amazon Redshift V2, you can configure the following advanced options for the file.

Amazon S3 v2

The following table lists the options that you can configure for an Amazon S3 source object:

Option	Description
Source Type	Type of the source from which you want to read data. You can select the following source types: - File - Directory Default is **File**. For more information about the source type, see Source types in Amazon S3 V2 sources.
Folder Path	Optional. Overwrites the bucket name or folder path of the Amazon S3 source file.If applicable, include the folder name that contains the source file in the `<bucket_name>/<folder_name>` format. If you do not provide the bucket name and specify the folder path starting with a slash (/) in the `/<folder_name>` format, the folder path appends with the folder path that you specified in the connection properties. For example, if you specify the `/<dir2>` folder path in this property and `<my_bucket1>/<dir1>` folder path in the connection property, the folder path appends with the folder path that you specified in the connection properties in `<my_bucket1>/<dir1>/<dir2>` format.
File Name	Optional. Overwrites the Amazon S3 source file name.
Allow Wildcard Characters	Use the ? and * wildcard characters to specify the folder path or file name if you run a mapping in advanced mode to read data from an Avro, flat, JSON, ORC, or Parquet file.
Enable Recursive Read	Use the recursive read option for flat, Avro, JSON, ORC, and Parquet files. The files that you read using recursive read must have the same metadata. Enable recursive read when you specify wildcard characters in a folder path or file name. To enable recursive read, select the source type as Directory.

Option	Description
Incremental File Load	Incrementally load source files in a directory to read and process only the files that have changed since the last time the mapping task ran.
Staging Directory	Optional. Path of the local staging directory. Ensure that the user has write permissions on the directory. In addition, ensure that there is sufficient space to enable staging of the entire file. Default staging directory is the /temp directory on the machine that hosts the Secure Agent. When you specify the directory path, the Secure Agent create folders depending on the number of partitions that you specify in the following format: `InfaS3Staging<00/11><timestamp>_<partition number>` where, 00 represents read operation and 11 represents write operation. For example, `InfaS3Staging000703115851268912800_0` The temporary files are created within the new directory. The staging directory in the source property does not apply to an advanced cluster. However, you must specify a staging directory on Amazon S3 in the advanced configuration. For more information, see Administrator.
Hadoop Performance Tuning Options	Optional. This property is not applicable for Amazon S3 V2 Connector.
Compression Format	Decompresses data when you read data from Amazon S3. You can choose to decompress the data in the following formats: - None - Gzip Default is None. **Note:** Amazon S3 V2 Connector does not support the Lzo and Bzip2 compression format even though the option appears in this property. For more information about the compression format, see Data compression in Amazon S3 V2 sources and targets.
Download Part Size	Downloads the part size of an Amazon S3 object in bytes. Default is 5 MB. Use this property when you run a mapping to read a file of flat format type.
Multipart Download Threshold	Minimum threshold size to download an Amazon S3 object in multiple parts. To download the object in multiple parts in parallel, ensure that the file size of an Amazon S3 object is greater than the value you specify in this property. Default is 10 MB.
Temporary Credential Duration	The time duration during which an IAM user can use the dynamically generated temporarily credentials to access the AWS resource. Enter the time duration in seconds. Default is 900 seconds. If you require more than 900 seconds, you can set the time duration maximum up to 12 hours in the AWS console and then enter the same time duration in this property.

Azure Data Lake Store Gen2

The following table lists the options that you can configure for an Azure Data Lake Store source object:

Option	Description
Concurrent Threads	Optional. Number of concurrent connections to load data from the Microsoft Azure Data Lake Storage Gen2. When writing a large file, you can spawn multiple threads to process data. Configure **Block Size** to divide a large file into smaller parts. Default is 4. Maximum is 10.
Filesystem Name Override	Optional. Overrides the default file name.
Source Type	Type of the source from which you want to read data. You can select the following source types: - File - Directory Default is **File**. For more information about the source type, see Directory Source in Microsoft Azure Data Lake Storage Gen2 Sources.
Allow Wildcard Characters	Use the ? and * wildcard characters to specify the folder path or file name if you run a mapping in advanced mode to read data from an Avro, flat, JSON, ORC, or Parquet file.
Directory Override	Optional. Microsoft Azure Data Lake Storage Gen2 directory that you use to write data. Default is root directory. The Secure Agent creates the directory if it does not exist. The directory path specified at run time overrides the path specified while creating a connection.
File Name Override	Optional. Target object. Select the file from which you want to write data. The file specified at run time overrides the file specified in Object.
Block Size	Optional. Divides a large file or object into smaller parts each of specified block size. When writing a large file, consider dividing the file into smaller parts and configure concurrent connections to spawn required number of threads to process data in parallel. Default is 8 MB.
Compression Format	Optional. Compresses and writes data to the target. Select `Gzip` to write flat files.
Timeout Interval	Optional. The number of seconds to wait when attempting to connect to the server. A timeout will occur if the connection cannot be established in the specified amount of time.
Interim Directory	Optional. Path to the staging directory in the Secure Agent machine. Specify the staging directory where you want to stage the files when you read data from Microsoft Azure Data Lake Store. Ensure that the directory has sufficient space and you have write permissions to the directory. Default staging directory is `/tmp`. You cannot specify an interim directory for an advanced cluster. You cannot specify an interim directory when you use the Hosted Agent.

Option	Description
Incremental File Load	Incrementally load source files in a directory to read and process only the files that have changed since the last time the mapping task ran.
Enable Recursive Read	Use the recursive read option for flat, Avro, JSON, ORC, and Parquet files. The files that you read using recursive read must have the same metadata. Enable recursive read when you specify wildcard characters in a folder path or file name. To enable recursive read, select the source type as Directory.

Snowflake Cloud Data Warehouse V2

The following table lists the options that you can configure for a Snowflake Cloud Data Warehouse V2 source object:

Option	Description
Database	Overrides the database specified in the connection.
Schema	Overrides the schema specified in the connection.
Warehouse	Overrides the Snowflake warehouse name specified in the connection.
Role	Overrides the Snowflake role assigned to user specified in the connection.
Table Name	Overrides the table name of the imported Snowflake Cloud Data Warehouse source table.

Amazon Redshift V2

The following table lists the options that you can configure for an Amazon Redshift V2 source object:

Option	Description
S3 Bucket Name	Amazon S3 bucket name for staging the data. You can also specify the bucket name with the folder path.
Enable Compression	Compresses the staging files into the Amazon S3 staging directory. The task performance improves when the Secure Agent compresses the staging files. Default is selected.
Staging Directory Location	Location of the local staging directory. When you run a task in Secure Agent runtime environment, specify a directory path that is available on the corresponding Secure Agent machine in the runtime environment.Specify the directory path in the following manner: `<staging directory>` For example, `C:\Temp`. Ensure that you have the write permissions on the directory. Does not apply to an advanced cluster.
Temporary Credential Duration	The time duration during which an IAM user can use the dynamically generated temporarily credentials to access the AWS resource. Enter the time duration in seconds. Default is 900 seconds. If you require more than 900 seconds, you can set the time duration up to a maximum of 12 hours in the AWS console and then enter the same time duration in this property.

Option	Description
Encryption Type	Encrypts the data in the Amazon S3 staging directory. You can select the following encryption types: - None - SSE-S3 - SSE-KMS - CSE-SMK You can only use SSE-S3 encryption in a mapping that runs on an advanced cluster. Default is None.
Download S3 Files in Multiple Parts	Downloads large Amazon S3 objects in multiple parts. When the file size of an Amazon S3 object is greater than 8 MB, you can choose to download the object in multiple parts in parallel. Default is 5 MB. Does not apply to an advanced cluster.
Multipart Download Threshold Size	The maximum threshold size to download an Amazon S3 object in multiple parts. Default is 5 MB. Does not apply to an advanced cluster.
Schema Name	Overrides the default schema name. **Note:** You cannot configure a custom query when you use the schema name.
Source Table Name	Overrides the default source table name. **Note:** When you select the source type as Multiple Objects or Query , you cannot use the Source Table Name option.

Microsoft Azure Synapse SQL

The following table lists the options that you can configure for a Microsoft Azure Synapse SQL source object:

Option	Description
Azure Blob Container Name	Microsoft Azure Blob Storage container name. Required if you select Azure Blob storage in the connection properties.
ADLS FileSystem Name	The name of the file system in Microsoft Azure Data Lake Storage Gen2. Required if you select ADLS Gen2 storage in the connection properties. You can also provide the path of the directory under given file system.
Schema Name Override	Overrides the schema specified in the connection.
Table Name Override	Overrides the table name of the imported Microsoft Azure Synapse SQL source table.
Field Delimiter	Character used to separate fields in the file. Default is `0x1e`. You can specify 'TAB' or 0-256 single-char printable and non-printable ASCII characters. Non-printable characters must be specified in hexadecimal.
Number of concurrent connections to Blob Store	Number of concurrent connections to extract data from the Microsoft Azure Blob Storage. When reading a large-size blob, you can spawn multiple threads to process data. Configure **Blob Part Size** to partition a large-size blob into smaller parts. Default is 4. Maximum is 10.

Option	Description
Blob Part Size	Partitions a blob into smaller parts each of specified part size. When reading a large-size blob, consider partitioning the blob into smaller parts and configure concurrent connections to spawn required number of threads to process data in parallel. Default is 8 MB.
Quote Character	The Secure Agent skips the specified character when you read data from Microsoft Azure Synapse SQL. Default is `0x1f`.
Interim Directory	Optional. Path to the staging directory in the Secure Agent machine.Specify the staging directory where you want to stage the files when you read data from Microsoft Azure Synapse SQL. Ensure that the directory has sufficient space and you have write permissions to the directory. Default staging directory is /tmp. You cannot specify an interim directory for an advanced cluster. You cannot specify an interim directory when you use the Hosted Agent.

Oracle Cloud Object Storage

The following table lists the options that you can configure for an Oracle Cloud Object Storage source object:

Option	Description
Folder Path	Overrides the folder path value in the Oracle Cloud Object Storage connection.
File Name	Overrides the Oracle Cloud Object Storage source file name.
Staging Directory	Path of the local staging directory. Ensure that the user has write permissions on the directory. In addition, ensure that there is sufficient space to enable staging of the entire file. Default staging directory is the `/temp` directory on the machine that hosts the Secure Agent. The temporary files are created within the new directory.
Multipart Download Threshold	Minimum threshold size to download an Oracle Cloud Object Storage object in multiple parts. To download the object in multiple parts in parallel, ensure that the file size of an Oracle Cloud Object Storage object is greater than the value you specify in this property. **Range** : - Minimum: 4 MB - Maximum: 5 GB Default is 64 MB.
Download Part Size	Downloads the part size of an Oracle Cloud Object Storage object in bytes. **Range:** - Minimum: 4 MB - Maximum: 1GB Default is 32 MB.

Profile Settings

You can choose a sampling option for the profile run. You can also choose whether to drill down on the profile results.

The following table lists the options that you can choose in the **Profile Settings** area:

Property	Description
Run profile on	Choose one of the following sampling options to run the profile: - All rows. The profile runs on all the rows in the source object. - First *n* rows. The profile runs on the first *n* number of rows in the source. - Random sample *n* rows. The profile runs on the configured number of random rows.
Drilldown	Choose one of the following drill-down options: - Choose **On** to drill down on the profile results to display specific data. In the profiling results, when you choose a data type, pattern, or value, Data Profiling displays the relevant data in the **Data Preview** area. If you choose this option, you can run queries on the source object after you run the profile. - Choose **Off** to not drill down on the source object. To drill down and to query the source object, you need Data Preview privileges in Data Profiling. **Note:** You cannot perform drill down on the profile results or queries if you select the Avro or Parquet source object for Amazon S3 and Azure Data Lake Store connections.

The following table lists the connections and supported sampling options:

Connection	Sampling Option
Azure Data Lake Store Gen2	All Rows
Amazon S3 v2	All Rows
Microsoft Azure Synapse SQL	All Rows First N Rows Random N Rows
Flat File	All Rows
Google Big Query v2	All Rows
Oracle	All Rows First N Rows
Salesforce	All Rows First N Rows
SQL Server	All Rows First N Rows
Snowflake Cloud Data Warehouse V2	All Rows First N Rows Random N Rows
Amazon Redshift V2	All Rows Random N Rows

Columns

The **Columns** tab displays the columns that are supported by Data Profiling. You can select or clear the columns on the **Columns** tab. The profile runs on the selected columns to extract column statistics.

Data Profiling supports the following data types and column precision:

- Non-numeric data types. Supports columns with a precision from 0 through 4000.
  By default, Data Profiling selects the columns with a precision from 0 through 150 on the **Columns** tab.
- Numeric data types. Supports columns with a precision from 0 through 38.
  By default, Data Profiling selects the columns with a precision from 0 through 15 on the **Columns** tab.

Note: If you select columns with precision greater than 255 in the **Columns** tab, Data Profiling truncates the value frequency and calculates the statistics based on the first 255 characters on the **Results** page.

The following table lists the properties that you can view on the **Columns** tab:

Property	Description
Columns	The column names in the selected source object.
Type	The data type that appears in the transformations. They are internal data types based on ANSI SQL-92 generic data types, which the Secure Agent uses to move data across platforms. Transformation data types appear in all transformations in a mapping
Precision	The number of characters in a column.
Native Data Type	The data type specific to the source database or flat files.
Scale	The number of numeric characters after the decimal point.
Nullable	Indicates whether the column can accommodate a NULL value.
Key	Indicates whether the column has been designated as a primary key in the data source.

To sort the list of columns in ascending or descending order, click the column name or a property name. Use the **Find** field to search for columns.

When a column is added or deleted in the data source, the list of columns on the **Columns** tab gets updated when you edit the profile. The changes appear only if the runtime environment is up and running. When a column is added, the column appears on the **Columns** tab and is unmarked. You can choose the column for the next profile run. When a column is deleted, the deleted column does not appear on the **Columns** tab.

Example

You are a data steward user. You have created and run a profile called *CustP* on the Customer table. You want to modify the profile based on a business need to classify customers for a new rewards program. To accomplish this task, add the rule to the profile and select the columns in the profile that meet the business need.

To add the rules and select the relevant columns in the profile, perform the following steps:

1. Modify the *CustP* profile.
2. Choose the columns that meet the business need.
3. In Data Quality, create a rule specification for the business need.
4. In Data Profiling, add the rule specification to the profile.
5. Run the profile.

6. View the profile results to measure the quality of data.
7. Export the results to a Microsoft Excel file for further analysis.

Override Column Metadata

You can edit the column metadata of delimited files from the flat file, Azure Data Lake Store Gen2, and Amazon S3 v2 connections. Edit the column metadata of delimited files before you run a profile.

You can edit metadata if you want to change the data type, precision, or scale of the columns in the source object. For example, when you run a profile on a flat file connection that does not have an embedded schema, the profile results sometimes display inaccurate inferred data types. You might want to identify such columns and edit the metadata of the columns to change the data type, precision, or scale values.

You can edit the **Native Data Types**, **Precision**, and **Scale** properties of a column. When you edit metadata, you can change the precision and scale, if applicable for the data type.
To edit the column properties, select a column and click the property value, and then edit metadata based on your requirements. Alternatively, you might want to edit the column properties in bulk.

To edit the column properties in bulk, perform the following steps:

1. Click **Actions** > **Override Column Metadata**.
   The **Override Column Metadata** window appears.
2. In the **Define Metadata** section, specify the following options:
   - Native Data Type. Click the menu icon and select the data type.
   - Precision. Enter a precision value. The precision must be greater than or equal to 1.
   - Scale. Enter a scale value. Scale must be greater than or equal to 0. The scale of a number must be less than its precision.
3. In the **Columns to Override** section, select the columns that you want to edit and click **Apply**.

Filters

You can use filters to select the values that a profile can read in a column of source data. You can create filters based on the simple and query filter types.

When you add a filter to a data column, the profile runs only on the data values that meet the filter criteria that you specify. You can add, delete, or update the filters in subsequent runs. After you add the filters, you can choose the filter that you want for the next profile run.

When you delete a column from the source object, any filter on the column is deleted from the profile during the profile run. When a filter applies to more than one column and you delete one of the columns, Data Profiling ignores the filter or filter condition that uses the deleted column during the profile run.

You can create the following types of filter:

Simple filter

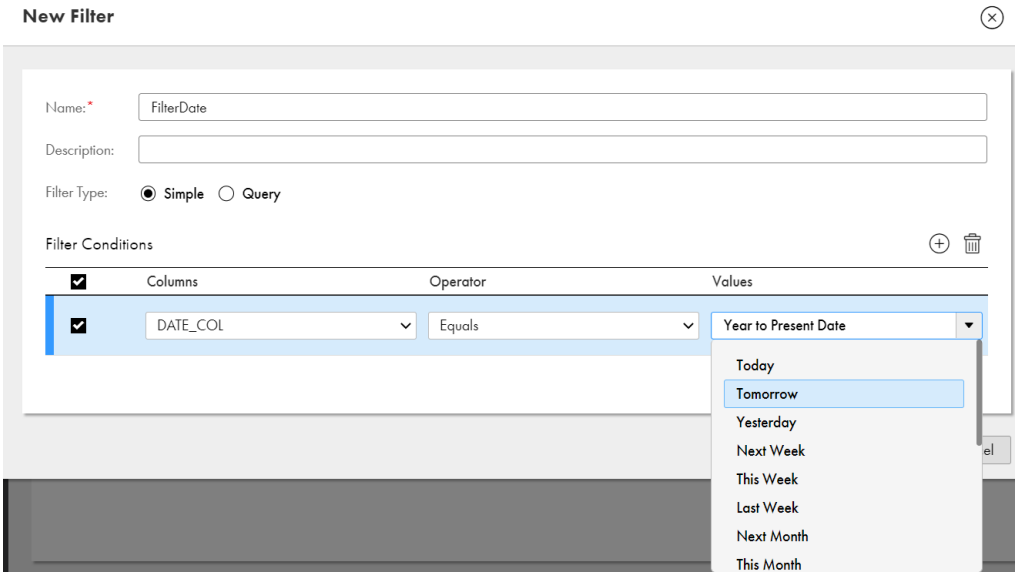
When you create a simple conditional filter, you can select operators such as Equals, Less Than, Less Than or Equals, Greater Than, Greater Than or Equals, Not Equals, Is Null, and Is Not Null.

For example, you are a data analyst and you created a profile on a Sales table. You want to extract the sales details for New York and share it with the business team. To accomplish this task, you create a filter with the filter condition `City = New York` and add it to the profile. You run the profile and export the profile results to share with the business team.

You can also create dynamic filters for relational data sources to filter the date and timestamp columns. The dynamic filter includes options such as Today, Tomorrow, Yesterday, Next Week, Next Month, and Custom.

For example, assume that you want to profile the sales orders that were created last month, and run the profile every month. To accomplish this task, you create a filter with the dynamic filter condition `COLUMN_DATE = Last Month` and add it to the profile. By doing this, you need not change the filter condition every month and Data Profiling resolves the right date at the runtime when the profiling task runs.

The following image displays a sample of the simple dynamic filter:



Query filter

You can define a custom SQL query to apply a complex filter condition to the column data. You can create an SQL filter for relational data sources such as Oracle, Amazon Redshift, and Snowflake. You must enter the SQL query with just the WHERE clause, but not the entire query statement.

You can enter the SQL query starting with the query condition as shown in the following example, `Id IN (SELECT Id FROM TABLE_2 WHERE Id > '35') AND City='Chicago'`.

Tip: Test the SQL statement you want to use as a filter condition before you create a saved query. Data Profiling does not display specific error messages for invalid SQL statements.

Note: To filter Google BigQuery source objects, use the SQL Override Query in the advanced options.

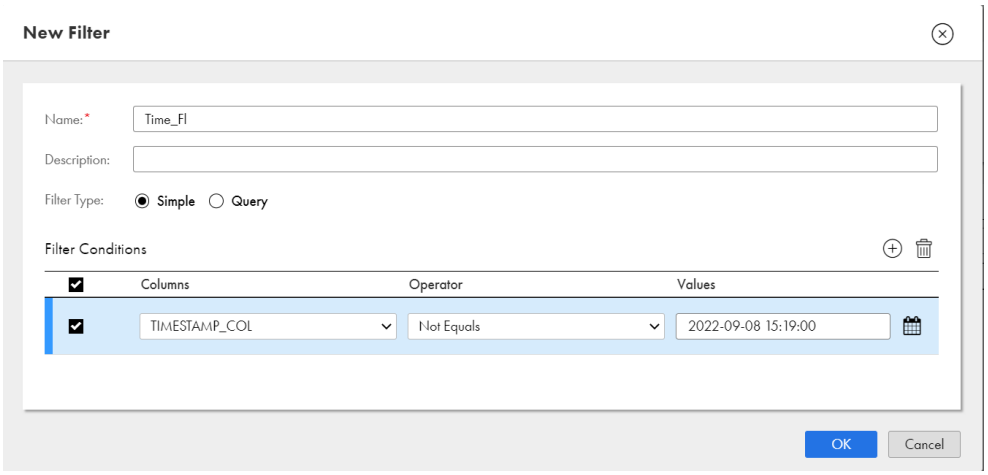
Creating Filters

You can create one or more filters in a profile. You can add a simple conditional or SQL filters to a profile.

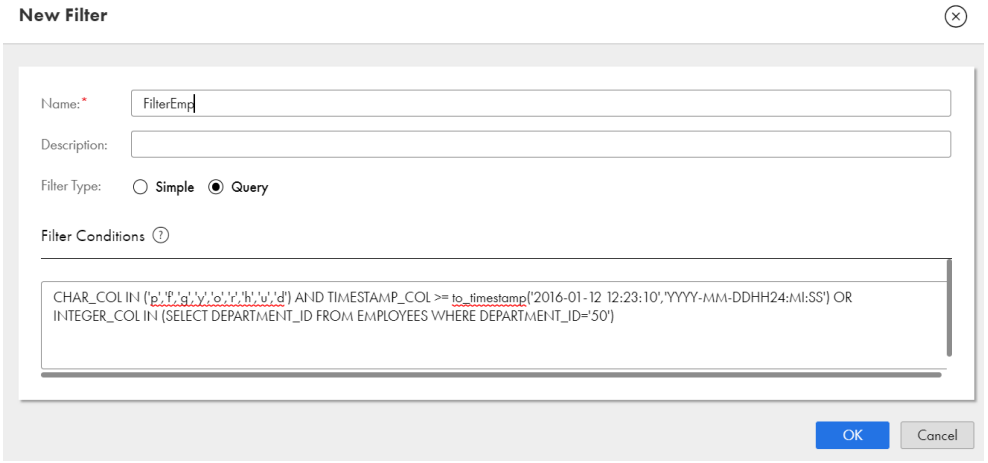
Note: You can create filters on the partitioned fields for profiles that you create with Avro and Parquet source objects for Amazon S3 or Azure Data Lake Store connection.

1. On the Filter tab, click Add ( ).
2. In the **New Filter** dialog box, enter a name for the filter. Optionally, add a description for the filter.
3. In the **New Filter** dialog box, create the following filter types:

- **Simple**. To enter a simple filter condition, click Add ( ). Choose a column and an operator, and enter a valid value. If required, continue to add more filter conditions.
  The following image shows a sample **New Filter** dialog box with a simple filter condition:



- **Query**. To add a SQL query as a filter, select **Query** from the filter type options, and then type in or paste an SQL query in the text box.
  The following image shows a sample **New Filter** dialog box with a query filter:



4. Click **OK**.

   The filter appears on the **Filter** tab. Add multiple filters if required.

Adding a filter to the profile run

You can add one filter to the profile for a profile run. You can change the filter in subsequent runs. On the **Filters** tab, after you create one or more filters, the **Use in Profile** option is enabled and selected by default.

1. Choose the required filter for the profile run.
2. Click **Save**.

   When you run the profile, the filter is applied to the source object and the profile runs on the filtered results.

Data preview

The **Data Preview** tab displays the first 10 rows and all the columns in the source object. To view this tab, you need the **Data Profiling - Data Preview** role in Data Profiling.

The **Data Preview** tab displays a checkmark for the selected columns if the column data type is supported by Data Profiling. The profile scope shows the number of rows that the profile runs on.

Note: You cannot preview data of columns if you select Avro or Parquet source object for Amazon S3 and Azure Data Lake Store connections.

Rules

On the **Rules** tab, you can add Data Quality assets as rules to a profile. Data Profiling also assigns rules automatically to the profile based on the chosen source object and its attributes. You can choose one or more rules for a profile run.

You can open a data quality asset from the Explore page or from within a profile in Data Profiling.

Add rules to the profile

You can add rule specification, cleanse, parse, and verifier assets as rules to a profile. You create these assets in Data Quality. You can add a Data Quality asset as a rule if you have Read permission on the asset. You can also profile passive mapplets, which may or may not have Data Quality assets. Profiling will calculate the statistics on all the output ports of the mapplet, including value frequencies.

You can add one or more rules for a data profiling task. You can also run a profile without a rule. Data Profiling displays column statistics and rule results in collapsible sections in the results area. The results for each rule output appear in a separate row.

In Data Quality, when you create rule specification, cleanse, parse, or verifier assets, you configure inputs, rule logic, and outputs for the asset. When you add the asset as a rule in Data Profiling, the input appears as input column and the output appears as rule output. You can add single input, single output and multiple input, single output rules to profiles. When you add a rule to the profile, you assign a source column to the input column. When you run the profile, Data Profiling generates statistics based on the rule logic. The **Results** tab shows the rule output statistics in a separate row.

For example, a rule specification 'Validity' has an input called in_value, a rule logic, and an output called out_validity. You want to perform an analysis on a source column called 'customer-national_ID' in the Customer table. To accomplish this task, you perform the following steps:

1. On the **Rules** tab, you click Add to add a rule to the profile.
2. In the **Add Rule**, you select the 'Validity' rule.
3. In the **Rule Settings** dialog box, you select the column 'customer-national_ID' as the input column. Data Profiling assigns the selected column to input 'in_value'.
4. You run the profile.
5. Data Profiling generates the rule statistics based on the rule logic.
6. On the **Results** tab, the rule statistics appear in the 'out_validity' row.

When you add a single input rule, you can assign multiple columns to it. Data Profiling replicates the rule for each column. When you add a multiple input rule to a profile, you can add a column for each input in the rule. Data Profiling displays results for each selected column in a separate row.

You can add the following Data Quality assets as rules to a profile:

Rule specification

Use this asset to define a business rule with a set of conditions that you can use to evaluate your data. You can add rule specifications that have a single output.

A rule specification can also contain a single passive mapplet or nested passive mapplets. You can use mapplets that contain passive transformations in a rule specification. You can use the following assets in a mapplet:

- Parse
- Cleanse
- Labeler
- Rule specification
- Verifier
- Expression
- Java
- Mapplet that contains passive transformations

For more information about using mapplets in rule specifications, see *Rule specification assets* in the Data Quality documentation.
For example, you are a sales analyst and you want to analyze the retail sales in the Sales table.

1. In Data Quality, you perform the following steps:
    a. Create a rule specification named Reg_pyr.
    b. Add Region and SalesYear as the inputs.
    c. Create the rule logic and test it.
    d. Save the rule specification.
2. In Data Profiling, you perform the following steps:
    a. Create a profile on the Sales table.
    b. Add Reg_pyr rule to the profile and choose Region and SalesYear source columns for the rule.
    c. Save and run the profile.
    d. View the results on the **Results** tab. Optionally, export the results to a Microsoft Excel file or run a query that generates the content into a delimited file for further analysis.

Cleanse

Use this asset as a rule to standardize the appearance of your data, replace incorrect values in your data, and remove unwanted values from your data.

For example, you are a data analyst and you want to convert the FirstName and LastName columns in the Customer table to title case for better readability. To accomplish this task, you can perform for the following steps:

1. In Data Quality, you perform the following steps:
    a. Create a cleanse asset named FN_SenC.
    b. Add a step sequence and choose **Title Case** as casing style.
    c. Save the asset.
    d. Test the asset with sample data.

2. In Data Profiling, you perform the following steps:
    a. Create a profile on the Customer table.
    b. Add FN_SenC rule to the profile and choose FirstName and LastName columns for the rule.
    c. Save and run the profile.
    d. View the results on the **Results** tab. Optionally, export the results to a Microsoft Excel file or run a query that generates the content into a delimited file for further analysis.

Verifier

Use this asset as a rule to measure and enhance the quality of your postal address data. You can add a Verifier asset in the **Verification only** mode to a profile.

For example, you are a data analyst and the marketing department wants to send new product brochures to potential customers in California state. They want to evaluate the accuracy and deliverability of the address records in the Leads table before they send the brochures. To accomplish this task, you perform the following steps:

1. In Data Quality, you perform the following steps:
    a. Create a verifier asset named Cal_addr.
    b. Select appropriate address model for the input address structure and specify the input and output fields.
    c. In the Process tab properties, choose **Verification only** as the verification mode.
    d. Save the asset.
2. In Data Profiling, you perform the following steps:
    a. Create a profile on the Leads table.
    b. Add Cal_addr rule to the profile and choose Address1 and Address2 columns for the rule.
    c. Save and run the profile.
    d. View the results on the **Results** tab. Optionally, export the results to a Microsoft Excel file or run a query that generates the content into a delimited file for further analysis.

Parse

Use a parse asset to improve the structure of your data. A parse asset defines a set of operations that can identify discrete values in an input field and write the values to appropriate output fields.
For example, you are a data analyst and you need to find out information about potential customers from the list of email addresses. The data source includes emails of people who contacted your organization. You need to share the results with the sales department so that they can pursue the new customers. To accomplish this task, you perform the following steps:

1. In Data Quality, you perform the following steps:
    a. Create a parse asset named Email_parse.
    b. Add the **Regular Expression** parse step.
    c. Select the **Parse Email** built-in regular expression.
    d. Enter `Name`, `Company`, and `Domain` as the output fields.
    e. Save the asset.
2. In Data Profiling, you perform the following steps:
    a. Create a profile on the customer details table.
    b. Add Email_parse rule to the profile and choose Email_ID source column for the rule.

c. Save and run the profile.

d. View the results on the **Results** tab. Optionally, export the results to a Microsoft Excel file or run a query that generates the content into a delimited file for further analysis.

You cannot add rules if the rule input or rule output name exceeds 4000 bytes. When you open a Data Quality asset that is associated to a profile, the **Used by** section on the **Asset References** tab shows the profile name.

For information about creating a rule specification, cleanse, verifier, or parse asset, see *Data Quality* in Data Quality help.

Mapplet

Use a mapplet to transform the source data. You can add passive mapplets as rules to a profile. A mapplet is reusable transformation logic that you can use to transform source data before it is loaded into the target. For example, you are a data analyst and you want to concatenate the first name and last name of customers in the Customer table to get the full name of customers. To accomplish this task, perform the following steps:

1. In Data Integration, you perform the following steps:
   a. Create a mapplet asset named Concatenate_mapplet.
   b. Add FirstName and LastName as the mapplet inputs.
   c. Add expression transformation to the mapplet.
   d. Add FullName as the mapplet output.
   e. Validate and save the mapplet.
2. In Data Profiling, you perform the following steps:
   a. Create a profile on the Customer table.
   b. Add Concatenate_mapplet rule to the profile and choose FirstName and LastName source columns for the rule.
   c. Save and run the profile.
   d. View the results on the **Results** tab. Optionally, export the results to a Microsoft Excel file or run a query that generates the content into a delimited file for further analysis.

For information about creating mapplets, see Mapplets in Data Integration.

Note:

- You cannot add active mapplets to a profile.
- Mapplets work only for profiles on native engine and do not work for profiles on spark engine.
- Mapplets are of three types: Data Integration, PowerCenter and SAP. Only Data Integration and PowerCenter mapplets can be used in Data Profiling.
- Mapplets that support parameters or require connection for lookups are not supported in Data Profiling.
- You can use the following list of assets in a mapplet:
  - Parse
  - Cleanse
  - Labeler
  - Rule specification
  - Verifier

- Expression
- Java
- Nested mapplet

- There are other transformations available in Data Integration that you can use in a mapplet. However, these transformations are not used in Data Profiling as they make the mapplet active. For information about other transformations, see Transformations in Data Integration.

Adding rules to a profile

You can add one or more rules to a profile run. You can add or delete the rules in subsequent runs.

1. On the **Rules** tab, click Add ().
2. In the **Add Rule** dialog box, choose a rule specification, cleanse, or verifier asset.
3. Click **Select**.
4. In the **Rule Settings** dialog box, perform either of the following actions:
   - If the Data Quality asset has a single input, choose one or more columns for the input rule.
   - If the Data Quality has multiple inputs, choose one column for each input.
5. Click **OK**.

   The rule appears on the **Rules** tab.
6. Continue to add more rules to the **Rules** tab as necessary.
7. Click **Save**.

Automatic rule association with source objects

Data Profiling automatically associates Data Quality assets as rules with columns, based on the column and source object name match. By default, Data Profiling associates rules with columns of Oracle, Flat File, ODBC, and Amazon S3 V2 connections.

To enable automatic rule association, make sure that you have a valid DataQualityClairRule package license for your organization. The DataQualityClaireRule package contains the connection-specific JSON files and a default JSON file.

Data Profiling uses the connection-specific JSON file for all the supported connections. To enable automatic rule association for the remaining connections, you can configure the `DefaultAutoAssignRulesConfig.json` file.

Data Profiling automatically associates rules with columns after you configure the `<connection_type>AutoAssignRulesConfig.json` file for the connection. You can configure the JSON file in the following location: `<secureagentlocation>/apps/Data_Integration_Server/data/profiling/AutoRuleAssignmentConfig/`.

Note: You need not restart the Secure Agent after you configure or customize the Config.json files.

When you configure the `AutoAssignRulesConfig.json` file for a specific connection, the Data Quality assets are assigned as rules to the matching column and source object names.

If the column names and source object names do not match the auto assign criteria in the connection `AutoAssignRulesConfig.json` file. Data Profiling assigns rules to matching results from the `DefaultAutoAssignRulesConfig.json` file.

If column and source names in the `AutoAssignRulesConfig.json` file do not match the automatic rule association criteria, you can edit the connection-specific JSON file to change the source object name and column names.

Example

You created a profile with a source object that contains columns named ID, First Name, and Last Name. You might want to assign the Employee_details rule to the columns automatically.

To achieve this goal, you must ensure that you have the DataQualityClaireRule package license in your organization, and then copy the ClouldDataQuality_Bundles from Administrator Service to the ClouldDataQuality_Bundles project. If the column names in the `AutoAssignRulesConfig.json` file match the source column names in the source object, Data Profiling automatically assigns the Employee_details rule to the columns.

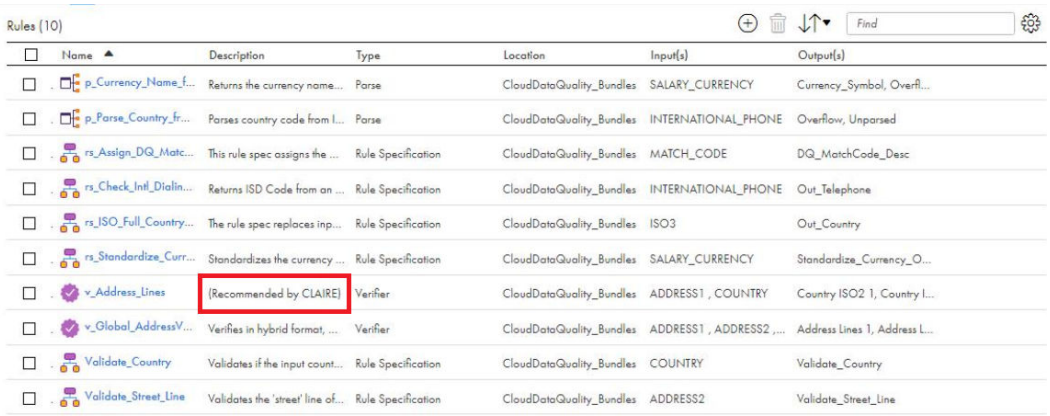
Automatic rule association steps

1. Ensure that the organization has the *DataQualityClaireRule* package license enabled.
2. In Data Profiling service, create a project named *CloudDataQuality_Bundles*.
3. Copy the *CloudDataQuality_Bundles* bundle from **Administrator service** > **Add-On Bundles** > **Available Bundles** to the *CloudDataQuality_Bundles* project. After you copy the bundle to the project, the project displays all the Data Quality assets that you can use for automatic rule association.
4. In the `<secureagentlocation>/apps/Data_Integration_Server/data/profiling/AutoRuleAssignmentConfig/` location, configure the `<connection_type>AutoAssignRulesConfig.json.file` with the data source information based on your requirements as shown in the following sample image:

```
  "sourceType": "Oracle",
  "ruleAssignments": [
    {
      "assignmentType": "ColumnNameMatch",
      "sourceName": "FRENCH_COMPANY_NAMES",
      "rule": {
        "ruleType": "RULE_SPECIFICATION",
        "name": "Validate_Longitude",
        "path": "/CloudDataQuality_Bundles"
      },
      "inportPortMappings": [
        {
          "portName": "Input_Longitude",
          "columnName": "longitude"
        }
      ],
      "outputPortMappings": [
        {
          "portName": "Longitude_Validate",
          "isProfileable": true
        }
      ]
    },
    {
      "assignmentType": "ColumnNameMatch",
      "sourceName": "EMPLOYEE_DATA",
      "rule": {
        "ruleType": "RULE_SPECIFICATION",
        "name": "Validate_Street_Line",
        "path": "/CloudDataQuality_Bundles"
      },
      "inportPortMappings": [
        {
          "portName": "Input_Street_Line",
          "columnName": "address2"
        }
      ],
      "outputPortMappings": [
        {
          "portName": "Validate_Street_Line",
          "isProfileable": true
```


5. View the associated rules in Data Profiling.

The following image shows the associated rules with the source objects:



Rules (10)

Name	Description	Type	Location	Input(s)	Output(s)
p_Currency_Name_f...	Returns the currency name...	Parse	CloudDataQuality_Bundles	SALARY_CURRENCY	Currency_Symbol, Overfl...
p_Parse_Country_fr...	Parses country code from l...	Parse	CloudDataQuality_Bundles	INTERNATIONAL_PHONE	Overflow, Unparsed
rs_Assign_DQ_Matc...	This rule spec assigns the ...	Rule Specification	CloudDataQuality_Bundles	MATCH_CODE	DQ_MatchCode_Desc
rs_Check_Intl_Dialin...	Returns ISD Code from an ...	Rule Specification	CloudDataQuality_Bundles	INTERNATIONAL_PHONE	Out_Telephone
rs_ISO_Full_Country...	The rule spec replaces inp...	Rule Specification	CloudDataQuality_Bundles	ISO3	Out_Country
rs_Standardize_Curr...	Standardizes the currency ...	Rule Specification	CloudDataQuality_Bundles	SALARY_CURRENCY	Standardize_Currency_O...
v_Address_Lines	(Recommended by CLAIRE)	Verifier	CloudDataQuality_Bundles	ADDRESS1 , COUNTRY	Country ISO2 1, Country l...
v_Global_AddressV...	Verifies in hybrid format, ...	Verifier	CloudDataQuality_Bundles	ADDRESS1 , ADDRESS2 ,...	Address Lines 1, Address L...
Validate_Country	Validates if the input count...	Rule Specification	CloudDataQuality_Bundles	COUNTRY	Validate_Country
Validate_Street_Line	Validates the 'street' line of...	Rule Specification	CloudDataQuality_Bundles	ADDRESS2	Validate_Street_Line

- The **Rules** tab displays *(Recommended by CLAIRE)* as a suffix in the rule description.

Customize an AutoAssignRulesConfig.json file

In this scenario, Data Profiling contains a profile with a source object named *Employee* and column named *First Name*. The column names and source names that are present in the source object do not match in `AutoAssignRulesConfig.json` file. You might want to customize the `AutoAssignRulesConfig.json` file to add rules to columns to match the source names and source objects.

Existing Field Value	Customized Field Value
Change the sourceName field value from *French_Company_Names*	Change to *Employee*
Change the columnName field value from *longitude*	Change to *First Name*

The following image shows a sample `AutoAssignRulesConfig.json` file with existing source and column names:
`{`
`"sourceType": "Oracle",`
`"ruleAssignments": [`
`{`
`"assignmentType": "ColumnNameMatch",`
`"sourceName": "FRENCH_COMPANY_NAMES",`
`"rule": {`
`"ruleType": "RULE_SPECIFICATION",`
`"name": "Validate_Longitude",`
`"path": "/CloudDataQuality_Bundles"`
`},`
`"inportPortMappings": [`
`{`
`"portName": "Input_Longitude",`
`"columnName": "longitude"`
`}`
`],`
`"outputPortMappings": [`
`{`
`"portName": "Longitude_Validate",`
`"isProfileable": true`
`}`
`]`
`},`
	The following image shows the changes made to the `AutoAssignRulesConfig.json` file:
	`{`
	`"sourceType": "Oracle",`
	`"ruleAssignments": [`
	`{`
	`"assignmentType": "ColumnNameMatch",`
	`"sourceName": "Employee",`
	`"rule": {`
	`"ruleType": "RULE_SPECIFICATION",`
	`"name": "Validate_Longitude",`
	`"path": "/CloudDataQuality_Bundles"`
	`},`
	`"inportPortMappings": [`
	`{`
	`"portName": "Input_Longitude",`
	`"columnName": "First Name"`
	`}`
	`],`
	`"outputPortMappings": [`
	`{`
	`"portName": "Longitude_Validate",`
	`"isProfileable": true`
	`}`
	`]`
	`},`

Rule occurrences and scorecards

A rule occurrence is a set of metrics you can create from a rule specification linked to a profile. A rule specification is the building block for a rule occurrence. You can configure multiple rule occurrences for a rule specification associated with a profile. You cannot use other Data Quality assets or mapplets to create rule occurrences. After you create rule occurrences in a profile, you can run the profile and view scorecards.

A scorecard is the graphical representation of valid values for a column in a profile. A scorecard is a collection of rule occurrences and represents data quality scores calculated when you profile a source dataset. You can use scorecards to measure data quality scores and monitor data quality progress. A measure of data quality in the source data is critical information in the management of the data asset in an enterprise. You can drill down on live data in a scorecard.

You can use scorecards to measure data quality progress for existing and new profiles. To view the scorecard, use the scorecard dashboard in Data Governance and Catalog.

Prerequisites to view scorecards

The following prerequisites must be fulfilled to view the scorecard dashboard in Data Governance and Catalog:

- You must have Intelligent Cloud Data Management and Data Governance and Catalog licenses.
- The Intelligent Cloud Data Management user must be assigned the **Governance User** role.

Important: Scorecard feature depends on the availability of Data Quality and Data Governance and Catalog on the pod where the organization is located.

Prerequisites to create rule occurrences

Before you create a rule occurrence, you must verify the configuration and output of the rule specification that you link to the profile. Verify the following prerequisites:

- The rule specification that you select must be defined with a dimension. A dimension is a one-word summary of the data quality issue that a rule specification represents. The dimension reflects the primary purpose of the business rule.
- The output of the rule specification must be one of the following values for rows that pass the quality check:
  - Valid
  - True
  - 1
  - Yes
  - Ok

Note:

- For rows that do not pass the quality check, the output of the rule specification can be any value.
- The source of truth for the dimension is the rule specification in Data Quality.
- You cannot create rule occurrences on a profile with rules that have multiple output ports.
- You can create rule occurrences on a profile with rules that have multiple input ports but when you run the profile, the scorecard dashboard displays the scores corresponding to only one column selected randomly from the input ports.

Creating rule occurrences

Create rule occurrences on the **Metrics** tab. The tab appears after you select **Scorecard Metrics** from the **Menu** option. After you create rule occurrences, you must save and run the profile.

Perform the following steps to create a rule occurrence:

1. On the **Metrics** tab, click Add ().
2. In the **Create Rule Occurrence** dialog box, choose a rule specification output to measure. You can use only rule specifications to create a rule occurrence and other Data Quality assets or mapplets cannot be used.
3. Click **Next**.
4. In the **Create Rule Occurrence** dialog box, perform the following steps:
   a. Specify a name and description for the rule occurrence in the **General Details** section.
   b. Define valid threshold values for scorecard generation in the **Rule Occurrence Thresholds** section. You can modify a threshold after you create a rule occurrence without running the profile again.
5. Click **OK**.

   The rule appears on the **Rule Occurrences** tab.
6. Continue to add more rule occurrences to the **Rule Occurrences** tab as necessary.
7. Choose one of the following options to save and run the profile:
   - Click **Save** to save the profile.
   - Click **Run** to save and run the profile.

   You can view the scorecard results after the data profiling task with a rule occurrence is complete.

   Note: When you delete a rule associated with a data profiling task, the corresponding rule occurrences are automatically deleted from the profile.

Viewing scorecards

Use scorecards to measure data quality scores and monitor data quality progress for existing and new profiles.

Click the **View Scorecard** button to view the scorecard dashboard in Data Governance and Catalog.

The following table lists the widgets that you can view with the scorecard dashboard:

Widget	Description
Average Latest Scores by Dimensions	Donut charts with round off values of the average latest data quality scores based on dimensions.
Number of Rule Occurrences by Dimensions	Number of rule occurrences for each dimension based on Good, Acceptable, and Not Acceptable threshold values.
Rule Occurrences	Shows the following details of rule occurrences: - Latest data quality score - Dimension of the rule specification - Date and time of latest profile run - Total number of rows processed - Total number of failed rows - Input column or primary data element - Preview of successful and unsuccessful rows

Every time you run a data profiling task with rule occurrences, the scores on the scorecard dashboard are updated. If you define a rule occurrence but do not execute the profile, then the rule occurrence appears on the scorecard dashboard without any score.

Note:

- Scorecards are created based on a profiling source. If you wish to create a scorecard with rule occurrences from a different source, you must use Data Governance and Catalog.
- When you run a data profiling task with a rule occurrence that has rules with multiple input ports, the scorecard dashboard displays the scores corresponding to only one column selected randomly from the input ports.

Example

You are a data analyst. You create and run profiles on a Customer table. You want to check the validity of the data available in the latest profile run.

You perform the following tasks:

1. Create a rule specification with the appropriate rule logic in Data Quality and set the dimension to **Validity**. When you apply a **Validity** dimension to a rule, the output data conforms to defined business rules and falls within allowable parameters when those rules are applied.
2. Create a profile and associate the rule specification.
3. Create a rule occurrence on the rule specification with Good, Acceptable, and Not Acceptable threshold values to be considered for scoring.
4. Save and run the profile.
5. View the metrics in the Data Governance and Catalog scorecard dashboard. You can use the metrics to verify the data quality progress in the Customer table.

Viewing stakeholder information

You can view users that have been designated as stakeholders for the rule occurrences on the **Overview** and **Stakeholder** tab in Data Governance and Catalog.

A stakeholder is an authorized user who is responsible for the rule occurrences, can approve or reject change requests for the occurrence, provide inputs to the properties of the rule occurrence, and are interested in following the asset to monitor changes.

To assign stakeholders to the rule occurrences, the organization administrator must enable the Data Governance and Catalog Stakeholdership and Participate in Change Approvals privileges for the user role. If you do not have the Stakeholdership and Participate in Change Approvals privileges, the organization administrator must provide your user role with Update or Create permission for the given asset type in Metadata Command Center.

For more information about the stakeholders, see the Asset Details and Working with Assets guides in the Data Governance and Catalog documentation.

Schedule and advanced options

On the **Schedules** tab, you can configure schedules, insights, runtime environment, email notifications, and advanced options for the profile.

Schedule details

You can configure the schedule details to run a profile on a schedule.

The following table lists the options that you can choose in the **Schedule Details** area:

Option	Description
Do not run this task on a schedule	Choose this option if you want to manually run the profile.
Run this task on a schedule	Choose a schedule to run the data profiling task. You can create, view, edit, and delete schedules in Administrator. To delete a schedule for a data profiling task, you must disassociate or delete the assets linked to the schedule.

Note: When you choose to run a profile on a schedule, the profile runs after the configured schedule offset time. For example, if you configure a schedule to run every hour from 8:00 a.m. to 12:00 p.m., and the schedule offset for your organization is 15 seconds. Your schedule runs at 8:00:15, 9:00:15, 10:00:15, 11:00:15, and 12:00:15. For information about schedule offset, see *Administrator* in the Administrator help.

Runtime environment

You can choose a runtime environment to run the task. If you do not choose a runtime environment, the profile runs on the default runtime environment configured for the connection.

You can create, view, edit, or delete runtime environments in Administrator. Displays runtime environments based on the source object that you select. For example, if the source object that you select is Avro, Parquet, or JSON, Data Profiling lists all the runtime environments that has the Elastic Server service enabled. If you select any other source object, Data Profiling lists all the runtime environments that has the Data Integration Server service enabled.

Serverless runtime environment

A serverless runtime environment is an advanced serverless deployment solution that does not require downloading, installing, configuring, and maintaining a Secure Agent or Secure Agent group. You can use a

serverless runtime environment in the same way that you use a runtime environment when you configure a connection or some types of tasks in Data Profiling.

The following table lists the options that you can choose in the **Serverless Usage Properties** area:

Option	Description
Max Compute Units	Maximum number of serverless compute units corresponding to machine resources that the task can use. Overrides the corresponding property in the serverless runtime environment. By default, for a data profiling task, the maximum number of compute units is set to two.
Task Timeout	Amount of time in minutes to wait for the task to complete before it is terminated. The timeout ensures that serverless compute units are not unproductive when the task hangs. By default, the timeout is the value that is configured in the serverless runtime environment.

For more information, see the Runtime environments document.

Advanced clusters

An advanced cluster is a Kubernetes cluster that provides a distributed processing environment on the cloud. Fully-managed and self-service clusters can run data logic using a scalable architecture, while local clusters use a single node to quickly onboard projects for advanced use cases.

To use an advanced cluster, you perform the following steps:

1. Set up your cloud environment so that the Secure Agent can connect to and access cloud resources.
2. In Administrator, create an advanced configuration to define the cluster and the cloud resources.
3. In Monitor, monitor cluster health and activity while developers in your organization create and run jobs on the cloud.

To run a profile with Avro, Parquet, or JSON file format type, you need to configure the Amazon S3 V2 or Azure Data Lake Store connection with the respective Advanced cluster.

For more information about setting up the AWS, Microsoft Azure, and local cluster, see the Advanced Clusters document.

Email notification options

When you run a profile, you can choose to send email notifications based on the profile job status. The job status for which you can send notifications include warning, failure, and success. You can choose default and custom email addresses to send the notifications.

The following table lists the email notification options that you can choose for a profile:

Option	Description
Use the default email notification options for my organization	Data Profiling sends the email notification to the default email address of the logged-in user. You can configure the default email addresses on the Organization page of Administrator. For more information, see *Administrator* in the Administrator help.
Use custom email notification options for this task	Choose to send the notifications to different email addresses based on the job status. Enter one or more comma-separated, valid email addresses to receive email notification for the following job status: - Failure - Warning - Success

Advanced options

You can configure the advanced options to detect outliers, infer the date and time, and infer other profile-related parameters.

The following table lists the advanced options that you can configure for a profile:

Option	Description
Maximum Number of Value Frequency Pairs	Number of column values with the highest frequencies appear in the profile results. Default is 500. For example, if you set the value to 100, only the top 100 values appear in the profile results. **Note:** If you do not want to save the value frequency information of a profile in the profiling warehouse, set the value to 0.
Maximum Number of Patterns	Number of patterns with the maximum number of occurrences appear in the profile results. The rest of the patterns appear under the **Patterns** > **Others** category on the **Results** area. Default is 10. For example, if you set the value to 3, the top 3 patterns appear with their statistics, and the rest of the patterns are consolidated under the **Others** category.
Pattern Threshold Percentage	Maximum percentage of values used to derive a pattern in the profile results. Default is 5. For example, when you set the value to 4, the patterns that are 4% and higher appear individually with their statistics and the rest of the patterns are consolidated under the **Others** category.
Infer Date and Time	Infers the date and time for a column of date or time data type. Default is Yes.
Detect Outliers	Detects pattern and value frequency outliers in the source object. Default is Yes.

Option	Description
Minimum Number of Rows for Split Process per Column	If the source object contains more rows than the minimum number of rows that you enter here, Data Profiling uses one subtask for each source column when the profile is run. Default is 100,000,000.
Maximum Number of Columns per Mapping	Number of columns for each mapping when the number of source rows is fewer than the **Minimum Number of Rows for Split Processing per Column** value. Default is 50.
Maximum Memory per Mapping*	Maximum amount of memory that you want to allocate for each mapping. Default is 512 MB.
Default buffer block size	Size of buffer blocks used to move data blocks from sources to targets. Default is Auto. Enter one of the following options: - Auto. Uses automatic memory settings. When you use Auto, configure **Maximum Memory per Mapping**. - A numeric value. Enter the numeric value that you want to use. The default unit of measure is bytes. Append KB, MB, or GB to the value to specify a different unit of measure. For example, 512MB.
DTM Buffer Size	Amount of memory allocated to the task from the DTM process. Default is Auto. By default, a minimum of 12 MB is allocated to the buffer at run time. Use one of the following options: - Auto. Uses automatic memory settings. When you use Auto, configure **Maximum Memory per Mapping**. - A numeric value. Enter the numeric value that you want to use. The default unit of measure is bytes. Append KB, MB, or GB to the value to specify a different unit of measure. For example, 512MB.
Line Sequential Buffer Length	Number of bytes that the task reads for each row in a flat file source. Default is 1024.
* The mapping is a type of subtask. Data Profiling creates and runs for a data profiling task to process the data concurrently.	

The default values for the advanced options have been derived to provide the best performance. However, you can configure the values based on your requirements. To optimize the data profiling task performance, see Chapter 4, “Tuning data profiling task performance” on page 98.

Note: You can configure the following advanced options for a profile with Avro or Parquet source objects:

- Maximum Number of Value Frequency Pairs
- Maximum patterns
- Threshold percentage for patterns
- Detect outliers

Session Options

You can configure the session options to specify the number of non-fatal errors the data profiling tasks can encounter before Data Profiling stops the session.

You can configure the following session option for a profile:

- Stop on Errors. Enter the number of non-fatal errors the data profiling tasks can encounter before Data Profiling stops the session. Enter 0 to continue the session irrespective of the number of non-fatal errors. Default is 100. For example, if you edit metadata of a delimited file and a metadata conversion error occurs. You can enter the number of rows that have the conversion error so that data profiling task can ignore these errors.

Insights

Insights is a method for discovering data quality issues in your data. The issues can range from anomalous data values to complex inconsistencies. Insights in Data Profiling automates the process to detect data quality issues.

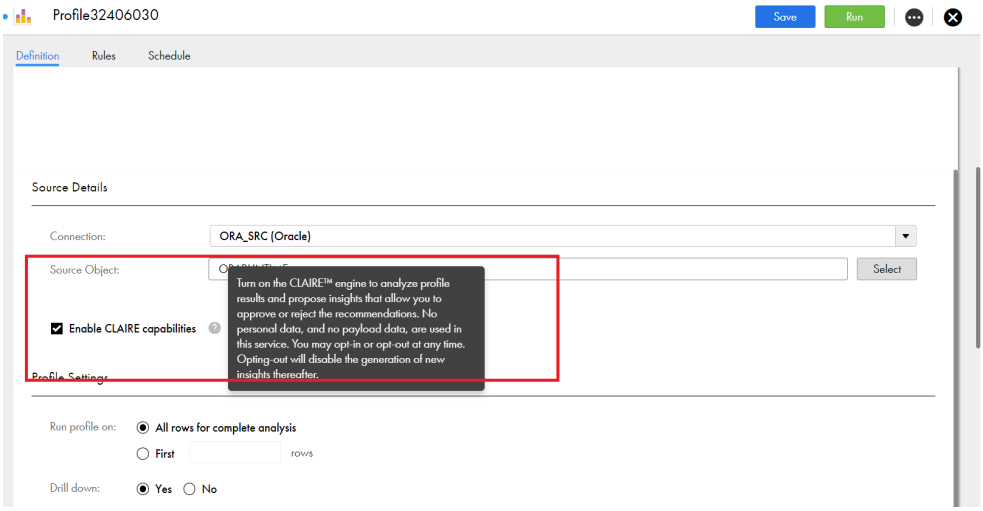
The CLAIRE™ artificial intelligence engine provides insights and generates recommendations for your data that you can approve or reject. CLAIRE can run automatically on profiles and display recommendations that enable you to detect data quality issues. If you approve the recommendations, the Data Quality and Data Profiling services automatically create data quality rules and apply them to your profile to detect the issues.

Note: Insights are available for all source objects that support profiling, except for Avro, Parquet, and JSON source objects that include hierarchical columns.

Generate insights

You can generate and view the inferred insights of a profile on the **Insights** tab.

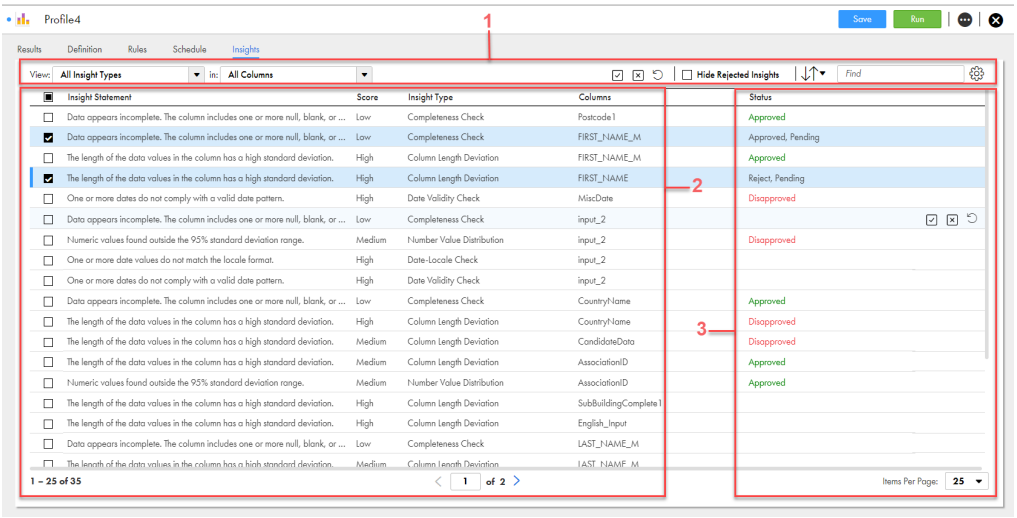
1. In Data Profiling, select the **Enable CLAIRE capabilities** checkbox on the **Definition** tab as shown in the following image:



   By default, the **Enable CLAIRE capabilities** checkbox is enabled for profiles.

2. After you enable CLAIRE capabilities, save and run the profile.

The inferred insights generated for the profile appear on the **Insights** tab.
The following image shows the areas on the **Insights** tab of a profile:



1. View options
2. Insights
3. Insight status

View options

The following table lists the view and sort options:

Option	Description
View	Shows the following options: - All Insight Types. View all the insight types in the profile run. - Completeness Check. View the insights generated where the data appears incomplete. The column can include one or more null, blank, or empty values or values that contain only zeros. - Uniqueness Check. View the insights generated when the majority of the data values in the column are unique. - Column Length Deviation. View the insights generated when length of the data values in the column has a high standard deviation. - Number Value Distribution. View the insights generated for numeric values found outside the 95% standard deviation range. - Date Validity Check. View the insights generated for one or more dates that do not comply with a valid date pattern. - Date-Locale Check. View the insights generated for one or more date values that do not match the locale format. - Day-Date Distribution. View the insights generated for unusual distribution of day values in a date column. - Month-Date Distribution. View the insights generated for unusual distribution of month values in a date column. - Year-Date Distribution. View the insights generated for unusual distribution of year values in a date column. - Top Pattern Stability. View the insights generated when the topmost pattern of the column decreases by a large amount when compared to the previous profile run. - Spelling Analysis. View the insights generated for the data values that are phonetically similar and contain inconsistent spelling. - Distribution Shift. View the insights generated for the distribution of the data that might have shifted more than expected based on the mean and standard deviation of the profile that has run over multiple times. - Column Token Deviation. View the insights generated for the number of tokens in a column that has a high standard deviation. - Special Characters. View the insights generated when columns with special characters are not included in the top 80% of the patterns. - Null Date Analysis. View the insights generated when the string data type columns contain values from a default date pattern, such as 00/00/0000 and 99/99/9999.
In	Shows the following options: - All Columns. View the insight types for all columns in the profile run. - Columns included in the profile run. Choose a filter in the **In** option after you choose a filter in the **View** option.
Sort	Choose any of the following options to sort the insights in ascending or descending order: - Insight Statement - Score - Insight Type - Columns
Find	Enter a keyword to view the relevant search results.
Menu	Choose Comfortable, Cozy, or Compact to adjust the row width on the Insights area.

Insights

The Insights area displays the CLAIRE generated recommendations for your data that you can approve or reject.

The following table lists the properties that you can view on the **Insights** tab:

Property	Description
Insight Statement	Description or statement that explains the inferred insight.
Score	Shows the following scores for the inferred insights: - High. Data anomaly is high. - Medium. Data anomaly is medium. - Low. Data anomaly is low. You can classify and review the inferred insights from higher sores to lower scores.

Insight Type	Shows the following types of insights:
	- Completeness Check. Data appears incomplete. The column includes one or more null, blank, or empty values or values that contain only zeros.
	- Uniqueness Check. The majority of the data values in the column are unique.
	- Column Length Deviation. The length of the data values in the column has a high standard deviation.
	- Number Value Distribution. Numeric values found outside the 95% standard deviation range.
	- Date Validity Check. One or more dates do not comply with a valid date pattern.
	- Date-Locale Check. One or more date values do not match the locale format.
	- Day-Date Distribution. Unusual distribution of day values in a date column.
	- Month-Date Distribution. Unusual distribution of month values in a date column.
	- Year-Date Distribution. Unusual distribution of year values in a date column.
	- Completeness Variation.
	  - For integer or decimal data types: Unusual variation in the number of null values and values that contain only zeros in the column between the current profile run and the previous one to five profile runs.
	  - For string, date, or timestamp data types: Unusual variation in the number of blank values, null values, and empty values in the column between the current profile run and the previous one to five profile runs.
	- Distinct Variation. Greater than 70% increase in the number of distinct values in the column between the current profile run and the previous one to five profile runs.
	- MinMax Variance. Greater than 70% increase in the difference between the minimum and maximum values in the column across the current and previous profile runs.
	- Top Pattern Stability. The topmost pattern of the column decreased by a large amount when compared to the previous profile run.
	- Spelling Analysis. The data values that are phonetically similar and contain inconsistent spelling.
	- Distribution Shift. The distribution of the data that might have shifted more than expected based on the mean and standard deviation of the profile that has run over multiple times.
	- Column Token Deviation. The number of tokens in a column that has a high standard deviation.
	- Special Characters. Columns with special characters that are not included in the top 80% of the patterns.
	- Null Date Analysis. String data type columns that might contain values from a default date pattern, such as 00/00/0000 and 99/99/9999.

Columns	Column name for which the insight is relevant.
Status	Status of the insight. When insights are generated for the first time, the status appears blank.

Each insight type follows an algorithm to look into columns, independently of other columns on the data set. These algorithms are based on the metrics calculated on every profile run. The following table lists the algorithm and the logic used to arrive at the scores for the inferred insights:

Insight Type	Algorithm	Score Interpretation
Completeness Check	Computes the percentage of total rows with null values, blank values, empty values, or values that contain only zeros in a column. This insight type is applicable for columns with any of the following data types: - String - Date - Integer	- 0 - OK - 92 to 100 - OK - 0 to 3 - Low - 3 to 5 - Medium - 5 to 8 - High
Uniqueness Check	Computes the percentage of non-unique rows based on the following formula: Percentage of non-unique rows = (Total Rows - Unique Rows) / Total Rows * 100 Insights are generated if the computed percentage of non-unique rows is less than 3%. **Note:** If a column contains one or more null values, then the insight is not generated. This insight type is applicable for columns with any of the following data types: - String - Date - Integer	- 0 - OK - 0 to 2 - High - 2 to 3 - Low
Column Length Deviation	Computes the length of alphanumeric values or numeric values on value frequency that falls more than two times the standard deviation from the mean value. This insight type is applicable for columns with any of the following data types: - String - Integer - Decimal	- 0 - OK - 0 to 1 - Low - 1 to 5 - Medium - 5 - High
Number Value Distribution	Computes the percentage of value frequency values in relation to the total number of rows profiled that falls more than two times the standard deviation or falls out of 95% of the mean value. This insight type is applicable for columns with any of the following data types: - String with all numeric patterns - Date - Integer	- 0 - OK - 15 to 100 - OK - 0 to 1 - Low - 1 to 5 - Medium - 5 to15 - High **Note:** The score can never be 100%.
Date Validity Check	Checks only for columns that have Date as the inferred data type and computes the percentage of values with dates that do not comply with a valid date pattern. Null values are ignored in the computation. This insight type is applicable for columns that have String with date content as the data type.	- 0 - OK - 0 to 5 - Low - 5 to 10 - Medium - 10 - High
Date-Locale Check	Checks for columns that have one or more date values that do not match the locale format. Computes the number of values that follow different date locale formats. This insight type is applicable for columns with String data type.	- 1 - OK - 2 - Medium - 3 - High

Insight Type	Algorithm	Score Interpretation
Day-Date Distribution	Extracts the day for the dates on the value frequency and calculates the mean and standard deviation. Computes the dates where the days fall over two times the standard deviation or falls out of 95% of the mean value. This insight type is applicable for columns with any of the following data types: - String with date patterns - Date - Timestamp	- 0 - OK - 15 to 100 - OK - 0 to 1 - Low - 1 to 5 - Medium - 5 to15 - High **Note:** The score can never be 100%.
Month-Date Distribution	Extracts the month for the dates on the value frequency and calculates the mean and standard deviation. Computes the dates where the days fall over two times the standard deviation or falls out of 95% of the mean value. This insight type is applicable for columns with any of the following data types: - String with date patterns - Date - Timestamp	- 0 - OK - 15 to 100 - OK - 0 to 1 - Low - 1 to 5 - Medium - 5 to15 - High **Note:** The score can never be 100%.
Year-Date Distribution	Extracts the year for the dates on the value frequency and calculates the mean and standard deviation. Computes the dates where the days fall over two times the standard deviation or falls out of 95% of the mean value. This insight type is applicable for columns with any of the following data types: - String with date patterns - Date - Timestamp	- 0 - OK - 15 to 100 - OK - 0 to 1 - Low - 1 to 5 - Medium - 5 to15 - High **Note:** The score can never be 100%.
Completeness Variation	Computes the variation on the number of null values and the values that contain only zeros in the column between the current profile run and the truncated mean of the last five profile runs, discarding the lowest and highest values. Uses the actual mean if there are less than four previous profile runs. Insights are not generated if there are no previous profile runs. Percentage of completeness variation = (Current Mean - Previous Mean) / Previous Mean * 100 - If the previous mean value is zero, then the completeness variation percentage increases to 100%. - If the completeness variation percentage is negative, then insights are not generated. The following values are considered as null values for data types: - Integer - 0 and null values - Decimal - 0.0 and null values - String - Blank or empty string and null values - Date - Null values - Timestamp - Null values	- [0 to 80] - OK - (80 to 90] - Medium - (90 to ∞) - High

Insight Type	Algorithm	Score Interpretation
Distinct Variation	Checks if there is more than 70% increase on the number of distinct values in the column between the current profile run and the truncated mean of the last five profile runs, discarding the lowest and highest values. Uses the actual mean if there are less than four previous profile runs. Insights are not generated if there are no previous profile runs. Percentage of distinct variation = (Current Mean - Previous Mean) / Previous Mean * 100 If the previous mean value is zero, then the distinct variation percentage increases to +∞. If the distinct variation percentage is negative, then insights are not generated.	- (-∞ to 70] - OK - (70 to 90] - Low - (90 to 200] - Medium - (200 to +∞) - High
MinMax Variance	Checks if there is more than 70% increase on the difference between the minimum values and the maximum values in the column when compared to the previous profile run. CLAIRE does not consider columns for insight recommendations in the following scenarios: - Difference between the minimum values and the maximum values in the column when compared to the previous profile run decreases. - Sources that have less than 1000 rows. - Columns that transition from 100% null to values. Percentage of min max variation = (Delta Current - Delta Previous) / Delta Previous * 100 Where: - Delta Previous = Maximum value in the first run - Minimum value in the first run - Delta Current = Maximum value in the second run - Minimum value in the second run For example, the following are the minimum and maximum values in the po_create_date column for two profile runs: - Previous run: Minimum = 01/01/1998, Maximum = 03/03/2013 - Current run: Minimum = 02/01/2003, Maximum = 12/07/2025 Delta Previous = 5540 days Delta Current = 8345 days Percentage of min max variation = (8345 - 5540) / 5540 = 50.6% The 50.6% score interprets the data anomaly for the column as OK.	- [0 to 70] - OK - (70 to 100] - Medium - (100 to ∞) - High
Top Pattern Stability	Checks if the top pattern with ≥30% compliance decreases by a large amount in comparison to the previous profile run. A large decrease may indicate that shape of data changed more than expected. The decrease is measured as a negative number computed using the following formula: `CurrentPercent - PreviousPercent / PreviousPercent * 100` The insight considers columns that contain a major pattern in the previous run. The same filter must be used for the both runs	- (-99, -70] - High - (-70, -60] - Medium - (-60, -30] Low - (-30, 0] - OK

Insight Type	Algorithm	Score Interpretation
Spelling Analysis	Creates a fingerprint for each string value and compares the number of non-null unique fingerprints to the number of non-null values. CLAIRE runs the insight if the difference as a percentage is too high, which indicates several misspellings. To qualify, the top 80% patterns must contain only letters (X) and up to 3 spaces and hyphens. This is to accommodate names. The insights get generated if 95% of the value frequencies in the values have five or more characters.	- [0,0.5] - OK - (0.5,1] - Low - (1, 2] - Medium - (2,100] - High
Distribution Shift	Tracks the mean and standard distribution of values over four or more profiles. The expectation is either the mean and standard deviation remains constant or shifts consistently up or down. For example, a table containing population size information that might shift consistently up or down at the same rate. This insight type is applicable for columns with any of the following data types: - Integer - Decimal	- [0,2] - OK - (2,3] - Medium - (3,∞] - High
Column Token Deviation	The number of tokens in the value frequency string values that fall more than two standard deviations from the mean. A token is any sequence of alpha-numeric characters separated by white space and the following special characters: . , / -.	- 0 - OK - (0,1] - Low - (1, 5] - Medium - (5,100] - High
Special Characters	Checks data for special characters that are not included in the top 80% of the patterns. CLAIRE considers this data anomalous. Additionally, CLAIRE does not consider the string data types when Data Profiling infers the numeric data type such as decimal, integer, or float as 100%.	- 0 - OK - (0,1] - Low - (1, 3] - Medium - (3,100] - High
Null Date Analysis	Checks string data type columns that might include one of all the zeros or nine values from a default date pattern. The insight type is applicable for columns of string data type. If a string data type column contains all of the zeros and nines from the default date pattern, the insight considers the values as invalid. For example, - 0000-00-00 or 9999-99-99 (year-month-day or year-day-month) - 00/00/0000 or 99/99/9999 (month/day/year or day/month/year) - 00000000 or 9999999 (YYYYMMDD) If a string data type column contains a valid date, month, or year part from the default date pattern, the insight considers the values as valid. For example, - 21/99/9999 - 99/02/9999 - 99/99/1994 The insight also considers a NULL value as a valid date pattern.	- 0 - OK - (0,1] - Low - (1,2] - Medium - (2,100] - High

Insight status

The Insights status area displays the status of the insights. When insights are generated for the first time, the status appears blank. You can approve or reject the generated insights and save the profile.

When you approve an insight, the status of the insight changes to "Approved, Pending". When you save the profile, the status of the insight changes to "Approved". When you reject an insight, the status of the insight changes to "Reject, Pending". When you save the profile, the status of the insight changes to "Disapproved".

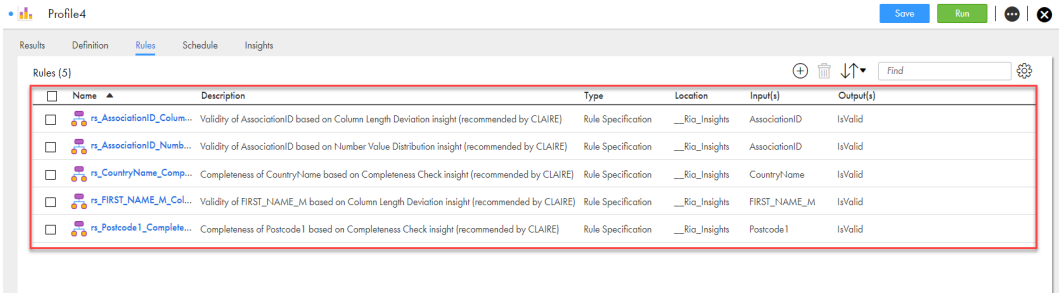
Review and act on insights

You can review the inferred insights generated by the CLAIRE engine in Data Profiling. Hover the mouse over the insight to approve or reject the insight. After you review the insights, you need to save and run the profile.

You can click on the ☑ or ☒ icons to approve or reject the insights. When you approve an insight, a rule specification is created in Data Quality and the rule is assigned to the profile automatically. After approval, the status of the insight changes to "Approved, Pending". When you save the profile, the status of the insight changes to "Approved". A rule specification is automatically created and appears on the **Rules** tab for the columns to which the insight is referred to. You must save the profile to persist the status and the rule association. Once an insight is approved, the insight cannot be removed from the profile unless you delete the corresponding rule specification from the **Rules** tab of the profile.

When you approve an insight, a rule specification is created that monitors the quality of the column corresponding to the insight. For example, an algorithm to detect outlier based on the value frequency length. CLAIRE recommends that any value frequency with length greater than 5 is invalid. A new rule specification is automatically created with the logic to verify for values greater than 5 and tags them as invalid. The rule specification is also automatically assigned to the column in the profile.

The following image shows the automatically created rule specifications assigned to the corresponding source columns in the profile on the **Rules** tab:



Note:

- The name of the automatically created rules follow the rs_<source column name>_<insight type>_< sequential number> pattern.
- The description of the automatically created rules are appended with the "recommended by CLAIRE" text.

When you reject an insight, the status of the insight changes to "Reject, Pending". You must save the profile to complete the rejection. When you save the profile, the status of the insight changes to "Disapproved".

To review multiple insight statements, select the insights and click on the ☑ or ☒ icons on the top of the Insights area.

You can reset a pending insight review. An insight with "Approved, Pending" or "Reject, Pending" status can be cleared. Hover the mouse over the insight and click on the ↺ icon to reset the review. The status of the insight disappears. You can click on the ↺ icon on the top of the Insights area to reset all the pending insights. You can select the **Hide Rejected Insights** checkbox to hide all the rejected insights.

Note: If an insight is approved or rejected, the same algorithm is not used again for the same columns.

After review of the inferred insights, you can drill down and view the anomalous data values and complex inconsistencies on the **Results** tab. For more information about profile results, see Chapter 3, “Profile results” on page 66. You can also create rule occurrences and view scorecards to measure data quality scores and monitor data quality progress for profiles. For more information about rule occurrences and scorecards, see “Rule occurrences and scorecards” on page 49.

Creating a profile

You can create a profile to view and analyze the content and structure of a source object.

1. In Data Profiling, click **New**.
2. In the **New** dialog box, click **Data Profiling Task**.

   The **Definition** tab for the profile appears.
3. On the **Definition** tab, enter the asset, source, and profile details. You can also choose columns and add a filter for the profile.
   - You can add a cleanse, labeler, parse, or rule specification asset to view the impact of the corresponding data quality operations on the source data.
4. On the **Rules** tab, add one or more Data Quality assets as rules to the profile.
5. On the **Schedules** tab, optionally choose a runtime environment and schedule. You can also change the default email notification options and advanced options for the profile run as necessary.
6. Choose one of the following options to save and run the profile:
   - Click **Save** to save the profile.
   - Click **Run** to save and run the profile.
   - Save the profile and choose a schedule on the **Schedules** tab to run the profile.

Exception management task

An exception record is a record that contains unresolved data quality issues. You can use a rule specification to identify exception records in your data set as part of an exception management process.

You can create an exception task from the profiling task. When you can create a data profiling task, you can add one or more rule specifications as rules to the task. Configure the profiling task to read the data set that contains the exception records.

You can create an exception task in Data Profiling or in Data Quality. Add one or more rules from the profiling task to the exception task. Include the rule specification that you configured earlier to find and update exception records.

For more information, see the Exception Management guide in the Data Quality documentation.

CHAPTER 3

Profile results

When you run a data profiling task on a data source, the profile extracts and displays the column statistics from the data source, such as null values, distinct values, data types, and patterns. You can analyze the profile results to make business decisions.

For example, you are a data analyst and you want to analyze and report potential data issues in the Customer table such as the completeness and validity of addresses and email IDs. The accuracy of this critical data impacts the effectiveness of the company's marketing campaign. To accomplish this task, you create a data profiling task on the table, add Verifier and Cleanse assets as rules, and then run the profile. You can view the results or export them to a file to analyze the data.

When you open a profile after you run it, the following tabs appear:

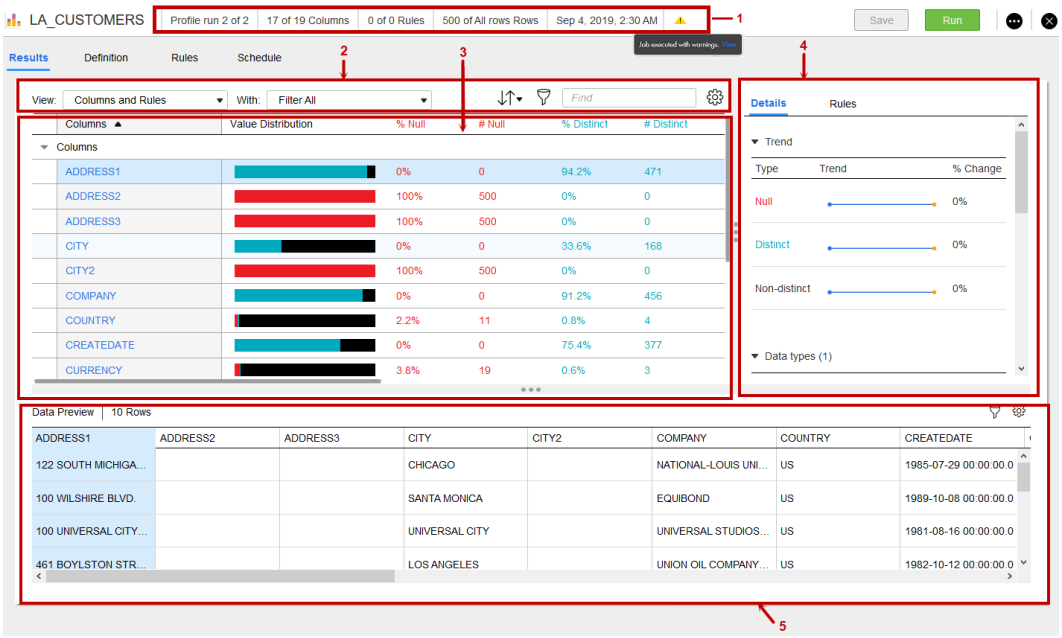
- Results. You can view the profile results on this tab.
- Definition. You can view and edit the profile definition on this tab.
- Rules. You can view and edit the rules on this tab.
- Schedule. You can view and edit the schedule, email notification options, and advanced options on this tab.

You can edit the profile-related options, rules, schedules, and advanced options as necessary and run the profile again. You can run a profile multiple times and view the profile results for each run. You can also compare profile runs, compare columns in a profile run, and export the profile results to a Microsoft Excel file.

View profile results for a profile run

You can view the profile results for a profile run on the **Results** tab. The tab appears after you run the profile. The **Results** tab contains a header area with profile run details, filter and sort area, profile results area, details and rules area, and data preview area. The profile results area shows the profile results for all the columns and rules in summary view. When you click a column, a detailed view of the profile results for the column appears in the area.

The following image shows the areas on the **Results** tab:



1. Header
2. Filter and sort
3. Profile results
4. Details and Rules
5. Data Preview

Note: You can also open a profile from the **Explore** page in Data Quality and perform the following:

- Edit a profile
- Run a profile
- View profile results
- Create and run queries on the source object
- Drill down on the profile results

Header

The header area shows the profile run details, which include the profile name, run number, number of columns and rules in the profile run, number of rows in the profile run, and run timestamp. The header area also displays a warning icon if the profile job runs with a warning. To view the job that ran with a warning, hover over the warning icon, and then click **View**.

Filter and sort

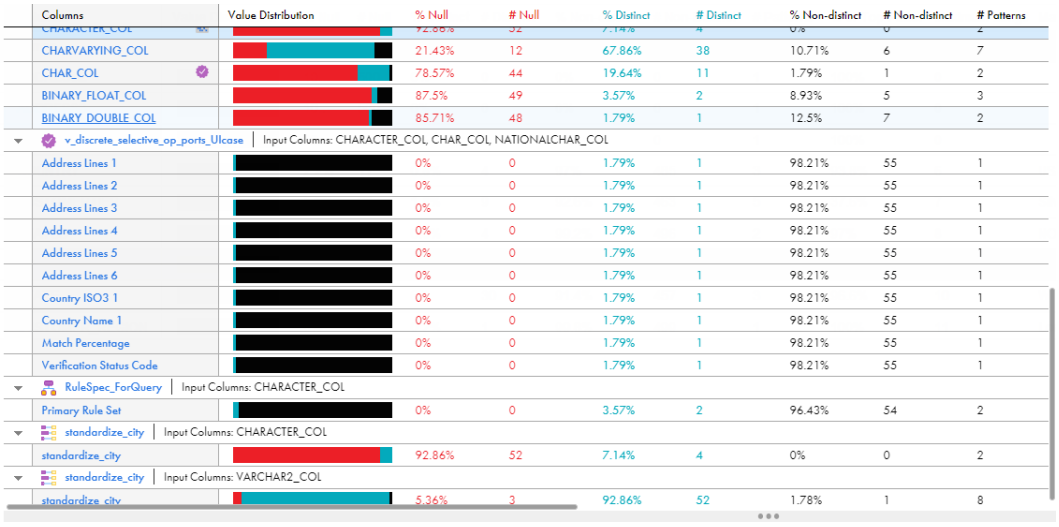
The following table lists the filter and sort options:

Option	Description
View	Shows the following options: - Columns and Rules. View the results for all the columns and rules in the profile run. - Columns. View the results for the columns in the profile run. - Rules. View the results for the rules in the profile run.
With	Shows the following options: - All Statistics. View the complete profile results for the profile run. - 100% Null <number_of_rows>. View the results for the columns and rules that have only null values. - 100% Distinct <number_of_rows>. View the results for the columns and rules that have only distinct values. - 100% Constant <number_of_rows>. View the results for the columns and rules that have the same value for all the rows. - Conflicting Data types <number_of_rows>. View the results for the columns and rules where the documented data type and inferred data type do not match. - Value Frequency Outliers <number_of_rows>. View the results for the columns or rules with value frequency outliers. - Pattern Outliers <number_of_rows>. View the results for the columns or rules with pattern outliers. Choose a filter in the **With** option after you choose a filter in the **View** option.
Sort	Choose a column statistic to sort the results in ascending or descending order.
Filter	To filter the results, you can perform one or both of the following actions: - Add a column and enter a valid value. Add more columns with valid values as necessary. - Add a column statistic and enter a valid value. Add more column statistics with valid values as necessary.
Find	Enter a keyword to view the relevant search results.
Menu	Choose Comfortable, Cozy, or Compact to adjust the row width in the profile results area.

Profile results: summary view

When you open a data profiling task or choose a profile run, the summary view of the profile results appears. The summary view shows all the columns and rules and their statistics in the profile run.

The following image shows the summary view of profile results for columns and rules and the results are sorted by minimum value:



Columns	Value Distribution	% Null	# Null	% Distinct	# Distinct	% Non-distinct	# Non-distinct	# Patterns
CHARACTER_COL		92.86%	52	7.14%	4	0%	0	2
CHARVARYING_COL		21.43%	12	67.86%	38	10.71%	6	7
CHAR_COL		78.57%	44	19.64%	11	1.79%	1	2
BINARY_FLOAT_COL		87.5%	49	3.57%	2	8.93%	5	3
BINARY_DOUBLE_COL		85.71%	48	1.79%	1	12.5%	7	2
v_discrete_selective_op_ports_Ulcase | Input Columns: CHARACTER_COL, CHAR_COL, NATIONALCHAR_COL								
Address Lines 1		0%	0	1.79%	1	98.21%	55	1
Address Lines 2		0%	0	1.79%	1	98.21%	55	1
Address Lines 3		0%	0	1.79%	1	98.21%	55	1
Address Lines 4		0%	0	1.79%	1	98.21%	55	1
Address Lines 5		0%	0	1.79%	1	98.21%	55	1
Address Lines 6		0%	0	1.79%	1	98.21%	55	1
Country ISO3 1		0%	0	1.79%	1	98.21%	55	1
Country Name 1		0%	0	1.79%	1	98.21%	55	1
Match Percentage		0%	0	1.79%	1	98.21%	55	1
Verification Status Code		0%	0	1.79%	1	98.21%	55	1
RuleSpec_ForQuery | Input Columns: CHARACTER_COL								
Primary Rule Set		0%	0	3.57%	2	96.43%	54	2
standardize_city | Input Columns: CHARACTER_COL								
standardize_city		92.86%	52	7.14%	4	0%	0	2
standardize_city | Input Columns: VARCHAR2_COL								
standardize_city		5.36%	3	92.86%	52	1.78%	1	8

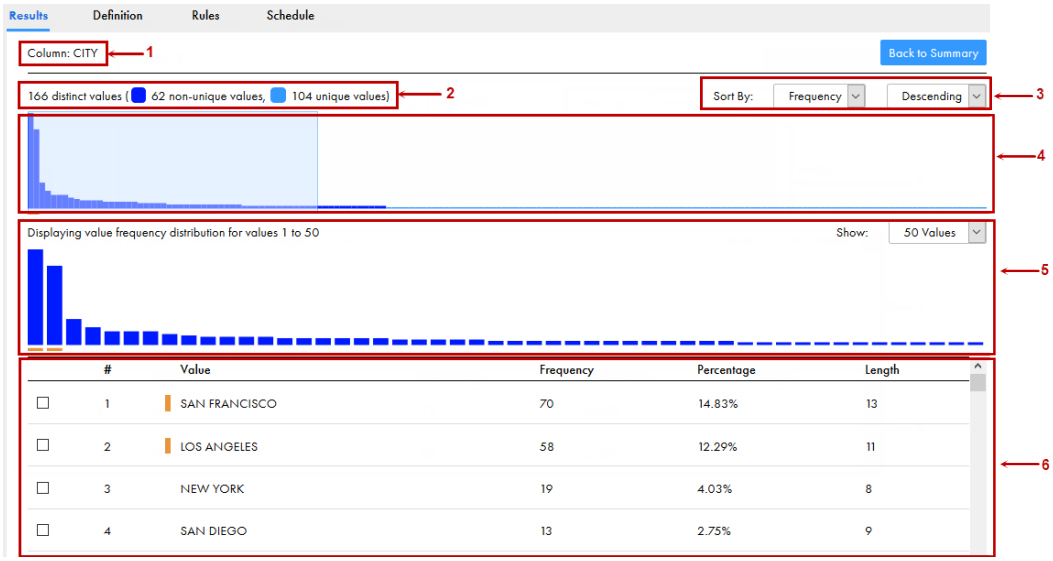
You can view the columns and rules area in collapsible sections. You can view the value distribution, number and percentage of null, distinct, and non-distinct values, number of patterns, percentage of top pattern, maximum value and length, and minimum value and length in the column or rule.

You can sort the columns and rules based on one of the statistics. To sort the columns and rules, click one of the statistics. For example, if you want to view the maximum value in ascending order, click *Maximum Value*. The columns are sorted in ascending order of maximum values.

Profile results: detailed view

When you click a column in the summary view, the detailed view of the profile results for the column appears. The area shows the column values in a graphical mode. The null values appear as red vertical bars.

The following image shows the detailed view of the profile results area:



1. Column or rule output name
2. Number of distinct values, which includes non-unique values and unique values
3. Sort By
4. Bar chart
5. Detailed chart
6. Value distribution table

The following table lists the properties in the detailed view:

Property	Description
Column <column_name> Rule <rule_output_name>	Shows the column name or rule output name.
Back to Summary	Click the button to go back to the summary view of profile results.
<total_number> distinct values (<number_of_non-unique_values>, <number_of_unique_values>	Shows the total number of distinct values in the column or rule. This property also shows the number of non-unique and unique values, with the color legend, in the column or rule.
Sort By	You can sort the value frequency distribution based on the date, integer, and decimal data types. Choose **Frequency** or **Value**, and then choose **Ascending** or **Descending** to sort the value frequency distribution as required.
Bar chart	Shows the values as a vertical bar chart. You can view a maximum of 16,000 values in the upper area. You can slide the slider over the values in the upper area. The lower area displays the values in the slider. The outlier values appear with an orange underline.

Property	Description
Detailed chart	Shows the values in the slider in the upper area. By default, 50 values appear in the lower area. You can choose to view 75 or 100 values at a time. The outlier values appear with an orange underline.
Value distribution table	Shows the following statistics in a tabular format: - #. Row or field number in the source object. - Value. List of values in the column. - Frequency. Number of times the value appears in the column, expressed as a number. - Percentage. Value percentage in the column. - Length. Length of the column value. The outlier values appear with a vertical bar.

By default, you can view 500 values in the detailed view. To increase or decrease the number of the values that you can view, configure the **Maximum Number of Value Frequency Pairs** option on the **Schedules** page and then run the profile.

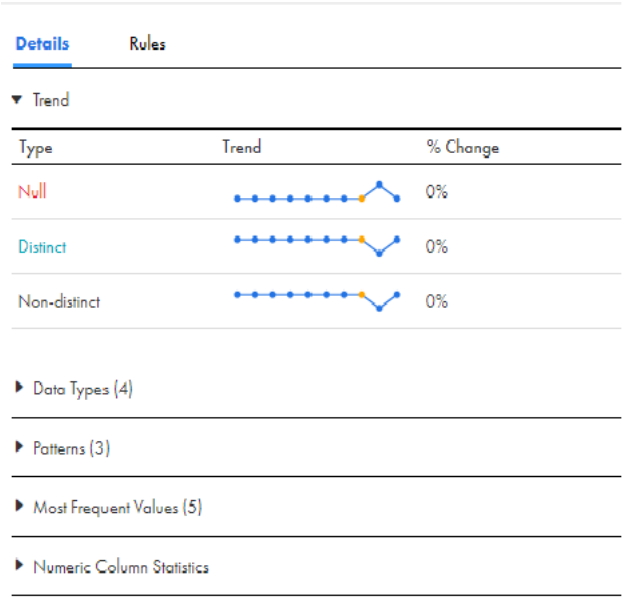
To view the drilldown results for a value, perform the following steps:

1. Select a value in the detailed view.
   The value appears as a filter in the **Data Preview** area.
2. Click **Apply.**
   The drill down results for the value appears in the **Data Preview** area.

Details and Rules

When you select a column or rule in the profile results area, the **Details** tab shows the trend of values across multiple profile runs, documented and inferred data types, inferred patterns, and most frequent values for the selected column. If the column has a numeric documented data type, the **Numeric Column Statistics** section also appears for the column. The **Rules** area shows the rules associated with the column in the profile run.

The following image shows the **Details and Rules** area:



The following table lists the sections and statistics that appear in the **Details and Rules** area:

Section	Description
Trend	Trend chart for percentage change in null, distinct, and non-distinct values. The trend chart shows the change for a maximum of 10 profile runs in a line chart. The chart displays the trend based on the profile run you have selected. For example, consider that there are 20 profile runs, and you are viewing the tenth profile run. In this case, the trend appears for five profile runs before the tenth profile run and four runs after the tenth profile run.
Data Types <number_of_inferred_data_type>	Shows the documented data type for the column in the data source. The section also shows the inferred data type, frequency percentage in which it appears in the column or rule, and a horizontal bar chart which is a virtual representation of data type distribution. Hover over the bar chart to view the number of rows that has the inferred data type. Select a data type to drill down and view the drilldown results in the **Data Preview** area.
Patterns <number_of_inferred_patterns>	Shows the inferred pattern, frequency percentage in which it appears in the column or rule, and a horizontal bar chart which is a virtual representation of pattern distribution. Hover over the bar chart to view the number of rows that has the inferred data type. Select a pattern to drill down and view the drilldown results in the **Data Preview** area.
Most Frequent Values	Shows the top five values that appear frequently in the column.
Numeric Column Statistics	Shows the following statistics for columns with numeric documented data type: - Average. Displays the average of the values for the column. - Sum. Displays the sum of all the values in the column. - Standard Deviation. Displays the standard deviation or variability between column values for all values of the column. - #Zero. Number of rows that contain the value 0 in the column or rule. - %Zero. Percentage of rows that contain the value 0 in the column or rule.
Rules	Shows the associated rules for the column and the rule details.

Data Preview

When you open a profile, the **Data Preview** area shows a maximum of 10 rows in the profile run results. When you select a column in the summary view of profile results, the column is highlighted in the area.

To view the drilldown results in the **Data Preview** area, perform one of the following actions:

- Choose a value in the detailed results area.
- Choose a pattern or data type in the **Details and Rules** area.

After you choose a value, pattern, or data type, it appears as a filter in the **Data Preview** area. Continue to add statistics or values if required. Click **Apply** to view the filtered drilldown results. Optionally, if you want to change the selected data type, pattern, or value, click the drop-down list to select the required statistics or values. Data Profiling creates and runs a subtask when you click **Apply** after you add or change a statistic or value.

For example, you are a data analyst and you want to view duplicate data for *SSN* in the Customer table. To accomplish this task, you perform the following actions:

1. Create a data profiling task for the Customer table.

2. Run the profile.
3. In the profiling results, click the pattern for *SSN* which is `999-99-9999`.

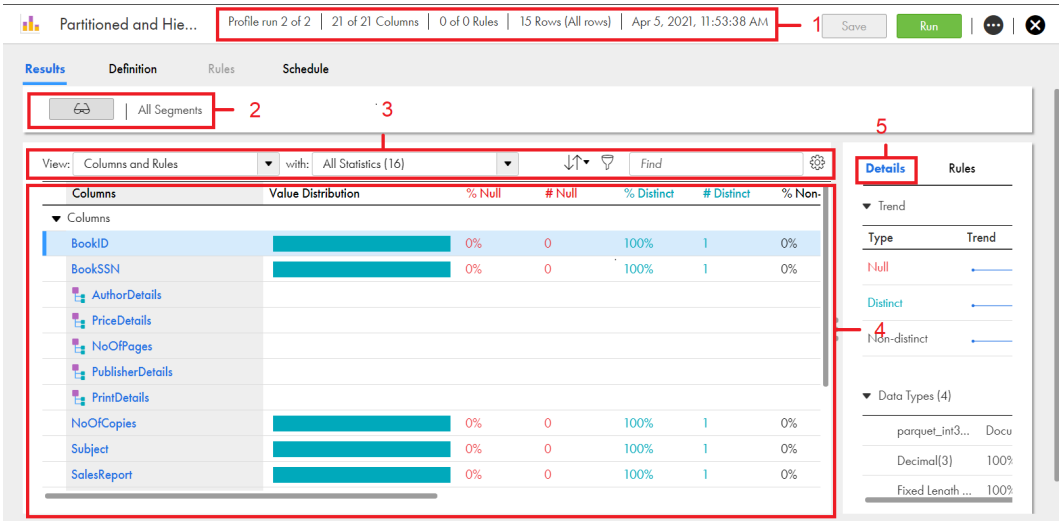
The **Data Preview** area shows all the rows with the pattern `999-99-9999`.

View tree previewer for hierarchical columns

You can view the tree previewer ( ) for a profile run that includes hierarchical columns from Avro and Parquet source objects in the **Results** tab. Hierarchical columns are classified as columns of data types such as an array, struct, map, or union. Use the tree previewer to view all the nested hierarchical columns within the hierarchical columns.

The **Results** tab contains a header area with profile run details, tree previewer area, filter and sort area, profile results area, and details area.

The following image shows the areas on the **Results** tab:

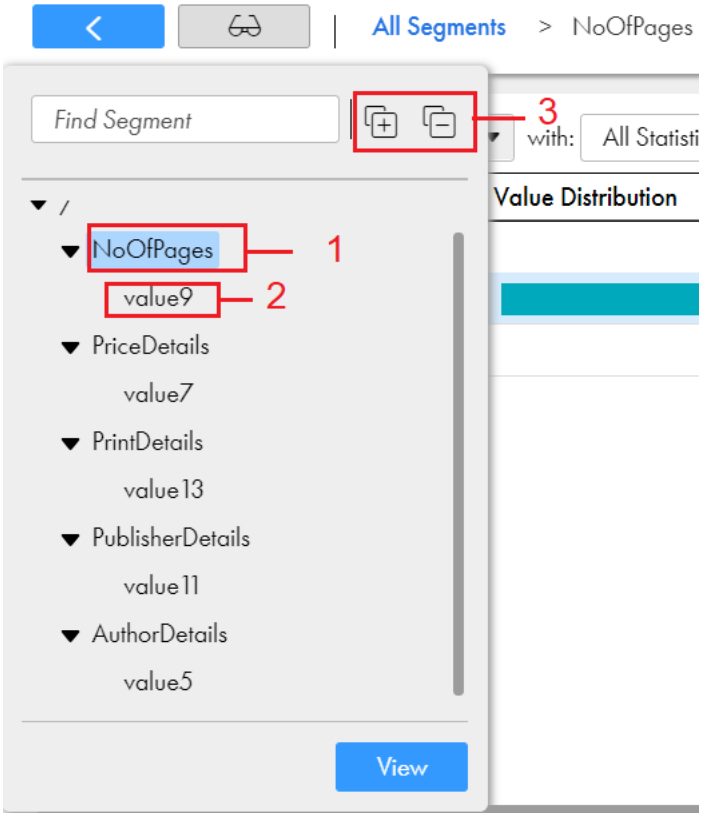


1. Header
2. Tree previewer
3. Filter and sort
4. Profile results
5. Details

Tree previewer

The tree previewer area displays the hierarchical and nested hierarchical columns. To view profile results for nested columns, click the tree previewer icon ( ), and then click the hierarchical column name or nested hierarchical column name from the tree previewer window.

The following image shows the sample tree previewer window:

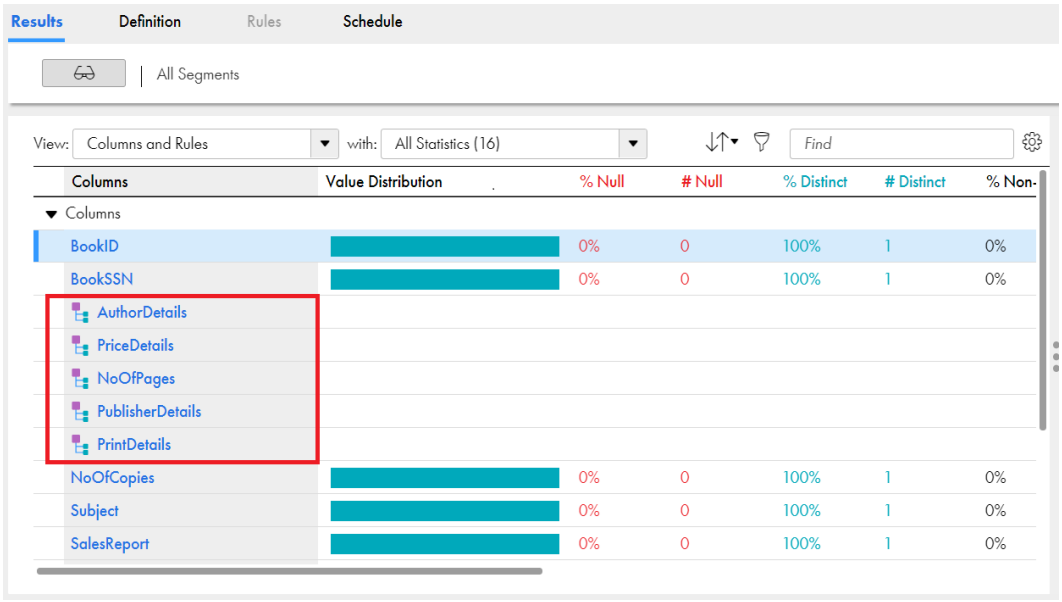


1. Hierarchical column
2. Nested hierarchical column
3. Expand and collapse view

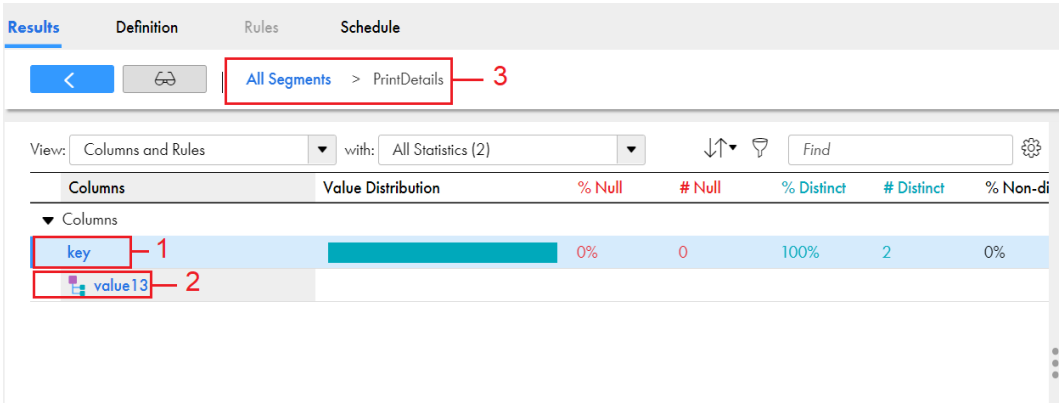
Profile results

The profile results area, by default, shows the profile results for all the columns in the summary view. When you click a non-hierarchical column, a detailed view of the profile results for the column appears. To view details of a hierarchical column, click the hierarchical column name. The hierarchical columns details view can include nested columns and other nested hierarchical columns. When you click the nested column name, a detailed view of the profile results for the nested column appears.

The following image shows the hierarchical columns in the **Results** tab:



The following image shows the nested hierarchical columns and nested columns in the **Results** tab:



1. Nested column
2. Nested hierarchical column
3. Breadcrumb to navigate between the segments and show the flow of the parent and child segments.

Edit a profile

You can edit a profile for the next profile run. You can change the profile definition, add or remove filters, add or remove rules, choose another runtime environment, edit schedule details, edit email address for notifications, and edit advanced options.

Definition

On the **Definition** tab, you can edit the following options for the next profile run:

Asset Details

Change or edit the **Name** and **Description** options.

The **Asset Details** area shows the location of the asset, user who created the profile, timestamp of the profile run, and used storage.

The **Used Storage (MB)** field shows the storage space consumed in the profiling warehouse for the profile results that you view on the **Results** tab. The storage space depends on the sampling option, columns, filter, and advanced options that you select for the profile run. It also depends on the identified number of unique values and outliers. Data Profiling stores the profile results in the profiling warehouse. The profiling warehouse is an Informatica Intelligent Cloud Services repository where Data Profiling stores the profile results.

Connection and Source Details

Switch between connections of the same database type in a profile definition. Data Profiling displays all the connections in the profile definition and does validation checks to validate the connection that you select. Choose a different connection or source object for the next profile run. The following list describes the different combinations with which you can edit a connection or source object:

- If you retain the same connection and choose a different source object that includes the same details as the previous source object, Data Profiling preserves the configuration settings of the columns that you select to profile, filters, and rules.
- If you retain the same connection and choose a different source object that does not include the same details as the previous source object, you need to select the columns to profile, and then fix the filters and rules that are not valid.
- If you choose a different connection and source object, you need to select the columns to profile, configure the filters and rules again from scratch.
- If you choose a different connection and a source object with same name and includes the same details as the previous source object, Data Profiling preserves the configuration settings of the columns that you select to profile, filters, and rules.

Profile Settings

Change **Run profile on** or **Drill down** options.

Columns

Select or clear one or more columns.

Filters

Choose a different filter. Optionally, you can create, add, or delete filters.

Rules

On the **Rules** tab, you can choose the rules for the next profile run. Optionally, you can add, or delete rules. When you change the source object, Data Profiling automatically assigns rules if the source object attributes match the configuration file parameters. You can include or exclude the rules to the profile.

Schedule

On the **Schedule** tab, you can edit the following options:

Schedule Details

Change the runtime environment and choose a schedule for the next profile run.

Email Notification Options

Change or edit the email notification options.

Advanced Options

Edit the advanced options. For more information about the advanced options, see “Advanced options” on page 54.

Statistics extracted from source objects

After you run a profile, the profile extracts column statistics, patterns, data types, value frequencies, and outliers for columns and rules.

Column Statistics

The following table lists the column statistics that you can view after you run a profile:

Property	Description
Columns Rules	Columns and rules in the profile run appear in collapsible sections. You can collapse or expand the section to view the columns and their statistics. When you click a metric for a column, the metric is highlighted in the **Data Preview** area. When you click a column name, the detailed view for the column appears.
Value Distribution	Distribution of null values, distinct values, and non-distinct values in a horizontal bar chart for a column or rule.
% Null	Percentage of rows with null values in the column or rule.
# Null	Number of null values in the column or rule.
% Distinct	Percentage of rows with distinct values in the column or rule.
# Distinct	Number of distinct values in the column or rule.
% Non-distinct	Percentage of rows with non-distinct values in the column or rule.
# Non-distinct	Number of non-distinct values in the column or rule.
# Patterns	Number of patterns in the column or rule.
% of Top Pattern	Percentage of rows with the most frequent pattern in the column or rule.
Maximum Length	Length of the longest value in the column.
Maximum Value	Highest value in the column.
Minimum Length	Length of the shortest value in the column.
Minimum Value	Lowest value in the column.
% Blank	Has no value in the column or rule.
# Blank	Percentage of rows that have no value in the column or rule.

Patterns

You can view inferred patterns after you run a profile.

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.
'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. Data Profiling 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. Data Profiling 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 $.

Data Types

You can view the documented data type and inferred data types after you run a profile.

Value frequencies

You can view value frequencies for each column after you run a profile in summary view and detailed view of profile results.

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.

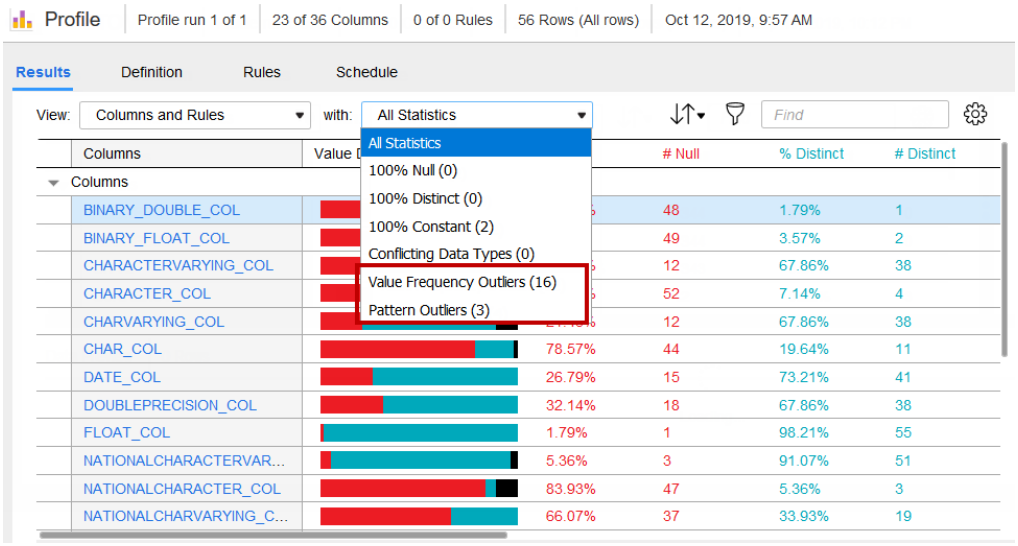
The **Detect Outliers** advanced option on the **Schedule** tab is enabled by default. During profile run, the profile identifies the columns with value frequency outliers and patterns outliers in the source object. The value frequency outliers are detected based on the values or frequencies in the column. The pattern outliers are detected based on the patterns in the column.

You can view the outliers in the source object in the following areas:

Profile results: summary view

In summary view, you can view the columns that contain outlier values. To view the columns with outliers, choose *Value Frequency Outliers* or *Pattern Outliers* filters in the results area.

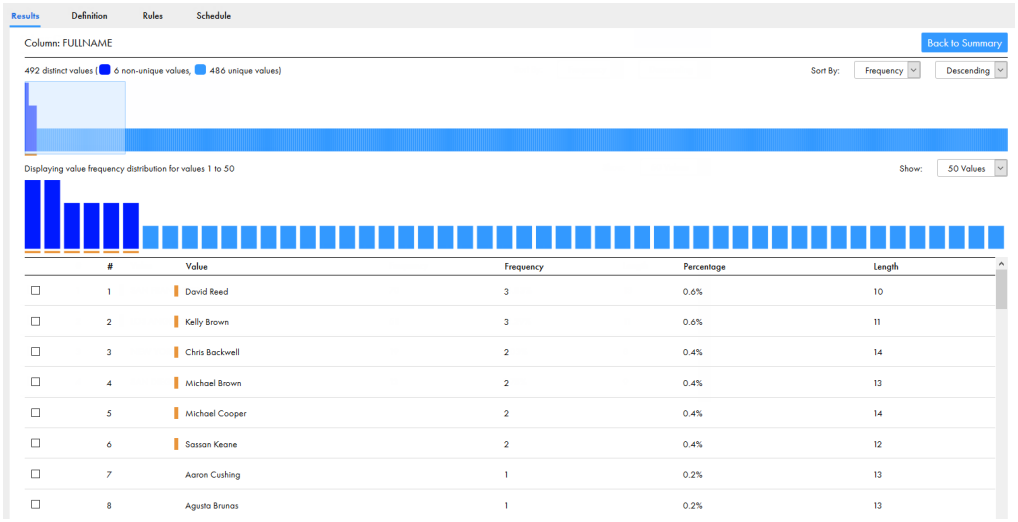
The following image shows an example of the *Value Frequency Outliers* and *Pattern Outliers* filters in the results area:



Profile results: detailed view

In the detailed view, you can view the outlier values in a column. The outlier values appear with an orange underline in the bar chart and a orange vertical bar in the value distribution table.

The following sample image shows the outlier values in the results area:



Queries

After you run a profile on a source object, you can create and run one or more queries on the source object.

You can create and run queries if you select the **Drilldown** option for the current profile run. You need the Query - Create privilege to create queries, and the Query - Submit privilege to run queries and view query results.

Create a query

On the **Results** tab, you can create one or more queries to retrieve the rows from a profiled source object that has a data quality problem. You can query based on field or column values, inferred patterns, data types, and rule outputs. For example, you can create and run a query that can retrieve source rows that are 'Invalid', where 'Invalid' is a business rule that you define in a rule specification, or if the postal code pattern is not 9(5).

You can add one or more query conditions to a query. The following table shows the attributes that you use to create a condition:

Attribute	Description
Columns	Choose a column. You can select columns in the source object and rule outputs in the current profile run. Columns and rule outputs might not appear in the list of columns if the data type of a column and rule output is not supported by Data Profiling.
Operator	Choose an operator to filter the results. You can select Equals, Not Equals, Less Than, Less Than or Equals, Greater Than, Greater Than or Equals, Between, In, Not In, Is Null, Is Not Null, Patterns, Data Types, Starts With, Ends With, or Contains operator for a condition. When you select the Patterns operator, Data Profiling shows the inferred patterns for the current profile run. When you select Data Types operator, Data Profiling shows the documented data type and inferred data types in the current profile run. Data Profiling does not show any inferred pattern if you select a column that is not included in the latest profile run. In this case, you can enter a pattern.
Values	Enter the values as necessary. When you choose the Patterns or Data Types operator, you can select one or more patterns or data types as values.

Run a query

You can run more than one query at a time. To run the queries, choose a flat file connection. Data Profiling runs the queries on the runtime environment that you chose for the flat file connection. When you use a flat file connection to create and run a profile on a flat file source, Data Profiling shows the flat file connections that use the same runtime environment that was used in the profile's flat file connection. You can create a dedicated flat file connection to run and save queries.

Data Profiling creates a job when you run a query. You can monitor the job progress on the **My Jobs** page. You can also monitor the job progress in Monitor and Operational Insights.

Note: The query runs on all the rows in the source object. If you chose a filter for the profile run or choose a filter and then create a query, Data Profiling filters the source object and then runs the query on the filtered results.

View query results

You can view the query results in the **Data Preview** area. When you run the query, Data Profiling generates a query results file named `query_<ProfileName>query<QueryName>.csv`. If the profile has associated rules, Data Profiling also generates a legend file named `query_<ProfileName>query<QueryName>.legend` which explains the column content in the query results file. Data Profiling saves the files in the directory that you specified in the flat file connection. Data Profiling. When you run a query multiple times, the query results are overwritten in the file.

Delete a query

When you delete a query, Data Profiling deletes the query from the profile. It does not delete the query results file and legend file related to the query. You can maintain, secure, and delete the files as required.

Example

You are a data analyst. You run a profile on the Order table, and you notice that the OrderID column has data types and patterns that are not valid. You want to generate a query to extract these specific results to analyze them. To accomplish this task, you complete the following steps:

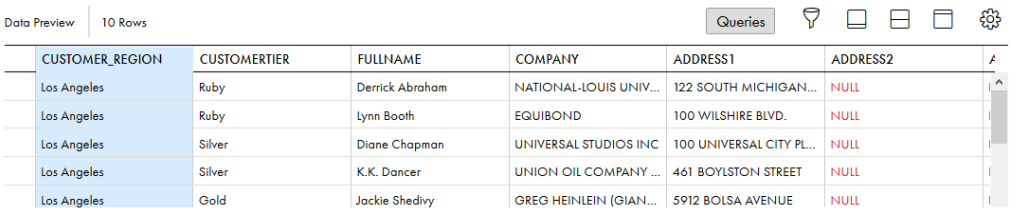
1. On the **Results** page, you create a query to meet one of the following conditions:
   a. You choose the Patterns operator for the OrderID column and then select the inferred patterns that are invalid.
   b. You choose the Data Types operator for the OrderID column and then select the inferred data types that are invalid.
2. You save and run the query.
   The complete query results appear in the **Data Preview** area.
3. Alternatively, to view the complete query results, you navigate to the query results file location to analyze the results.

Creating and running a query

You can create a query on the source object after you run a profile on it.

1. On the **Results** tab, click **Queries**.

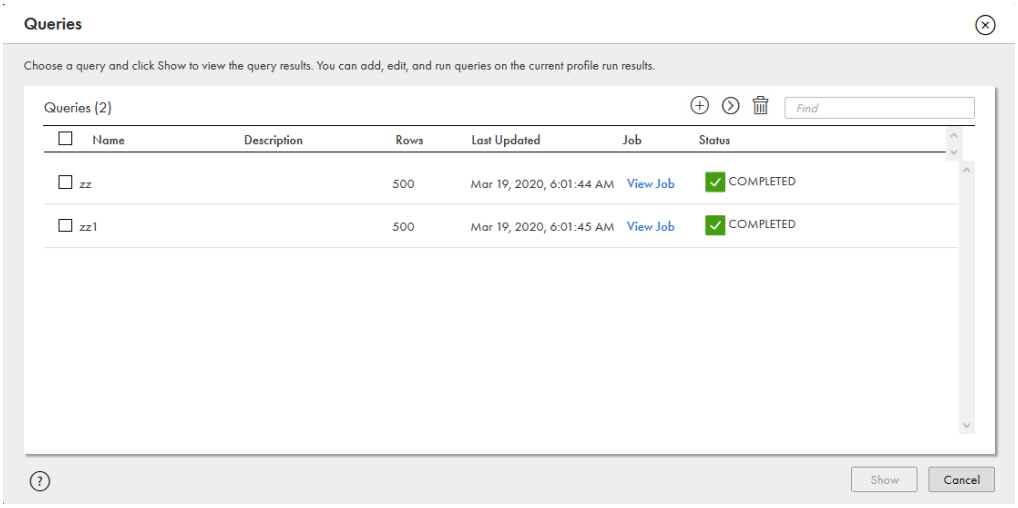
   The following image shows the **Queries** option on the **Results** tab:



CUSTOMER_REGION	CUSTOMERTIER	FULLNAME	COMPANY	ADDRESS1	ADDRESS2
Los Angeles	Ruby	Derrick Abraham	NATIONAL-LOUIS UNIV...	122 SOUTH MICHIGAN...	NULL
Los Angeles	Ruby	Lynn Booth	EQUIBOND	100 WILSHIRE BLVD.	NULL
Los Angeles	Silver	Diane Chapman	UNIVERSAL STUDIOS INC	100 UNIVERSAL CITY PL...	NULL
Los Angeles	Silver	K.K. Dancer	UNION OIL COMPANY ...	461 BOYLSTON STREET	NULL
Los Angeles	Gold	Jackie Shedivy	GREG HEINLEIN (GIAN...	5912 BOLSA AVENUE	NULL

2. In the **Queries** dialog box, click Add.

   The following image shows the **Queries** dialog box:



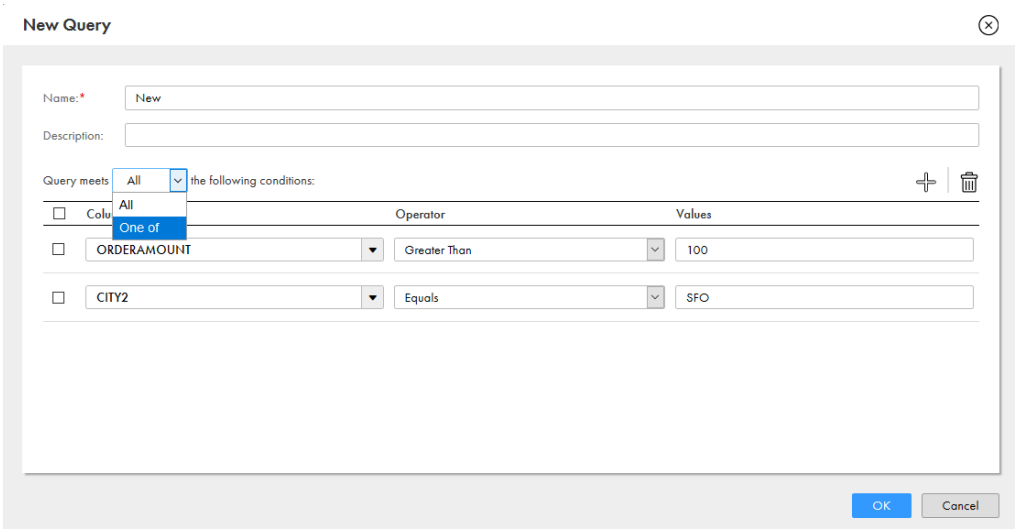
3. In the **New Query** dialog box, enter a name for the query.

   Optionally, you can add a description for the query.

4. Click Add to add a condition.
5. Choose a column, operator, and values as necessary.
6. Enter more conditions if required.
7. After you enter all the conditions for the query, choose one of the following options to generate query results:

   - All. Data Profiling retrieves the rows that meet all the conditions.
   - One of. Data Profiling retrieves those rows that meet at least one of the conditions.

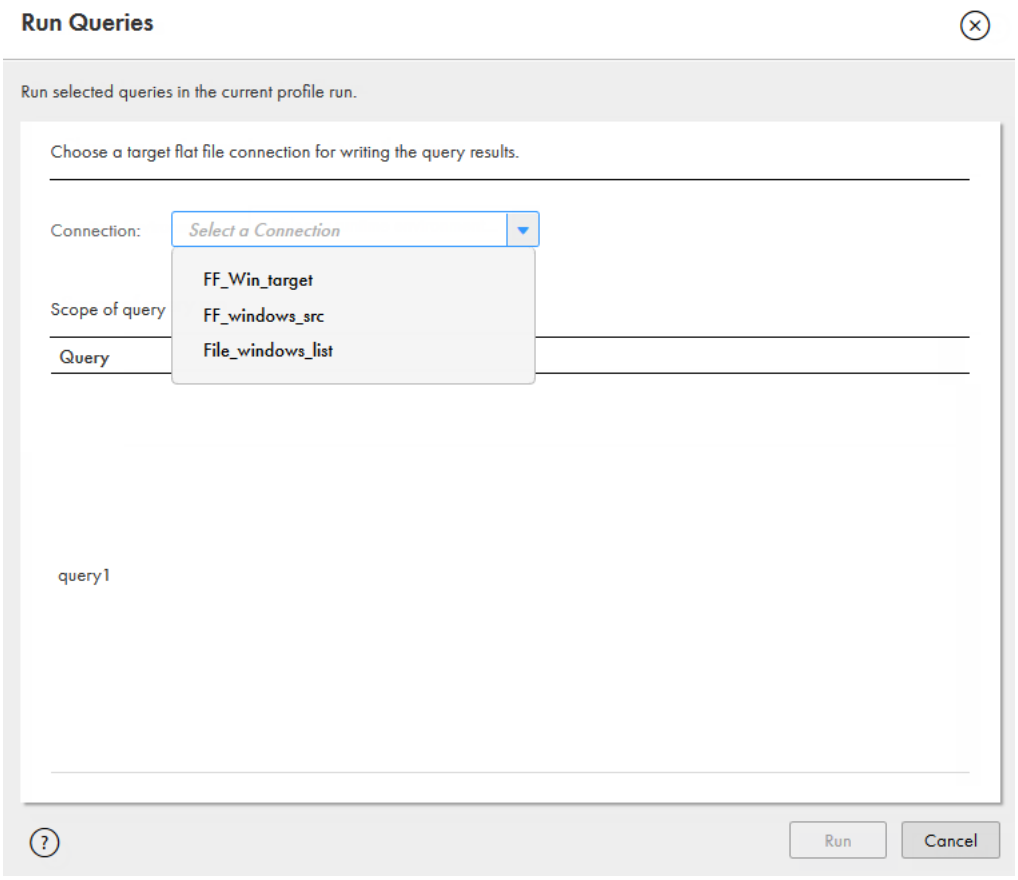
   The following image shows the **New Query** dialog box which contains the Add option to add conditions to the query, the Delete option to delete the conditions, and an option to choose all or one of the conditions for the query:



8. Click **OK**.
9. In the **Queries** dialog box, select one or more queries, and click Run.

10. In the **Run Queries** dialog box, choose a flat file connection. Data Profiling runs the query on the runtime environment associated with the flat file connection.

    The following image shows the **Run Queries** dialog box:



11. Click **Run**.
12. In the **Queries** dialog box, click **Show**.

    The **Data Preview** area shows the complete query results.
13. To view the query results as a .csv file, navigate to the directory that you used to create the flat file connection.

    Data Profiling generates a query results file named `query_<ProfileName>query<QueryName>.csv`. If the profile has associated rules, Data Profiling also generates a legend file named `query_<ProfileName>query<QueryName>.legend` which explains the column content in the query results file.

Choose a profile run

You can run a profile multiple times. The profile results for each run is saved in the Informatica Intelligent Cloud Services repository. You can choose to view the profile results for any profile run.

You can choose to view the profile results for the following profile runs:

- Latest profile run. You can view the latest profile run results after you run a profile. When you open a data profiling task that you have run, the latest profile run results appear.
- Historical profile run. You can view the profile results for one of the previous profile runs.

Example

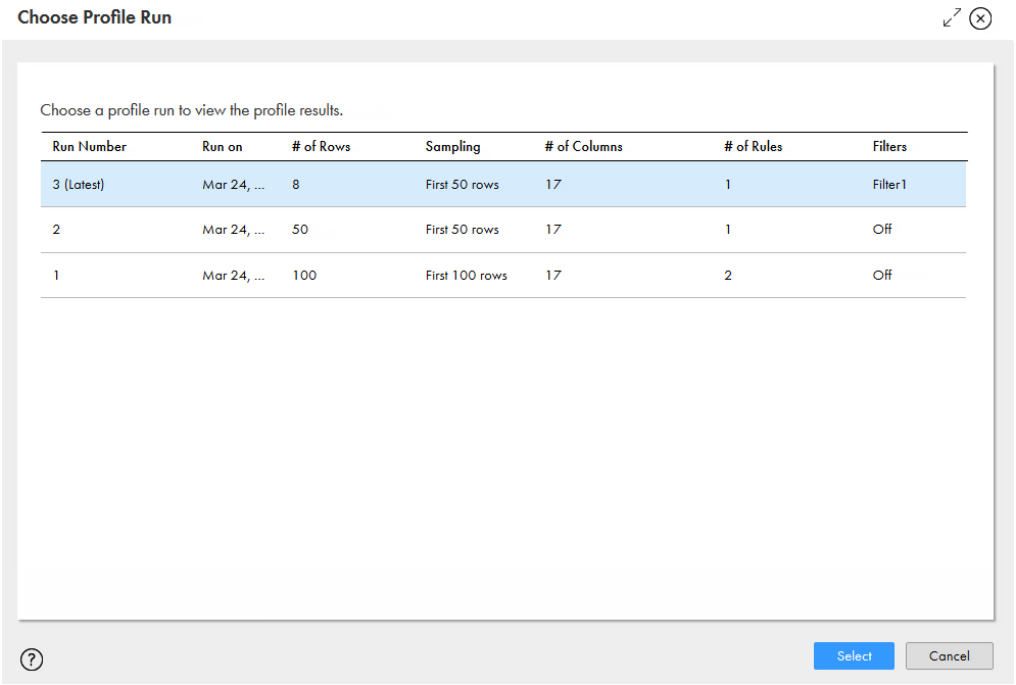
You are a data analyst. You create and run profiles on the Sales table every month. Based on a business need, you want to view the results in October 2018. To accomplish this task, you can open the profile run that you ran in October 2018.

Choosing a profile run

You can select the latest profile run or a historical profile run to view its profile results.

1. Open a profile.
2. Click **Actions** > **Choose Profile Run**.

   The following image shows the **Choose Profile Run** dialog box:



3. To choose the latest profile run, click the run with **(Latest)** appended to the run number, and click **Select**.
4. To choose a historical profile run, click any run other than the latest run, and click **Select**.

Compare profile runs

You can compare the results for two profile runs to analyze and compare the content and statistics. After you select the profile runs to compare, the comparison results appear on the **Compare Runs** tab.

The later profile run results are compared to the previous profile run results. If a column was added in the later run, the column name appears with the term **Added**. If a column was removed in the later run, the column name appears with the term **Removed**.

When you change the source object after multiple runs, Data Profiling retains the profile results for all the profile runs in the profiling warehouse. You can compare the profile results for the previous and current source object. The columns of the previous source object appears as **Removed** and the columns of the current source object appears **Added** on the **Compare Runs** tab.

Example

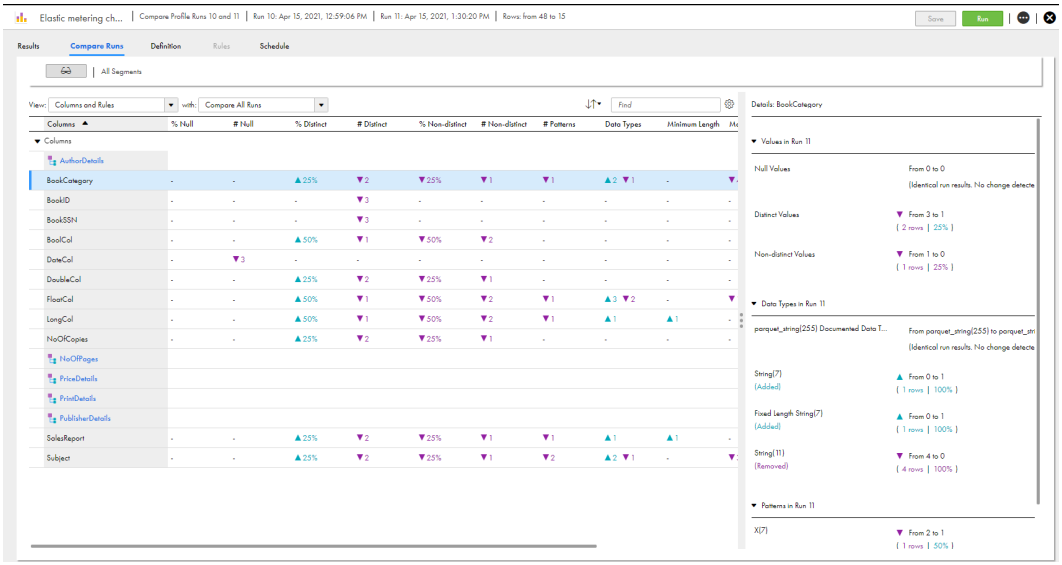
You are a data steward. You create a profile on the Customer table. You need to identify the customers who were added to or deleted from a subscription in a month.

To accomplish the task, perform the following tasks:

1. Run the profile on the Customer table on a monthly basis.
2. Compare the latest profile results with the previous one or as required.
3. Analyze the compare run results.

The **Compare Runs** tab displays a tree previewer to help you navigate to the profile runs of the nested columns for profiles that you create with Avro or Parquet source objects.

The following image displays a sample **Compare Runs** tab with a tree previewer:

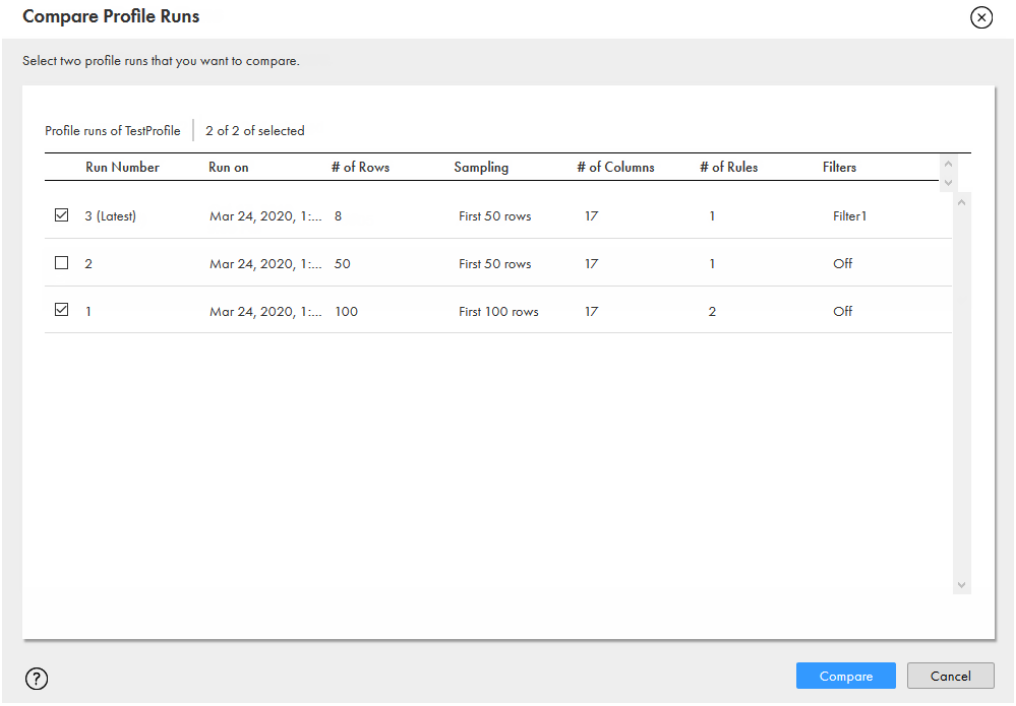


Comparing profile runs

You can select two profile runs to compare the profile results.

1. Open a profile and view the **Results** tab.
2. Click **Actions** > **Compare Profile Runs**.

The following sample image shows the **Compare Profile Runs** dialog box:

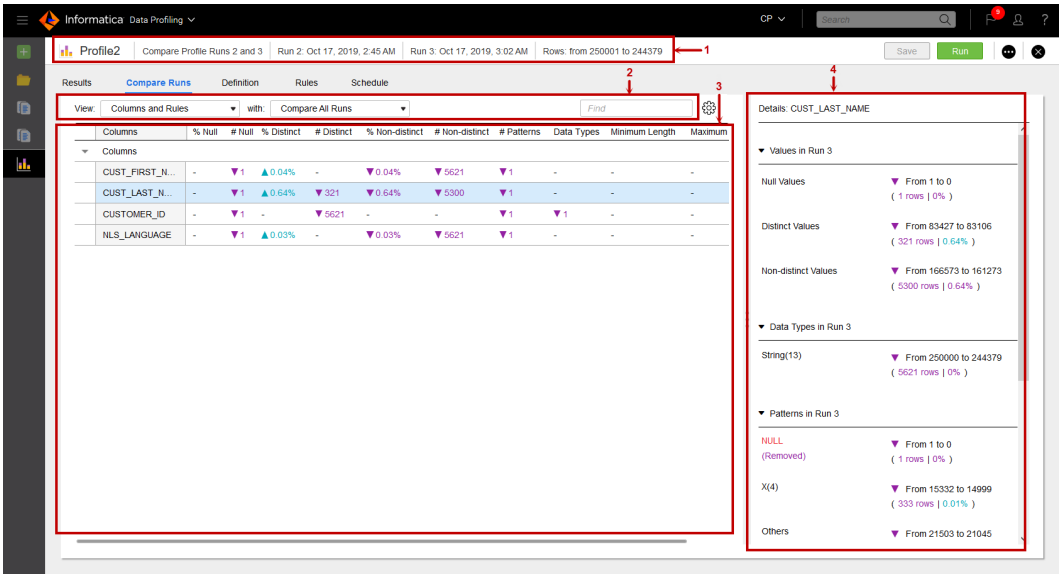


3. Choose two profile runs, and click **Compare**.

Compare run results

When you compare the results for two profile runs, the comparison results appear on the **Compare Runs** tab.

The following sample image shows the areas that you can view on the **Compare Runs** tab:



1. Header
2. Filter or find
3. Compare statistics
4. Details

Header

The header area shows the profile run details which include the profile run numbers, profile run timestamps, and number of rows in the earlier run as compared to the later run.

Filter or find

The following table explains the options that appear in the **Filter and find** area:

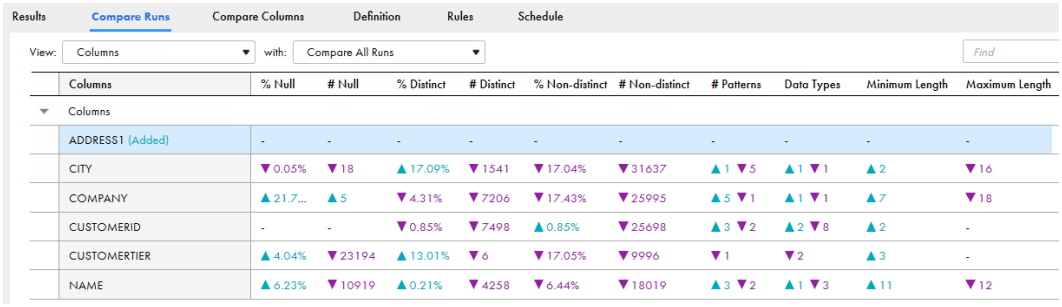
Option	Description
View	Shows the following options: - Columns and Rules. View the results for all the columns and rules in the profile run. - Columns. View the results for the columns in the profile run. - Rules. View the results for the rules in the profile run.
With	Shows the following options: - Compare All Runs. View the comparison results for both the runs. - Differences. View the differences in results in both the runs. - Matches. View the results that match in both the runs. - Added. Vew the results for columns that was added in the later run. - Removed. View the results for columns that was removed in the later run. Choose a filter in the **With** option after you choose a filter in the **View** option.
Find	Enter a keyword to view the relevant search results.
Menu	Choose Comfortable, Cozy, or Compact to adjust the row width in the profile results area.

Compare statistics

The compare statistics area shows the columns and rules in collapsible sections. The column statistics in both the runs are compared and displayed in the compare statistics area. An up arrow with a numeric count displays an increase in value for the statistic from the earlier run to later run. A down arrow with a numeric

count displays a decrease in value for a statistic. You can choose the statistics that you want to view in the area. To add or remove a statistic, right-click a statistic name and select or clear the statistic.

The following sample image shows the compare statistics area:



Columns	% Null	# Null	% Distinct	# Distinct	% Non-distinct	# Non-distinct	# Patterns	Data Types	Minimum Length	Maximum Length
Columns										
ADDRESS1 (Added)	-	-	-	-	-	-	-	-	-	-
CITY	▼0.05%	▼18	▲17.09%	▼1541	▼17.04%	▼31637	▲1 ▼5	▲1 ▼1	▲2	▼16
COMPANY	▲21.7...	▲5	▼4.31%	▼7206	▼17.43%	▼25995	▲5 ▼1	▲1 ▼1	▲7	▼18
CUSTOMERID	-	-	▼0.85%	▼7498	▲0.85%	▼25698	▲3 ▼2	▲2 ▼8	▲2	-
CUSTOMERTIER	▲4.04%	▼23194	▲13.01%	▼6	▼17.05%	▼9996	▼1	▼2	▲3	-
NAME	▲6.23%	▼10919	▲0.21%	▼4258	▼6.44%	▼18019	▲3 ▼2	▲1 ▼3	▲11	▼12

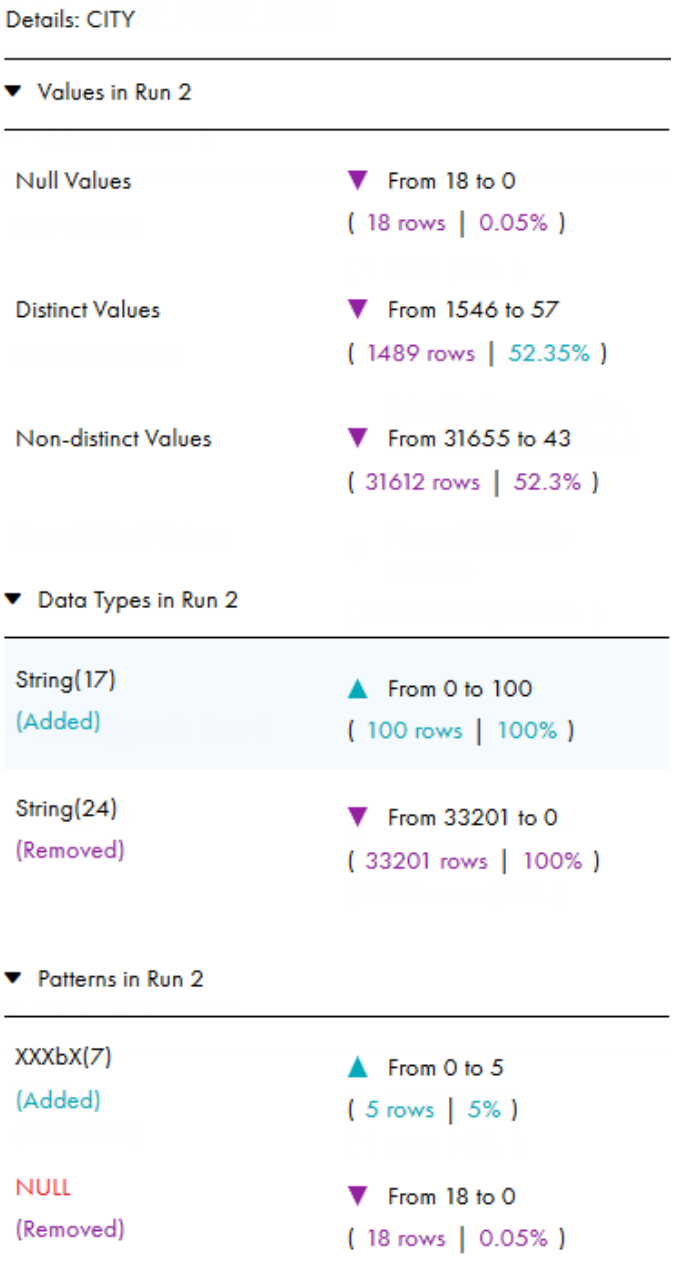
The compare statistics area shows column statistics, such as the value distribution, percentage and number of values, data types, patterns, and the minimum and maximum values.

When you click a column, the statistics for the column appear in the **Details** area for the later run.

Details

In the **Details** area, you can view the statistics and comparison results. The comparison results include the number of rows in both the runs, difference in row count and row percentage in the later run.

The following sample image shows the **Details** area:



In this area, you can view the following statistics in collapsible sections:

Values in *<later_run>*

Shows the comparison results for null values, distinct values, and non-distinct values.

Data Types in *<later_run>*

Shows the comparison results for inferred data types.

Patterns in *<later_run>*

Shows the comparison results for inferred patterns.

Compare columns in a profile

You can compare the column results for two or more columns in a profile run. You can compare the results for a maximum of 15 columns.

Example

You are a data quality user. You need to compare the date fields, such as *create date of agreement*, *contract start date*, *contract end date* and similar fields in the Contracts table to analyze the data.

To accomplish this task, you perform the following tasks:

1. Create and run a profile on the Contracts table.
2. Compare the required columns and view the results for further analysis.

The following image displays a sample **Compare Columns** dialog box for the nested hierarchical columns and nested columns:

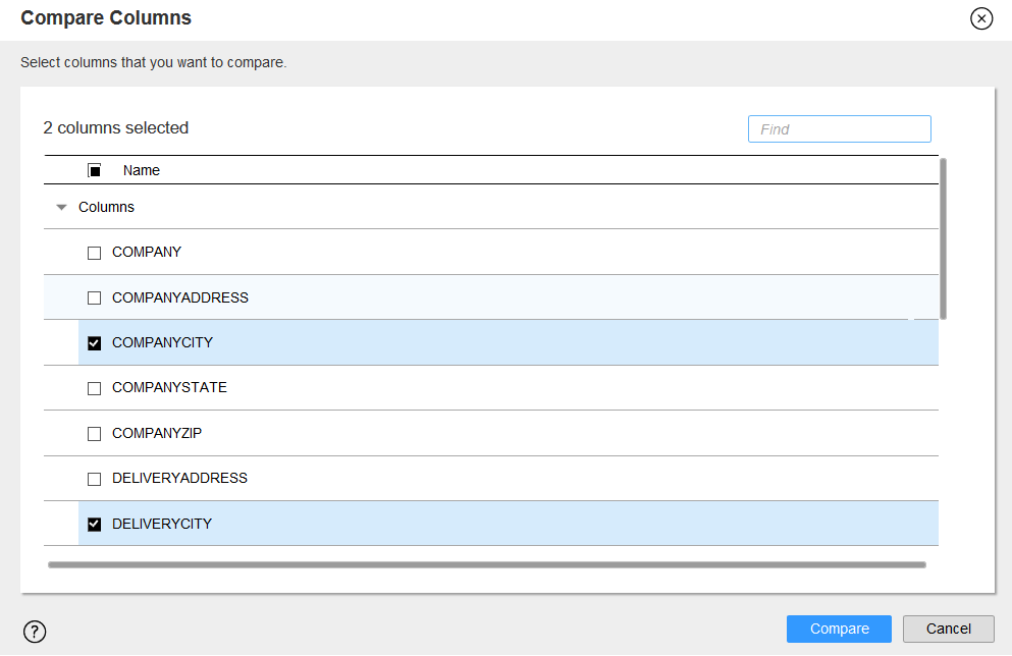
1. Nested hierarchical column
2. Nested column

Comparing multiple columns in a run

You can select two or more columns in a profile run to compare the column results.

1. Open a profile.
2. Click **Actions** > **Compare Columns**.

   The following sample image shows the **Compare Columns** dialog box:

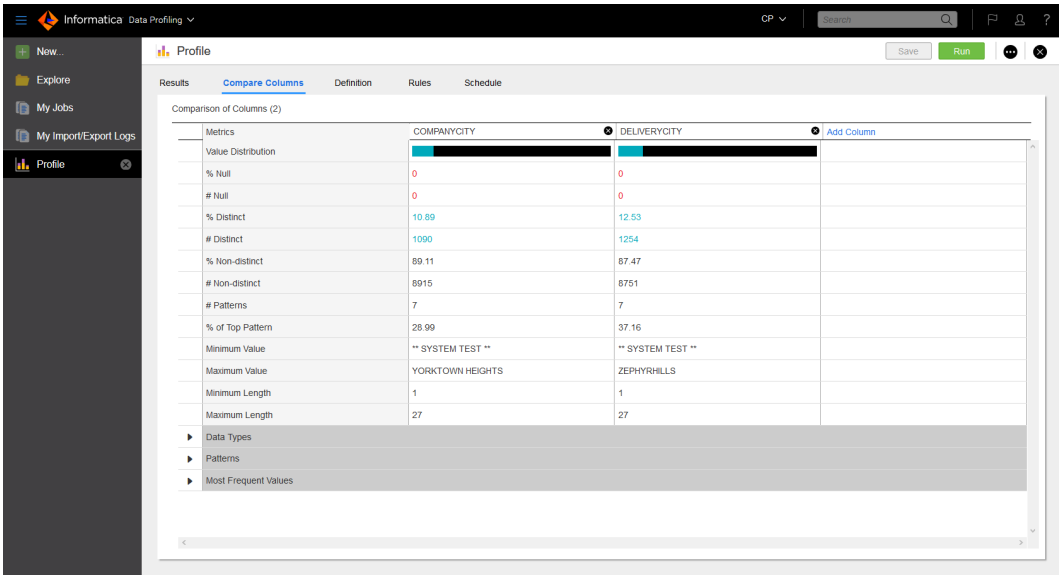


3. In the **Compare Columns** dialog box, you can perform one of the following actions to select the columns:
   - Select **Name** to select all the columns.
   - Choose two or more columns.
   - Enter a keyword in the **Find** field to search for the columns.
4. Click **Compare**.

Compare column results

When you compare the results for two or more columns in a profile run, the comparison results appear on the **Compare Columns** tab. Each column statistic appears in a row. You can add or delete columns on the **Compare Columns** tab.

The following sample image shows the **Compare Columns** tab:



The **Compare Columns** area shows the following results in collapsible sections:

Metrics

Shows column statistics, such as the value distribution, percentage and number of values, and the minimum and maximum values.

Data Types

Shows the percentages of documented data type and inferred data types in separate rows.

Patterns

Shows the percentages of inferred patterns in separate rows.

Most Frequent Values

Shows the percentages of all the values and their frequencies in a column in separate rows.

Sometimes, the **Most Frequent Values** might not display all the available values in a column when the number of most frequent values in a column is greater than the **Maximum Number of Value Frequency Pairs** value. To view all the available values, increase the **Maximum Number of Value Frequency Pairs** value as necessary.

Export profile results

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

You can export the profile results for any valid profile run. When you export the profile results for a profile run, Data Profiling saves the file name with the latest name of the profile. To export the profile results, verify that you have the enabled the **Export Data Profiling Results** feature for the user role in the Administrator.

Example

You are a data analyst and you have access to create and run profiles using the Data Profiling service. The sales team need the profile results for the Sales table to make some business decisions and they do not have access to Data Profiling.

To accomplish this task, you can create a profile on the Sales table, run the profile, and export the results to a Microsoft Excel file. You can share the file with the sales team.

Exporting profile results to a file

You can export the results for one or more columns or for all columns.

1. Open a profile to view the **Results** tab.
2. Click **Actions** > **Export Profile Results**.
3. In the **Export Profile Results to a File** dialog box, enter the following details:
   - File Name. You can retain the default file name, or enter a file name of your choice.
   - Choose **All Columns** to export the results for all the columns, or choose **Selected Columns** and select one or more columns to export the results for the chosen columns.
   - Choose one or more of the following options:
     - Summary. Exports all the profile results.
     - Value Frequency. Exports only the value frequencies.
     - Statistics. Exports only the statistics.
     - Patterns. Exports only the patterns.
     - Data Types. Exports only the data types, which include documented and inferred data types.

   By default, the **File Format** is set to Excel and the **Code Page** is set to **7-bit ASCII**.
4. Click **Export**.

   You can open or save the file to your local machine.

View exported profile results in the file

When you export the profile results, the profile results are exported to a Microsoft Excel file. The service saves the file in the ".xlsx" format.

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

Worksheet	Description
Column Profile	Summary of profile results appears in this worksheet where the results for columns and rules appear in collapsible panes. You can view the following results in the worksheet: - Profile name - Filter name - Sampling policy - Column name and its statistics appear in separate rows. - Rule and its statistics appear in separate rows.
Values	Contains values in a column with the value frequency, percentage, and maximum length. This worksheet appears when you choose the **Value Frequency** option in the **Export Profile Results to a File** dialog box.

Worksheet	Description
Statistics	Contains the following statistics for each column: - Maximum Length - Minimum Length - Bottom (5) - Top (5) This worksheet appears when you choose the **Statistics** option in the **Export Profile Results to a File** dialog box.
Patterns	Contains inferred patterns and their frequency and percentage for each column. This worksheet appears when you choose the **Patterns** option in the **Export Profile Results to a File** dialog box.
Data Types	Contains inferred data types and their frequency and percentage for each column. This worksheet appears when you choose the **Data Types** option in the **Export Profile Results to a File** dialog box.
Properties	Contains the following profile and profile run properties: - Profile name - Type - Description - Location - Link to profile - Source object - Row count - Filter name - Filter condition - Created by - Last Modified on - Date and time when the last profile was run

Export the value frequencies to a dictionary

You can export the value frequencies of a particular column to a dictionary from the detailed view. A dictionary is a reference data set that you can use to evaluate data in a mapping. You can use a dictionary to verify the accuracy and format of the value frequencies on a data source or an object in a mapping.

After you export the value frequencies to a dictionary, you can add the dictionary to an asset in Data Quality, and then use the asset as a rule in Data Profiling. For example, you can configure a rule specification or cleanse asset to read a dictionary.

You can export one column and the values at a time to a dictionary. If you want to export multiple columns values under different columns, you need to design the dictionary with the required columns, and then use the dictionary to export all the value frequencies from Data Profiling.

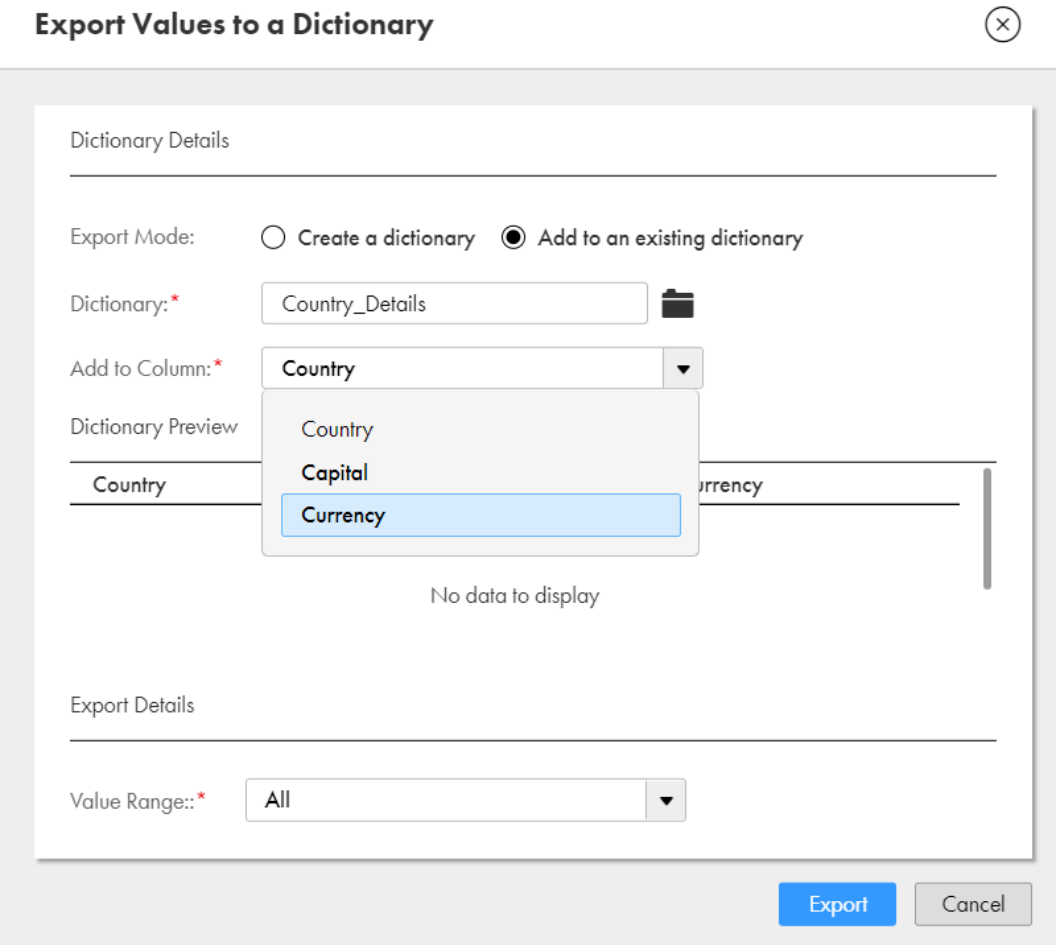
When you export value frequencies to the dictionary, Data Quality adds the values to the first empty column in the dictionary. When you export value frequencies to an existing dictionary, Data Quality adds the values to the blank rows of the column that you selected.

Example

You want to create a reference table of different countries, currencies, and capitals using a profile, and then use the dictionary in a rule specification or cleanse in Data Quality. To achieve this, you need to export the

value frequencies to a dictionary named **Country_Details**. The **Country_Details** dictionary includes columns such as `Country`, `Capital`, and `Currency`.

To export the value frequencies to a dictionary, you can create and run a profile on the source object that includes the required information. After the profile job completes, you can then export each value frequencies of a particular column to the columns in the **Country_Details** dictionary. The following image displays the **Export Values to a Dictionary** window:



Note: You can export the value frequencies of nested columns of the Avro and Parquet source objects to a dictionary from the detailed view.

Exporting column values to a dictionary

You can export one or more value frequencies of a column to a dictionary. To export the value frequencies of different columns, you need to design the dictionary with the required columns, and then export the values one at a time. Before you export the value frequencies to a dictionary, verify that you have enabled the dictionary permissions and privileges for the user role in Administrator.

1. From the summary view of the profile results section, select a column that you want to export to a dictionary.
2. Click the column name.

   The column appears in the detailed view.

3. Optionally, if you want to export specific value frequencies, select the values from the value distribution table. By default, Data Profiling exports all the value frequencies.
4. In the detailed view section, click the export values to a dictionary icon .

   The **Export Values to a Dictionary** window appears.
5. Specify the following dictionary details:
   - Export Mode. You can create a new dictionary, or you add to an existing dictionary. By default the **Create a dictionary** option is selected.
   - Name. Enter a name for the dictionary.
   - Description. Optionally, enter a description.
   - Location. Click **Browse** and then select a project or folder where you want to save the dictionary.
6. If you choose the **Add to an existing dictionary** option, specify the following details:
   - Dictionary. Click , and then select an existing dictionary from the **Select Dictionary** window.

     Note: Before you select an existing dictionary, verify that you have the required permissions for the specific dictionary.
   - Add to Column. Select the column to which you want to export the value frequencies in the dictionary.
7. Select the value range. You can choose to export **All**, **Selected**, **Unique**, or **Non-unique** value frequencies depending upon the values you selected in the value distribution table.
8. Click **Export**.

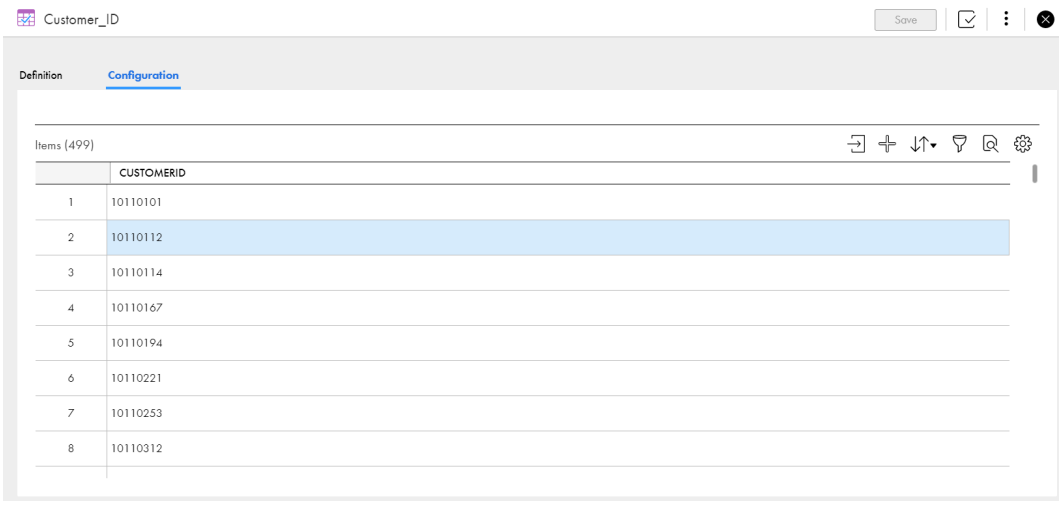
   Data Profiling exports the value frequencies to the dictionary in Data Quality. To view the exported values in the dictionary, click the notifications link that appears after you export.

View exported column values in a dictionary

When you export the value frequencies to a dictionary, the values appear on the **Configuration** view in Data Quality. You can define and update the content and structure of a dictionary on the **Configuration** view.

To open the dictionary on the **Configuration** view, open the folder where you saved the dictionary, and then click the dictionary name link.

The following image shows a sample dictionary on the **Configuration** view:



Profile Jobs

Data Profiling creates a job when you run a data profiling task. You can view the job statistics on the **My Jobs** page.

Click **Actions** > **Profile Jobs** on the **Results** page to view the job statistics for the data profiling task. For more information about the **My Jobs** page, see *Monitor*.

You can view the runtime environment and the Secure Agent for the following subtasks in Data Profiling, Monitor, and Operational Insights:

- Fetching the source row count
- s_profiling
- Drilldown
- Query

Note: The Runtime Environment field displays the name of the Secure Agent Group.

You can view the following details for the Secure Agent in the session log file for the profile mapping jobs:

- Task Name. The name of the profiling task.
- Agent Group Id. The ID of the Secure Agent Group.
- Agent Group Name. The name of the Secure Agent Group.
- Agent Id. The ID of the Secure Agent.
- Agent Name. The name of the Secure Agent.

Deleting profile runs for a profile

You can delete one or more profile runs for a profile. When you delete a profile run, Data Profiling deletes the profile results for the profile run from the profiling warehouse. You can delete a profile run to reclaim the storage space in the profiling warehouse.

1. Open a profile and view the **Results** tab.
2. Click **Actions** > **Delete Profile Runs**.
3. In the **Delete Profile Runs** dialog box, you can choose the following options to view the profile runs for the profile:
   - View. Choose All or a time period to view the profile runs.
   - Sort. Sort the profile runs based on their run number in ascending order.
4. Select the profile runs as necessary.

   You can view the amount of data that the selected profile runs occupy in the profiling warehouse.
5. Click **Delete**.

   Data Profiling permanently deletes the selected profile runs. You can delete a maximum of 50 profile runs for a profile at a time.
6. Click **Close**.

CHAPTER 4

Tuning data profiling task performance

You can tune a data profiling task by configuring the advanced options for a data profiling task in Data Profiling. You can also configure the number of concurrent tasks for a Secure Agent in Administrator.

To optimize the performance of data profiling tasks, Data Profiling creates subtasks for concurrent processing of profile jobs. The number of subtasks is based on the number of columns and rows in the data source and on the advanced options that you set for data profiling tasks.

By default, Data Profiling uses the following criteria to create subtasks:

Row-based

Creates one subtask for each column when the data source exceeds 100,000,000 rows. To modify the default value, configure the **Minimum Number of Rows for Split Process per Column** option. For example, the source object has 50 columns and 101,000,000 rows, Data Profiling creates 50 subtasks. If the rows in the source object exceed the default **Minimum Number of Rows for Split Process per Column** value, Data Profiling creates one subtask for each column in the source object.

Column-based

Creates one subtask for every 50 columns and rules when the data source contains 100,000,000 rows or lesser. To modify the default value, configure the **Maximum Number of Columns per Mapping** option. For example, the source object has 80 columns and 10,000,000 rows, Data Profiling creates 2 subtasks. If the columns in the source object exceed the default **Maximum Number of Columns per Mapping*** value, Data Profiling creates one subtask for every 50 columns and another subtask for the remaining columns.

Note: Data profiling prioritizes row-based criteria. To prioritize column-based criteria, set the **Minimum Number of Rows for Split Processing per Column** option to a value that is greater than the actual number of rows in the source.

You can configure the advanced options on the **Schedule** tab for each data profiling task. The following table lists the advanced options and recommendations for optimum performance:

Option	Recommendations
Maximum Number of Value Frequency Pairs	Default is 500. Decrease or increase this value based on the business need.
Maximum Number of Patterns	Default is 10. Decrease or increase this value based on the business need.
Pattern Threshold Percentage	Default is 5. Decrease or increase this value based on the business need.

Option	Recommendations
Infer Date and Time	By default, Data Profiling infers the date and time for a column of date or time data type. Clear this option if you do not want to infer the date and time for a column of date or time data type in the data source. **Note:** Data Profiling performance might be impacted because it consumes a lot of resources to infer date and time.
Detect Outliers	By default, outliers are detected in the profile results. Clear this option if you do not want to detect and view outliers in the data source.
Minimum Number of Rows for Split Process per Column	Default is 100,000,000. Increase or decrease this value based on the business need. Row-based criteria uses this option to optimize performance. For example, if you set the value to 100,000 and the number of rows in the source object is 100,500 and the columns is 30, Data Profiling creates 30 subtasks for each column in the source object.
Maximum Number of Columns per Mapping*	Default is 50. Increase or decrease this value based on the business need. Column-based criteria uses this option to optimize performance. For example, you set the value to 30 and **Minimum Number of Rows for Split Processing per Column** value to 100,000,000. If the source object contains 149 columns and 70,000 rows. Data Profiling creates a subtask for each 30 columns, which results in five subtasks. Four subtasks contain 30 columns each, and one subtask contains 29 columns.
Maximum Memory per Mapping	Default is 512 MB. Increase or decrease this value based on the business need.
Default buffer block size	Default is Auto. Enter a numeric value and append KB, MB, or GB to the value to increase or decrease the value based on the business need.
DTM buffer size	Default is Auto. Enter a numeric value and append KB, MB, or GB to the value to increase or decrease the value based on the business need. By default, a minimum of 12 MB is allocated to the buffer at run time. You might increase the DTM buffer size in the following circumstances: - When a task contains large amounts of character data, increase the DTM buffer size to 24 MB. - When a task contains *n* subtasks, increase the DTM buffer size to at least *n* times the value for the task with one subtask. - When a source contains a large binary object with a precision larger than the allocated DTM buffer size, increase the DTM buffer size so that the task does not fail.
Line Sequential Buffer Length	Default is 1024. Increase the value if the source flat file records are larger than 1024 bytes.
* The mapping is a type of subtask. Data Profiling creates and runs subtasks for a data profiling task to process the data concurrently.	

Configure Secure Agent concurrency

By default, the Secure Agent processes two concurrent tasks.

To configure the Secure Agent concurrency, perform the following steps:

1. In Administrator, open the **Runtime Environments** page.

2. Select the Secure Agent to view its details, and then click **Edit** on the Secure Agent page.
3. Specify the following details:
   - Service. Choose **Data_Integration_Server**.
   - Type. Choose **Tomcat**.
   - Name. Enter **maxDTMProcesses**.
   - Value. Enter the Secure Agent concurrency value.
4. Click **Save**.

Frequently Asked Questions

Sometimes, I see a low performance with the default Minimum Number of Rows for Split Process per Column and Maximum Number of Columns per Mapping values. How can I improve the data profiling task performance?

This issue might occur for large data sources where the number of rows is less than 100,000,000 and there are more than 50 columns. In this case, Data Profiling chooses column-based criteria and creates one subtask for every 50 columns. This consumes a lot of memory and processing power.

To resolve this issue, you can set the **Minimum Number of Rows for Split Process per Column** option to a higher value and the **Maximum Number of Columns per Mapping** option to a lower value.

For example, if a data source contains 10,000,000 rows and 100 columns, the data profiling task creates two subtasks with the default configuration. This consumes a lot of memory and results in a longer run time. In this case, you can retain the default value of **Minimum Number of Rows for Split Process per Column** and set the **Maximum Number of Columns per Mapping** option to 25. Data Profiling creates four subtasks which optimizes the performance and resource utilization. In addition, you can also increase the Secure Agent concurrency from the default 2 to (n), where n = Integer (0.8* number of cores) on the machine where Secure Agent runs.

Can I increase the Secure Agent concurrency to optimize the Data Profiling performance?

Yes, in addition to configuring the advanced options for data profiling tasks, you can configure the Secure Agent concurrency which impacts Data Profiling performance.

For example, assume that a data source contains 10,000,000 rows and 100 columns and the machine on which the Secure Agent runs has 4 cores.

To optimize the performance, perform the following steps:

1. In Administrator, configure **maxDTMProcesses** to a value **n**, where n = Integer (0.8 * number of cores) on the machine where Secure Agent runs. In this case, set **maxDTMProcesses** to 3.
2. In Data Profiling, create a profile for the data source.
3. On the **Schedule** page, set **Maximum Number of Columns per Mapping** to 15.
4. Save and run the profile.

In this case, Data Profiling generates 7 subtasks.

When I configure the Maximum Number of Columns per Mapping option to 20 for a data source with 100 columns, I see 8 subtasks for the profile on the My Jobs tab. Why do I see more subtasks than required?

When you create and run a profile, the following subtasks are generated and run:

- Fetching the source row count-<number_of_chosen_rows>. This task is generated only once for a profile run.

- Generating data profiling mappings. This subtask is generated only once for a profile run.
- s_profiling. The number of subtasks generated are based on the **Minimum Number of Rows for Split Process per Column** and **Maximum Number of Columns per Mapping** values. In this case, five subtasks are generated.
- Loading data from staging area to metric store. This task is generated only once for a profile run.

In this case, the total number of subtasks created for the task is eight subtasks.

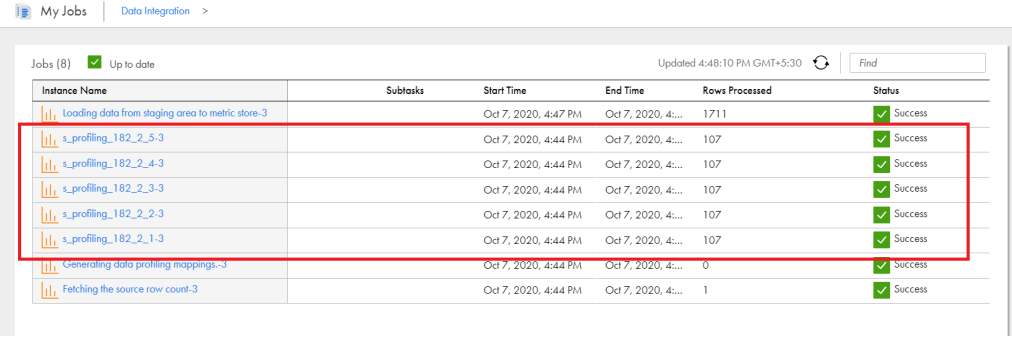
What are the connections that create multiple mapping subtasks for a profile job?

The following is a list of connections that create multiple mapping subtasks for a profile job:

- Oracle
- SQL Server
- Flat File
- Azure Synapse SQL (ODBC)
- Amazon Redshift v2
- Snowflake Cloud Data Warehouse V2

How can I confirm the number of mapping subtasks each profile job creates?

You can view the count of s_profiling jobs listed on the **My Jobs** tab. To view the s_profiling jobs, click the subtasks link on the **My Jobs** tab. For example, the following image displays the sample **My Jobs** tab with s_profiling jobs:



The profile job does not fail even if the Stop on Errors field value is set to a value that is less than equal to the sum of rows rejected.

This issue occurs when you apply multiple rules to a profile job. Data Profiling considers the maximum number of rows rejected by an individual rule. When a profile job includes multiple rules, Data Profiling stops the profile job on errors when the total number of reject rows cross the configuration of one of the rules, and not the sum of other rules.

I cannot view auto-assigned rules for the profiles after I change the source object or connection of my profile.

Data Profiling does not assign rules automatically when you edit and make changes to a profile.

The automatic rule association feature does not work for profiles that I created in R36.

Data Profiling associates rules automatically only if you create new profiles in 2021.07.M release (R37).

Profile job completes with the following warning message: `Record length [] is longer than line sequential buffer length [] for <file location>. Record will be rejected.` **How do I resolve this issue?**

To resolve this issue, increase the value of Line Sequential Buffer Length parameter till the error resolves. The parameter is located under the **Advanced Options** section on the **Schedule** tab of the profile.

Note:

- The line sequential buffer length is used for delimited and fixed width flat files.
- An easier way to calculate the line sequential buffer length is to increase the value by multiples of 2 of the record length value that appears in the warning message.

Why is the length of the value frequency higher than the actual value frequency length on the Results page?

This issue might occur if the column data contains special characters such as an apostrophe ('). To resolve this issue, you can remove the special characters from the column data and rerun the profile.

The number of successful and unsuccessful rows in the preview does not show correct number of rows in the scorecard dashboard when you run a profile with random sampling for a specific number of rows?

The profile runs honour the sampling options whereas the drill down feature does not honour the sampling options. Due to this, the number of successful and unsuccessful rows shown in the preview does not match with the total number of rows shown in the **Rule Occurrences** table of the scorecard dashboard.

Data Profiling creates automatic rule specifications when insight validations have failed for columns that are either deselected from the profile definition or are deleted from the source. Can I delete these rules?

Yes, you can manually delete these rules from the **Explorer** tab.

Can you run native profiles using the serverless runtime environment?

Yes, you can run native profiles using the serverless runtime environment, but you cannot run Spark profiles.

Can you drilldown or query a profile that runs on a serverless runtime environment?

No, you cannot drilldown or query a profile that runs on a serverless runtime environment.

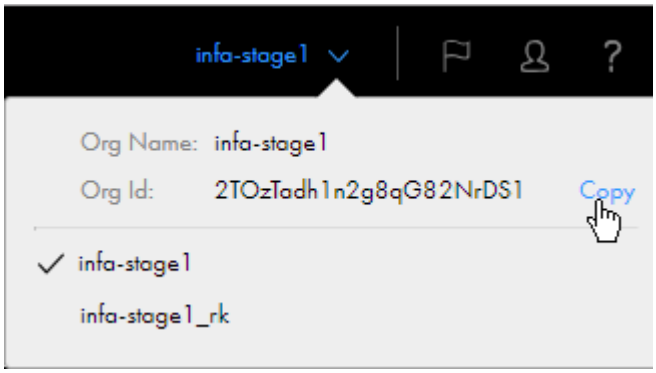
CHAPTER 5

Troubleshooting

Use the following sections to troubleshoot errors in Data Profiling.

Note: To get support for Data Profiling, you might need to give your organization ID to Informatica Global Customer Support. You can find your organization ID through the **Organization** menu in the upper right corner.

The following image shows the **Organization** menu:



To copy the organization ID, click the **Copy** option that appears when you hover the cursor to the right of the **Org ID** field.

You can also find your organization ID on the **Organization** page in Administrator.

Troubleshooting a data profiling task

Create and run profiles

The Review Insights option in the menu appears disabled if I open the Results tab before the Insights job is completed. How can I resolve this issue?

To resolve this issue, refresh the page. The **Review Insights** option appears enabled in the menu if insights are generated for the profile.

During profile creation, if I choose an ODBC connection and search for a source object, the search results do not show the source object even when it exists. How can I resolve this issue?

Searches are case-sensitive for ODBC. To search for the source object, enter the source object name using the correct case.

A profile run fails and the following error message appears in the session log: "The executor with id xx exited with exit code 137(SIGKILL, possible container OOM).". How do I resolve this issue?

To resolve this issue, perform the following steps:

1. Open the custom.properties file available in the following location on the machine where the Secure Agent runs: `/root/infaagent/apps/At_Scale_Server/<version>/spark/`
2. Add the following property: **spark.executor.memoryOverhead = 2048MB**
3. Save the custom.properties file.
4. In Data Profiling, run the profile.

A profile run fails and the following error message appears in the session log: "The node was low on resource: ephemeral-storage. Container spark-kubernetes-driver was using xxx, which exceeds its request of xx.". How do I resolve this issue?

To resolve this issue, increase the minimum and maximum EBS volume sizes to attach to a worker node for temporary storage during data processing.

To increase the minimum and maximum EBS volume sizes, perform the following steps in Administrator:

1. In Administrator, open the **Advanced Clusters** page.
2. Select the Advanced Configuration for which you want to change the EBS volume size.
3. Click **Edit**.
4. In the **EBS Volume Size** field of the **Platform Configuration** area, increase the values in the **Min GB** and the **Max GB** fields to **200**.
   By default, the minimum and maximum volume sizes are 100 GB.
5. Click **Save**.
6. Restart the Secure Agent.
7. In Data Profiling, run the profile.

Unable to save a profile using Databricks with an ODBC connection when I create tables with the same name under two different databases. How can I resolve this issue?

This issue occurs when you do not specify the schema name in the connection. To resolve this issue, specify the schema name in the connection to point to the correct database.

If columns contain a large number of rows, the profile job fails for a Microsoft Azure Synapse SQL connection and the following error message appears: "error "[FATAL] Exception: com.microsoft.sqlserver.jdbc.SQLServerException: Error 0x27 - Could not allocate tempdb space while transferring data from one distribution to another.". How can I resolve this issue?

To resolve this issue, increase the Data Warehouse Units (DWU) of the Microsoft Azure Synapse SQL instance.

A profile run fails with the error "Profile job failed with error java.lang.RuntimeException: Output Port Primary does not exist in specified rule". How do I resolve this issue?

This error appears when the following conditions are true:

1. In Data Profiling, you create a profile, add a rule R1, save, and run the profile.
2. In Data Quality, you modify the rule input or output name for rule specification R1 and save it.
3. In Data Profiling, you run the profile.

To resolve this issue, you can remove rule R1 from the profile and save the profile. Add the rule R1 again to the profile, save, and run the profile.

A profile run fails with the error "*ERROR: nsort_release_recs() returns -10 ". How do I resolve this issue?**

To resolve this issue, increase the disk space storage of the hard drive where Secure Agent is installed.

When you run a profile on an Amazon S3 source object, the profile run fails with an error "Cloud DQ Profiling failure ERROR: Unexpected condition at [file:[..\..\..\common\reposit\trepcnx.cpp|file://[.....commonrepositтrepcnx.cpp/]] line: [293]". How do I resolve this issue?

To resolve this issue, ensure that you have the valid license for the Amazon S3 connection in Administrator.

When I run a profile on a Salesforce source object, the profile run fails and an 'Out of Memory' error appears. How do I resolve this issue?

To resolve this issue, you can increase the Java heap size **-Xmx** value to twice its current value.

To increase the Java heap size, perform the following steps in Administrator:

1. In Administrator, open the **Runtime Environments** page.
2. Select the Secure Agent for which you want to change the Java heap size.
3. Click **Edit**.
4. In the **System Configuration Details** area, select the **Data Integration Server** service and choose the **DTM** type.
5. Click **Edit** in the row for the **INFA_MEMORY** property.
6. Increase the value of **Xmx** to twice its current value.
   For example, if the current value of **INFA_MEMORY** property is **-Xms256m -Xmx512m**, change it to **-Xms256m -Xmx1024m**.
7. Click **Save**.
8. Restart the Secure Agent.
9. In Data Profiling, run the profile.

The profile run fails with an "Out Of Memory" error. How do I resolve this issue?

To resolve this issue, you can increase the Java heap size **-Xmx** value to twice its current value.

To increase the Java heap size, perform the following steps in Administrator:

1. In Administrator, open the **Runtime Environments** page.
2. Select the Secure Agent for which you want to change the Java heap size.
3. Click **Edit**.
4. In the **System Configuration Details** area, select the **Data Integration Server** service and choose the **DTM** type.
5. Click **Edit** in the row for the **INFA_MEMORY** property.
6. Increase the value of **Xmx** to twice its current value.
   For example, if the current value of **INFA_MEMORY** property is **-Xms256m -Xmx512m**, change it to **-Xms256m -Xmx1024m**.
7. Click **Save**.
8. Restart the Secure Agent.
9. In Data Profiling, run the profile.

When I run a profile on a Google Big Query source object, the profile run fails and a 'GC overhead limit exceeded' error appears. How do I resolve this issue?

To resolve this issue, you can increase the Java heap size in the JVM options for type DTM. To increase the Java heap size, perform the following steps in Administrator:

1. In Administrator, open the **Runtime Environments** page.

2. Select the Secure Agent for which you want to change the Java heap size.
3. Click **Edit**.
4. In the **System Configuration Details** area, select the **Data Integration Server** service and choose the **DTM** type.
5. Click **Edit** in the row for the **INFA_MEMORY** property.
6. Set the available JVMOption fields to a minimum (**-Xms1024m**) and maximum (**-Xmx4096m**) Java heap size. For example, set JVMOption3 to **-Xms1024m** and JVMOption4 to **-Xmx4096m**.
7. Click **Save**.
8. Restart the Secure Agent.
9. In Data Profiling, run the profile.

When I run a profile on a Snowflake Cloud Data Warehouse V2 source object, the profile job runs with a warning or it fails.

To resolve this issue, you must increase the Java heap size in the JVM options. To increase the Java heap size, perform the following steps in Administrator:

1. In Administrator, open the **Runtime Environments** page.
2. Select the Secure Agent for which you want to change the Java heap size.
3. Click **Edit**.
4. In the **System Configuration Details** area, select the **Data Integration Server** service and choose the **DTM** type.
5. Set the available JVM Option fields to a maximum Java heap size value.
   - If the profile job runs with a warning due to large volumes of data in the source object, set the available JVM Option fields to a maximum Java heap size as per your requirements. For example, JVM Option fields to a maximum (**-Xmx2048m**).
   - If the profile job fails, set the available JVM Option fields to a maximum (**-Xmx2048m**) Java heap size.

   For more information, see the following KB article:
   https://knowledge.informatica.com/s/article/336913?language=en_US
6. Click **Save**.
7. Wait till the **Data Integration Server** service restarts.
8. In Data Profiling, run the profile.

Data Profiling rejects the rows that have conversion errors when you run a profile. How do I resolve this issue?

This issue occurs when you edit the column metadata to change the data type of a column that still includes rows with a few values of the previous data type. For example, if the data source includes a column with string and integer values and you change the column data type to integer.

To resolve this issue, you can configure the **Stop on Errors** option and enter the number of rows that include incorrect data type, and then run the profile.

How do I run a profile with Avro and Parquet file format types?

To run a profile with Avro or Parquet file format type, you need to configure the Amazon S3 V2 or Azure Data Lake Store connection with the respective secure agents for the Amazon or Azure cluster.

When I run a profile with Avro or Parquet file format types, the profile run fails and the following error message appears: `Columns tab error:[The file or partition directory[] is not valid. The parser`

`encountered the following error while parsing the content:[Only one hadoop distribution can be supported]. Select a valid [Parquet] file or partition directory.]`**. How do I resolve this issue?**

The Cloudera 6.1 package that contains the Informatica Hadoop distribution script and the Informatica Hadoop distribution property files is part of the Secure Agent installation. When you run the Hadoop distribution script, you need to specify the distribution that you want to use. To resolve the above issue, you need to perform the following steps:

1. Go to the following Secure Agent installation directory where the Informatica Hadoop distribution script is located: `<Secure Agent installation directory>/downloads/package-Cloudera_6_1/package/Scripts`
2. Copy the `Scripts` folder outside the Secure Agent installation directory.
3. From the terminal, run the `./infadistro.sh` command from the `Scripts` folder and proceed with the prompts.
4. In Administrator, open the **Runtime Environments** page.
5. Select the Secure Agent for which you want to configure the DTM property and click **Edit**.
6. Add the following DTM properties in the **Custom Configuration** section:
   - Service: Data Integration Service
   - Type: DTM
   - Name: INFA_HADOOP_DISTRO_NAME
   - Value: <distribution_version>
     The value of the distribution version can be given as CDH_6.1.
7. Restart the Secure Agent to reflect the changes.
8. In Data Profiling, run the profile.

For more information on the above steps, see https://docs.informatica.com/integration-cloud/cloud-data-integration-connectors/current-version/hive-connector/introduction-to-hive-connector/configure-hive-connector-to-download-the-distribution-specific-h.html.

When a profile run fails with the following error: "Either the Amazon S3 bucket *<xyz>* does not exist or the user does not have permission to access the bucket", the Amazon S3 test connection also fails for the same runtime environment:

To resolve the issue, perform the steps listed in the following KB article:https://kb.informatica.com/solution/23/Pages/76/626289.aspx

When you use the Snowflake ODBC connection to create a profile, the source columns do not load in Data Profiling and the following error message appears:

```
{"@type":"error","code":"APP_13400","description":"com.informatica.saas.rest.client.spring.RestTemplateExtended$SpringIOException: HTTP POST request failed due to IO error: Read timed out; nested exception is org.springframework.web.client.ResourceAccessException: I/O error on POST request for \" [https://iics-qa-release-pod2-r36-r1-cdi102.infacloudops.net:47813/rest/MetadataRead/getTableMetadata\|https://iics-qa-release-pod2-r36-r1-cdi102.infacloudops.net:47813/rest/MetadataRead/getTableMetadata/] ": Read timed out; nested exception is java.net.SocketTimeoutException: Read timed out","statusCode":403}
```

To resolve this issue, you must add the CLIENT_METADATA_REQUEST_USE_CONNECTION_CTX=true property in the odbc.ini file located at the $ODBCHOME directory.

Snowflake profiles with large volume like 10 million rows or more fails with the following error: "The target server failed to respond". How do I resolve this issue?

To resolve this issue, perform the following steps:

1. Create a file with name: logging.properties in the secure agent server at any location, and add the following line in the file, and save the file.
   `java.util.logging.ConsoleHandler.level=WARNING`
2. In Administrator, open the **Runtime Environments** page.
3. Select the Secure Agent and click **Edit**.
4. In the **System Configuration Details** area, select the **Data Integration Server** service and choose the **DTM** type.
5. Click **Edit Agent Configuration** and add the following value for an empty JVMOption property: `-Xmx6144m`

   Note: If the Java heap size **-Xmx** value is already configured, edit the value of the existing JVMOption property to `-Xmx6144m`.
6. Click **Edit Agent Configuration** and add the following value for an empty JVMOption property: `-Dnet.snowflake.jdbc.loggerImpl=net.snowflake.client.log.JDK14Logger`
7. Click **Edit Agent Configuration** and add the following value for an empty JVMOption property: `-Djava.util.logging.config.file=<absolute path along with file name created in step 1>`
8. Click **Save**.
9. Restart the Secure Agent.
10. In Data Profiling, run the profile.

A profile run fails for an Snowflake or Azure Synapse SQL connection and the following error message appears: `'com.informatica.profiling.jpa.model.ProfileableDataSourceColumn; nested exception is org.hibernate.HibernateException: More than one row with the given identifier was found'`**. How do I resolve this issue?**

This issue occurs if the following conditions are true:

- You do not specify a schema during the ODBC connection configuration for an Snowflake or Azure Synapse SQL subtype.
- There are multiple tables with the same name and columns exist within the different schemas of the connection.

To resolve this issue, you must add a schema in the connection properties to eliminate the duplicate source objects.

A few profile runs fail with the following service exceptions: `com.informatica.cloud.errorutil.MicroServiceException: Error parsing results file. com.opencsv.exceptions.CsvMalformedLineException: Unterminated quoted field at end of CSV line` **and** `java.sql.SQLException: Parameter index out of range (7 > number of parameters, which is 6).`. **How do I resolve the issues?**

To resolve the issues, you must set the following flag in the **Custom Configuration** section of the Secure Agent: ADD_ESCAPE_CHAR_TO_TARGET=true.

The following image displays the sample configuration details:

Custom Configuration

Service	Type	Name	Value	Sensitive
Data_Integration_Server64.0	TOMCAT_CFG	maxDTMProcesses	15	
Data_Integration_Server64.0	TOMCAT_JRE	ADD_ESCAPE_CHAR_TO_TARGET	true	

When I run a profile on a JSON source object, the profile run fails and the following error message appears: `<WorkflowExecutorThread40> SEVERE: The Integration Service failed to execute the mapping. java.lang.RuntimeException: java.lang.RuntimeException: [SPARK_1003] Spark task [InfaSpark0] failed with the following error: [Container [spark-kubernetes-driver] failed with reason [Error] and message [ehaus.janino.CodeContext.flowAnalysis(CodeContext.java:600) ++ at org.codehaus.janino.CodeContext.flowAnalysis(CodeContext.java:600)` **How do I resolve this issue?**

To resolve the issue, perform the following steps:

1. Stop the Secure Agent and the cluster that is associated with the Secure Agent.
2. Go to the following Secure Agent custom.properties file directory: `<AgentHome>/apps/At_Scale_Server/<latestversion>/spark`
3. Enter the following values:
   - `spark.driver.extraJavaOptions=-Djava.security.egd=file:/dev/./urandom`
   - `-XX:MaxMetaspaceSize=256M -XX:+UseG1GC -XX:MaxGCPauseMillis=500 -Xss75m`
   - `spark.executor.extraJavaOptions=-Djava.security.egd=file:/dev/./urandom`
   - `-XX:MaxMetaspaceSize=256M -XX:+UseG1GC -XX:MaxGCPauseMillis=500 -Xss75m`
   - `spark.driver.memory=10G`
   - `spark.executor.memory=12G`
4. Start the Secure Agent.
5. Re-run the profile.

When you run a profile that includes a mapplet with a Java transformation, the profile fails and the following error message appears: `400 : "{"code":"0","description":"Compilation failed for Java Tx: Java: 500 : \"{\"error\":{\"code\":\"APP_60001\",\"message\":\"Exception occurred during compilation: {\\ \"code\\\":\\\"TUNNEL_NOT_FOUND\\\",\\\"message\\\":\\\"No tunnels discovered for...` **How do I resolve this issue?**

Before you create a Mapplet with a Java transformation, perform the following steps:

1. In Administrator, navigate to the **Runtime Environments** page and select **Enable or Disable Services, Connectors** from the **Actions** menu of a Secure Agent or a Secure Agent group.
2. In the **Enable/Disable Components in Agent Group** window, select **Data Integration - Elastic**.
3. Click **Save**.

If the issue persists, perform the following steps:

1. In Data Integration, open the mapplet that contains the Java transformation.
2. Select the Java transformation in the **Design** workspace.
3. Compile and save the mapplet.

A profile run fails if the tenant initialization intermittently fails for a few orgs and the following error occurs: – `java.lang.RuntimeException: java.security.InvalidKeyException: Invalid AES key length: 56 bytes.` **How do I resolve this issue?**

You can re-run the profile if the first profile run in the org fails with the runtime exception error message.

Data types and patterns

For which data source does the Data Preview area show True or False for Boolean data type?

Data Profiling shows True and False for Salesforce columns that have the Boolean data type.

Does Data Profiling support all the data types in a Google BigQuery source object?

Data Profiling supports most of the data types in a Google BigQuery source object. The following table lists the known issues for Google BigQuery data types in Data Profiling:

Data types	Known issues
String	- When the column precision exceeds 255, Data Profiling truncates the column precision to 255 before the profile run. - Incorrect frequency of null values appear in the **Details** > **Data Types** section. - When you drill down on null values, blank values also appear in the **Data Preview** area.
Numeric	When the column precision exceeds 28, the profile run fails and the following error appears: [ERROR] Data Conversion Failed.
Time, Datetime, or Timestamp	- Milliseconds do not appear in the profile results. - Profile results contain duplicate values which results in incorrect frequency of values.
Geography	Profile run fails and the following error appears: [SDK_APP_COM_20000]
Float	When you drill down or create queries, an error appears if the column contains +inf, -inf, or NaN values.

Why do I see a pattern mismatch for INTERVALYEARTOMONTH and INTERVALDAYTOSECOND data types?

This issue occurs because Data Profiling reads the INTERVALYEARTOMONTH and INTERVALDAYTOSECOND data types as strings during pattern detection.

Binary float data types appear with extra decimal places. Do I need to do anything to round this off to two decimal places?

This is a known and accepted behavior for the binary float data type in Data Profiling. No action is required.

Profile results

If the drilldown results contain more than or equal to 100 rows, the Data Preview area does not display all the rows and the following error message appears in the session log: "Transformation Evaluation Error [<<Expression Fatal Error>> [ABORT]: DrillDown limit reached... i:ABORT(u:'DrillDown limit reached')]]". How do I resolve this issue?

If the drilldown results contain more than or equal to 100 rows, Data Profiling stops processing the job further and displays the top 100 results in the **Data Preview** area. To resolve this issue and to view the drilldown results of all the rows, you can use the **Queries** option in the **Data Preview** area.

Incorrect profile results appear for data sources that contain UTF-8 characters. How do I resolve this issue?

If the data source contains UTF-8 characters, you can set the `OdbcDataDirectNonWapi` parameter to **0** in Administrator. In Data Profiling, create and run the profile on the source object.

To configure the property in Administrator, open the **Runtime Environment** page, perform the following steps:

1. In Administrator, open the **Runtime Environments** page.
2. Select the Secure Agent for which you want to set this property.
3. Click **Edit**.
4. In the **System Configuration Details** area, select the **Data Integration Server** service and choose **DTM** type.
5. Click **Edit** in the row for the `OdbcDataDirectNonWapi` property and set the property to 0.
6. Click **Save**.

Why do I sometimes see no drilldown results for numeric columns?

This issue can occur when the data type is Integer and the column precision is greater than 28. Data Profiling does not display drilldown results for Integer data types with column precision greater than 28.

Why do I, sometimes, see incorrect column statistics for numeric and decimal columns that include average, sum, standard deviation, and most frequent values?

This issue can occur when the column precision for numeric columns or decimal columns is greater than 28. Data Profiling does not support column precision greater than 28 for numeric columns and decimal columns.

After I upgrade to Spring 2020 July, I do not see the existing query results. Why?

This issue occurs because the previous query results location `$PMCacheDir\profiling\query` is no longer valid. To view the query results, run the query again after you select a flat file connection. Data Profiling saves the query results to a file in the directory that you specified for the flat file connection.

After I upgrade to Fall 2020 October, I can still view the drill down results of Spring 2020 July in the Secure Agent Location. How do I clear the drill down results?

To clear the drill down results, open the Secure Agent installation directory `<Agent_installation_dir>/apps/Data_Integration_Server/data/temp/profiling/drilldown`, and then delete the Spring 2020 July drill down results manually.

I see incorrect profile results for columns that include escape characters. How do i resolve this issue?

To resolve this issue, you must set the following flag in the **Custom Configuration** section of the Secure Agent: ADD_ESCAPE_CHAR_TO_TARGET=true.

The following image displays the sample configuration details:

Custom Configuration

Service	Type	Name	Value	Sensitive
Data_Integration_Server64.0	TOMCAT_CFG	maxDTMProcesses	15	☐
Data_Integration_Server64.0	TOMCAT_JRE	ADD_ESCAPE_CHAR_TO_TARGET	true	☐

Rules

Why does profile run take a long time to complete when it contains a Verifier asset as a rule?

This issue occurs when the following conditions are true:

- You add the Verifier asset as a rule to the profile and run the profile.
- The Secure Agent is configured for a full country license.

- The reference data directory in the Secure Agent does not contain address reference data.

When you add the Verifier asset as a rule and run the profile, the Secure Agent downloads the address reference data for the first time which might impact the profile run time. The address reference data is the authoritative data for the postal addresses in the specified country.

Miscellaneous

How do I change the cache directory name in Administrator?

Perform the following steps to edit the cache directory name in Administrator:

1. In Administrator, open the **Runtime Environments** page.
2. Select the Secure Agent for which you want to change the cache directory name.
3. Click **Edit**.
4. In the **System Configuration Details** area, select the **Data Integration Server** service and choose **DTM** type.
5. Click **Edit** in the row for the **$PMCacheDir** property.
6. Remove the whitespaces in the property.
   For example, if the property contains `C:\Informatica Cloud Secure Agent\temp\cache`, change it to `C:\InformaticaCloudSecureAgent\temp\cache`.
7. Click **Save**.
8. In Data Profiling, run the profile.

Why does profile import fail if a profile with the same name exists in the folder?

This issue occurs because Data Profiling does not support overwriting of assets during import operation. To resolve this issue, rename the existing profile in the folder and then import the profile.

Why do I see the "ERROR: Document Artifact with Id \u003d jaAqeGnQc6phwrbCWkBW8D not found" error when I delete a profile run? How do I resolve it?

This issue occurs when a profile has an invalid frs ID association. To resolve this issue, you can re-import the profile asset if you have the export file. Or, you can move or copy the imported profile asset to a different folder or project.

Can I change the connection type, source object, and formatting options of a profile job?

Yes, you can edit the connection type, source object, and formatting options of the profile job in the following scenarios:

- You can change the connection type with the same connection type.
- You cannot change the source object to use a source object of a different connection.

I'm unable to choose a runtime environment for a profile of a flat file connection. Why?

Data Profiling does not support change in the runtime environment for a profile of a flat file connection. The profile runs on the default runtime environment configured for the flat file connection in Administrator.

How do I configure or override the runtime environment for Avro and Parquet file format types?

You must select a runtime environment that is associated with the advanced configuration.

Where do I find more information about advanced clusters and Informatica encryption for an Amazon S3 V2 connector on an advanced cluster?

- For more information about advanced clusters, see the Administrator help.

- For more information about Informatica encryption for an Amazon S3 V2 connector on an advanced cluster, see the Configuring Informatica Encryption for Elastic Mappings in Amazon S3 V2 Connector How-to-Library article.

INDEX