Data Profiling Guide

Informatica® PowerCenter®

(Version 8.6.1)

Informatica PowerCenter Data Profiling Guide

Version 8.6.1
December 2008

The information in this product or documentation is subject to change without notice. If you find any problems in this product or documentation, please report them to us in writing.

DISCLAIMER: Informatica Corporation provides this documentation "as is" without warranty of any kind, either express or implied, including, but not limited to, the implied warranties of non-infringement, merchantability, or use for a particular purpose. Informatica Corporation does not warrant that this software or documentation is error free. The information provided in this software or documentation may include technical inaccuracies or typographical errors. The information in this software and documentation is subject to change at any time without notice.

Part Number: PC-DPG-86100-0001

Table of Contents

Preface

The Informatica *PowerCenter Data Profiling Guide* provides information about building data profiles, running profile sessions, and viewing profile results. It is written for the database administrators and developers who are responsible for building PowerCenter mappings and running PowerCenter workflows.

This book assumes you have knowledge of relational database concepts, database engines, and PowerCenter. You should also be familiar with the interface requirements for other supporting applications.

Informatica Resources

Informatica Customer Portal

As an Informatica customer, you can access the Informatica Customer Portal site at http://my.informatica.com. The site contains product information, user group information, newsletters, access to the Informatica customer support case management system (ATLAS), the Informatica How-To Library, the Informatica Knowledge Base, Informatica Documentation Center, and access to the Informatica user community.

Informatica Documentation

The Informatica Documentation team takes every effort to create accurate, usable documentation. If you have questions, comments, or ideas about this documentation, contact the Informatica Documentation team through email at infa_documentation@informatica.com. We will use your feedback to improve our documentation. Let us know if we can contact you regarding your comments.

Informatica Web Site

You can access the Informatica corporate web site at http://www.informatica.com. The site contains information about Informatica, its background, upcoming events, and sales offices. You will also find product and partner information. The services area of the site includes important information about technical support, training and education, and implementation services.

Informatica How-To Library

As an Informatica customer, you can access the Informatica How-To Library at http://my.informatica.com. The How-To Library is a collection of resources to help you learn more about Informatica products and features. It includes articles and interactive demonstrations that provide solutions to common problems, compare features and behaviors, and guide you through performing specific real-world tasks.

Informatica Knowledge Base

As an Informatica customer, you can access the Informatica Knowledge Base at http://my.informatica.com. Use the Knowledge Base to search for documented solutions to known technical issues about Informatica products. You can also find answers to frequently asked questions, technical white papers, and technical tips.

Informatica Global Customer Support

There are many ways to access Informatica Global Customer Support. You can contact a Customer Support Center through telephone, email, or the WebSupport Service.

Use the following email addresses to contact Informatica Global Customer Support:

- support@informatica.com for technical inquiries
- support_admin@informatica.com for general customer service requests

WebSupport requires a user name and password. You can request a user name and password at http://my.informatica.com.

Use the following telephone numbers to contact Informatica Global Customer Support:

North America / South America	Europe / Middle East / Africa	Asia / Australia
Informatica Corporation Headquarters 100 Cardinal Way Redwood City, California 94063 United States	**Informatica Software Ltd.** 6 Waltham Park Waltham Road, White Waltham Maidenhead, Berkshire SL6 3TN United Kingdom	**Informatica Business Solutions Pvt. Ltd.** Diamond District Tower B, 3rd Floor 150 Airport Road Bangalore 560 008 India
Toll Free +1 877 463 2435	**Toll Free** 00 800 4632 4357	**Toll Free** Australia: 1 800 151 830 Singapore: 001 800 4632 4357
Standard Rate Brazil: +55 11 3523 7761 Mexico: +52 55 1168 9763 United States: +1 650 385 5800	**Standard Rate** Belgium: +32 15 281 702 France: +33 1 41 38 92 26 Germany: +49 1805 702 702 Netherlands: +31 306 022 797 Spain and Portugal: +34 93 480 3760 United Kingdom: +44 1628 511 445	**Standard Rate** India: +91 80 4112 5738

CHAPTER 1

Data Profiling Overview

This chapter includes the following topics:

Understanding Data Profiling

Data profiling is a technique used to analyze the content, quality, and structure of source data. Use PowerCenter Data Profiling to detect patterns and exceptions of source data during mapping development and during production. Use data profiling to make the following types of analyses:

- **Make initial assessments.** You can make initial assessments about data patterns and exceptions data during mapping development. As a result, you can design mappings and workflows on actual data, rather than make theoretical assumptions about sources.
- **Validate business rules.** You can validate documented business rules about the source data. For example, if you have a business rule requiring columns in a source table to contain U.S. ZIP codes, you can profile the source data to verify that the rows in this table contain the proper values.
- **Verify assumptions.** You can verify that the initial assumptions you made about source data during project development are still valid. For example, you may want to view statistics about how many rows satisfied a business rule and how many did not.
- **Verify report validity.** You can use data profiling to verify the validity of the Business Intelligence (BI) reports.

Data Profiling Components

To understand data profiling, you need to be familiar with the following components:

- **PowerCenter Client.** Use the PowerCenter Client to create and manage data profiles.
- **PowerCenter Data Profile.** Metadata that you generate in the PowerCenter Client that defines what types of statistics you want to collect for a source. It is comprised of a source definition, a profile mapping, and a profile session.
- **Data Profiling warehouse.** The Data Profiling warehouse stores results from profile sessions and reports that you run to view the results.
- **Data Profiling reports.** View data and metadata in Data Profiling reports.

PowerCenter Client

Use the following PowerCenter Client tools to create and manage data profiles:

- **Designer.** Create data profiles from the Source Analyzer or the Mapplet Designer. When you create a data profile, the Designer generates a profile mapping based on the profile functions. The PowerCenter repository stores the profile mappings and metadata. If the repository is versioned, profile mappings are versioned in the same way other PowerCenter mappings are versioned.
- **Profile Manager.** A tool in the PowerCenter Designer that you use to manage data profiles. You can edit and regenerate profiles, run profile sessions, and view profile results.

PowerCenter Data Profile

A data profile contains the source definitions, the functions and function parameters, and the profile session run parameters. To create a data profile, you run the Profile Wizard from the PowerCenter Designer. When you create a data profile, you create the following repository objects:

- **Profile.** A profile is a repository object that represents all the metadata configured in the wizard. You create the profile based on a mapplet or source definition and a set of functions.
- **Profile mapping.** When you create a data profile, the Profile Wizard generates a profile mapping. Select functions in the wizard that to help determine the content, structure, and quality of the profile source. You can use pre-defined or custom functions. The Profile Wizard creates transformations and adds targets based on the functions that you supply. You can view the profile mapping in the Mapping Designer.
- **Profile session.** After the Profile Wizard generates a profile mapping, you provide basic session information such as Integration Service name and connection information to the source and the Data Profiling warehouse. The Profiling Wizard creates a profile session and a profile workflow. You can choose to run the profile session when the wizard completes, or you can run it later. When you run a profile session, the Integration Service writes profile results to the Data Profiling warehouse.

While profiles are not versioned, the profile mappings and profile sessions are versioned objects.

Data Profiling Warehouse

The Data Profiling warehouse is a set of tables that stores the results from profile sessions. It also contains reports that you run to view the profile session results. You can create a Data Profiling warehouse on any relational database that PowerCenter supports as a source or target database. Create a Data Profiling warehouse for each PowerCenter repository you want to store data profiles in.

Data Profiling Reports

You can view the results of each function configured in the data profile. Based on the type of metadata you want to view, you can view reports from the following tools:

- **Profile Manager.** PowerCenter Data Profiling reports provide information about the latest session run. View them from the Profile Manager.
- **Data Analyzer.** Data Analyzer Data Profiling reports provide composite, metadata, and summary reports. View them from the Data Profiling dashboard in Data Analyzer. You can also customize the reports in Data Analyzer.

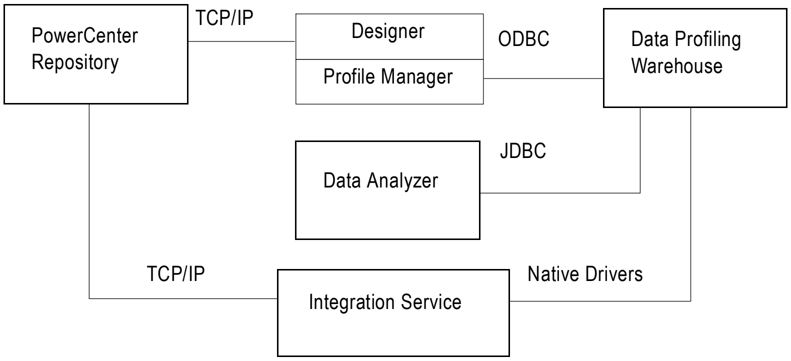
Data Profiling Connectivity

PowerCenter Data Profiling uses the following types of connectivity:

- **TCP/IP.** The PowerCenter Client and the Integration Service use native protocol to communicate with the Repository Service.
- **Native.** The Integration Service uses native database connectivity to connect to the Data Profiling warehouse when it loads target data from the profiling sessions.

- **ODBC.** The PowerCenter Client uses ODBC to connect to the Data Profiling warehouse when you run data profiling reports from the Profile Manager.
- **JDBC.** Data Analyzer uses JDBC to connect to the Data Profiling warehouse when you run data profiling reports.

The following figure shows the connectivity between the PowerCenter Data Profiling components:

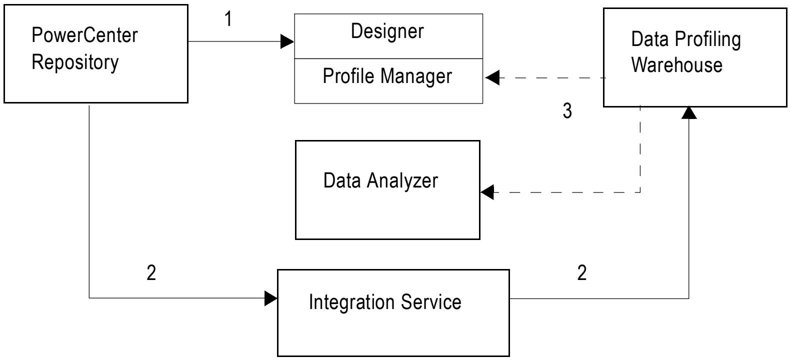


For more information about connectivity within PowerCenter, see the *PowerCenter Configuration Guide.*

Data Profiling Process

After you create a data profile, you can run the profile session and view the results in a report.

The following figure shows the Data Profiling process:



The following steps describe the data profiling process:

1. **Create a data profile.** Use the Profile Wizard in the Designer to create a data profile based on a source definition and a set of functions. The Profile Wizard generates a mapping and a session based on criteria that you provide.
2. **Run the profile session.** You can choose to run the profile session when you finish the Profile Wizard, or you can run it from the Profile Manager. The Integration Service runs the session and loads the profile results to the Data Profiling warehouse.
3. **View the reports.** View the Data Profiling report associated with the profile session. Based on the type of profile report, you can view reports from the Profile Manager or from Data Analyzer.

Steps for Profiling Source Data

After you create the Data Profiling warehouse, you create data profiles in PowerCenter. A data profile contains functions that perform calculations on the source data. When you create a data profile, the Designer generates a profile mapping and a profile session.

You can run profile sessions against the mapping to gather information about source data. The Data Profiling warehouse stores the results of profile sessions. After you run profile sessions, you can view reports that display the session results.

Complete the following tasks to profile a source, mapplet, or groups in a source or mapplet:

1. Create a data profile.
2. Run a profile session.
3. View profile reports.

The Designer provides a Profile Manager and Profile Wizard to complete these tasks.

Step 1. Create a Data Profile

To profile source data, you create a data profile based on a source or mapplet in the repository. Data profiles contain functions that perform calculations on the source data. For example, you can use a function to validate business rules in a data profile. You can apply profile functions to a column within a source, to a single source, or to multiple sources.

You can create the following types of data profiles:

- **Auto profile.** Contains a predefined set of functions for profiling source data. Use an auto profile during mapping development to learn more about source data.
- **Custom profile.** A data profile you define with the functions you need to profile source data. Use a custom profile during mapping development to validate documented business rules about the source data. You can also use a custom profile to monitor data quality or validate the results of BI reports.

You use the Designer to create a data profile. When you create a profile, the Designer generates a mapping and a session based on the profile information.

You can configure a data profile to write verbose data to the Data Profiling warehouse during a profile session. Verbose data provides more details about the data that results from a profile function. For example, for a function that validates business rules, verbose data may include the invalid rows in the source. For a function that determines the number of distinct values, verbose data can include a list of distinct values.

After you create a data profile, you can view profile details from the Profile Manager. You can also edit and delete the data profile.

Step 2. Run the Profile Session

After you create a data profile, you can run the profile session. The Integration Service writes the profile session results to the Data Profiling warehouse.

You can run profile sessions from the following places:

- **Profile Manager.** You can create and run temporary and persistent profile sessions from the Profile Manager. A temporary session runs on demand and is not stored in the repository. A persistent session can run on demand and is stored in the repository.
- **Workflow Manager.** If you create a persistent profile session when you create the data profile, you can edit and run the profile workflow from the Workflow Manager.

Step 3. View Data Profiling Reports

When you run a profile session, the Integration Service loads the session results to the Data Profiling warehouse. You can view the session results using PowerCenter Data Profiling reports.

Using the Profile Manager

The Profile Manager is a tool in the Designer that helps you manage data profiles. Use the Profile Manager to set default data profile options, work with data profiles in the repository, run profile sessions, view profile results, and view sources and mapplets with at least one profile defined for them. When you launch the Profile Manager, you can access profile information for the open folders in the repository.

There are two views in the Profile Manager:

- **Profile View.** The Profile View tab displays the data profiles in the open folders in the repository.
- **Source View.** The Source View tab displays the source definitions in the open folders in the repository for which you have defined data profiles.

Note: If the repository folder is read-only, you can view and run data profiles in the Profile View. You can also view Data Profiling reports. You cannot edit or delete data profiles.

From the Profile View and the Source View, you can complete the following tasks to manage, run, and view data profiles:

- Create a custom profile.
- View data profile details.
- Edit a data profile.
- Delete a data profile.
- Run a session.
- Regenerate a profile mapping.
- Check in profile mappings.
- Configure default data profile options.
- Configure domains for profile functions.
- Purge the Data Profiling warehouse.
- Display the status of interactive sessions.
- Display PowerCenter Data Profiling reports.

The Profile Manager launches immediately after you create a data profile. You can manually launch the Profile Manager from the following Designer tools:

- **Source Analyzer.** Click Sources > Profiling > Launch Profile Manager.
- **Mapplet Designer.** Click Mapplets > Profiling > Launch Profile Manager.
- **Repository Navigator.** Open a folder and select a source definition. Right-click on the source definition and select Launch Profile Manager.

Tip: If you do not want the Profile Manager to launch immediately after you create a data profile, you can change the default data profile options in the Profile Manager.

Profile View

The Profile View tab displays all of the data profiles in the open folder in the repository. Use the Profile View to determine the data profiles that exist for a particular repository folder.

Source View

The Source View displays the source definitions with data profiles in the open folder in the repository. A folder must be open before you can launch Profile Manager. Use the Source View to determine if a source definition already has data profiles defined. The Source View shows if the data profile is an auto profile or custom profile.

You can also use the Source View when you want to work with a data profile but are more familiar with the source name than the data profile name. For example, you want to run a profile session, and you know the source definition name but not the data profile name.

When you select the Source View tab in the Profile Manager, the Profile Navigator displays data profiles as nodes under the source definition for which you defined the data profile.

If you change or delete a data profile or a source or mapplet with a data profile, you can click View > Refresh to refresh the Source View.

CHAPTER 2

Managing Data Profiles

This chapter includes the following topics:

Overview

You can create, edit, and delete data profiles. Data profiles contain a set of functions that you apply to a specified set of source data. The functions return metadata about the profile sources that comprise the Data Profiling reports.

You can create the following types of data profiles:

- **Auto profile.** Contains predefined functions to profile source data. Use an auto profile during mapping or mapplet development to learn more about source data.
- **Custom profile.** Contains functions you create to profile source data. Use a custom profile during mapping or mapplet development to validate documented business rules about the source data. You can also use a custom profile to monitor data quality.

After you create a data profile, you can edit and delete the data profile.

Before you use Data Profiling, you can configure the default data profile options for the PowerCenter Client machine. Each data profile you create in the Designer uses these default options.

Profiling Sources and Mapplet Output Data

You can profile sources and output data from connected ports in mapplet output groups. You create or import source definitions in the Designer. You can profile the following types of sources:

- Relational database sources
- Flat file sources
- XML sources
- VSAM sources
- Application sources, such as SAP, PeopleSoft, and WebSphere MQ
- Mapplet output

When you profile a multi-group source, such as an XML source, you can select the groups in the source you want to profile, or you can profile the entire source.

When you profile mapplet output data, the Profile Wizard creates a data profile based on the data output from the connected ports in the Output transformation.

You can profile output data from mapplets that meet the following conditions:

- The mapplet contains a source definition for input.
- The mapplet contains no Input transformation.
- The mapplet contains no Transaction Control transformations.
- The ports in the Output transformation are connected.

Note: You can profile sources with ASCII and non-ASCII port names.

Profiling Multiple Sources

If you want to profile multiple sources, you can create a mapplet that combines multiple sources and create a data profile based on the mapplet output data. For example, you might use several Joiner transformations within the mapplet to join source data from multiple sources and profile the output data from this mapplet.

Profiling SAP R/3 Sources

When you create a data profile for SAP R/3 sources, you must generate an ABAP program for the profile mapping before you can run a session for the data profile.

Profile Functions

You can add multiple profile functions to a data profile. Profile functions are calculations you perform on the source data that return information about various characteristics of the source data.

When you add a function to a data profile, you can choose from the following types of functions:

- **Source-level functions.** Performs calculations on two or more columns in a source, source group, or mapplet group. For example, you can evaluate a business rule for groups in an XML source.
- **Column-level functions.** Performs calculations on one column of a source. For example, you can evaluate the data in a column to find patterns that frequently occur in the data.
- **Intersource functions.** Performs calculations on two or more sources. These functions generate information about the relationship between the sources. For example, you might compare the values of columns in two sources to determine the percentage of identical data that appears in both sources.

Each function type has a subset of functionality that you can configure when you add a function to the data profile.

Configuring Default Data Profile Options

You can configure the default data profile options for the PowerCenter Client machine. Each data profile you create uses these default options. Configure default data profile options from the Profile Manager.

You can configure the following types of default data profile options:

- **General.** Set basic data profile options.
- **Report.** Set the maximum number of rows for each report grid.
- **Advanced.** Define prefixes for mappings, workflows, and sessions. Set domain inference and structure inference default parameters.
- **Auto Profile Default Functions.** Select the default functions to include when you create an auto profile.

To configure default data profile options, you can launch the Profile Manager from the following Designer tools:

- **Source Analyzer.** Click Sources > Profiling > Launch Profile Manager.
- **Mapplet Designer.** Click Mapplets > Profiling > Launch Profile Manager.
- **Repository Navigator.** Open a folder and select a source definition. Right-click on the source definition and select Launch Profile Manager.

Some data profiling functions may impact profile session performance.

Configuring General Options

You configure general options in the General tab of the Options dialog box.

The following table describes general options you can configure in the Profile Manager:

Option	Description
Always Save Changes Before Interactive Run	Saves changes to the profile mapping before running a profile session interactively. If you clear this option, the Designer prompts you to save changes before you run an interactive session. Default is enabled.
Display Profile Manager After Creating a Data Profile	Launches the Profile Manager after you create a data profile. If you clear this option, the Profile Manager does not launch immediately after you create a data profile. Default is enabled.
Always Run Profile Interactively	Runs a profile session interactively when you create a data profile. If you clear this option, you can still run auto and custom profile sessions interactively from the Profile Manager. Default is enabled.
Check in Profile Mapping When Profile Is Saved	Checks in profile mappings when you save changes for versioned repositories. If you are using a non-versioned repository, the Designer ignores this option. Saving versions of profile mappings in the repository can consume large amounts of disk space. Make sure that you have enough disk space on the machine hosting the repository. Default is disabled.
Always Invoke Auto Profiling Dialog	Displays the Auto Profile Column Selection dialog box when you create a new auto profile. If you clear this option, the Auto Profile Column Selection and the Auto Profile Function Selection pages in the Profile Wizard do not display for sources with 24 or fewer columns when you create a new data profile. As a result, you cannot configure domain and structure inference tuning options. Also, you cannot select to load verbose data for the auto profile session. Default is enabled.
Use Source Owner Name During Profile Mapping Generation	Adds the table owner name to relational sources when the Designer generates a profile mapping. Default is disabled. If the owner name changes after you generate the profile mapping, you must regenerate the mapping. You can regenerate a profile mapping in the Profile Manager.

Option	Description
Launch Workflow Monitor When Workflow Is Started	Launches the Workflow Monitor when you start a profile session. The Workflow Monitor runs in the background whenever you run a session. You can launch the Workflow Monitor from the Tools menu or by clicking the Workflow Monitor button. Default is disabled.
Session Log File Editor	Path for a text editor for the session log file. By default, the Profile Manager selects Wordpad as the text editor.
Session Log File Location	Location where the Integration Service writes the session log files.
Reset All	Restores default options.

Configuring Report Options

You configure Data Profiling report options in the Report tab of the Options dialog box.

The following table describes the report option you can configure in the Profile Manager:

Option	Description
Maximum Number of Rows per Report Grid	Maximum number of rows for a report grid. Default is 10.

Configuring Advanced Options

You can modify most of the advanced options in the Profile Settings dialog box when you create data profiles.

You configure prefix, domain inference, and structure inference options in the Advanced tab of the Options dialog box.

The following table describes advanced options you can configure in the Profile Manager:

Option	Description
Profile Mapping	Prefix to use for all profile mapping names. Profile mappings use the following naming convention: `<prefix><Profile Name>` Default prefix is m_DP_. Prefix must be 1 to 10 characters. It cannot contain spaces.
Profile Workflow Prefix	Prefix to use for all profile workflow names. Profile workflows use the following naming convention: `<prefix><Profile Name>` Default prefix is wf_DP_. Prefix must be 1 to 10 characters. It cannot contain spaces.
Profile Session Prefix	Prefix to use for all profile session names. Profile sessions use the following naming convention: `<prefix><Profile Name>` Default prefix is s_DP_. Prefix must be 1 to 10 characters. It cannot contain spaces.
Maximum Size of Column-Set to Analyze	Maximum size of the column set to be analyzed. This value limits the candidate key analysis, redundancy analysis, functional dependency analysis and intersource structure analysis to the column sets with this number of columns or fewer. For example, if you enter 3, the Integration Service does not return redundant column sets of four or greater. Enter a value from 1 to 7. Default is 2. The following functions use this setting: - Redundancy Analysis - Candidate Key - Intersource Structure Analysis - Functional Dependency Analysis

Option	Description
Maximum Allowable Error Candidate Key (%)	Maximum allowable error percentage required to determine unique candidate keys. The maximum allowable error percentage filters the results for the candidate key analysis and intersource structure analysis and restricts the analysis of the related super-sets. Enter a value from 0.00% to 99.99%. Default is 15%. The following functions use this setting: - Candidate Key - Intersource Structure Analysis
Maximum Allowable Error in Redundancy Analysis (%)	Maximum allowable error percentage required in redundancy analysis. For example, a value of 50% means that 50% or more of the values in a column must be redundant with another column or column set to determine a redundant column. If a column set meets the requirements for redundancy, the Integration Service does not return the related subsets. Enter a value from 0.00% to 99.99%. Default is 50%. The Redundancy Evaluation function uses this setting.
Maximum Allowable Error in Functional Dependencies (%)	Maximum allowable error percentage to determine functional dependencies among columns and column sets. For example, if you enter 10%, the data profile results show columns where 90% or more of the values exhibit a functional dependency across columns or column sets. The maximum allowable error percentage filters the results for functional dependency analysis and restricts the analysis of the related super-sets. Enter a value from 0.00% to 99.99%. Default is 10%. The Functional Dependencies Analysis function uses this setting.
Minimum Confidence Required for PK-FK or PK-PK Relationship (%)	Minimum confidence level required for the primary key-foreign key or primary key-primary key relationship. You define an acceptable percentage of accuracy called a confidence measure. Data Profiling uses this confidence measure to filter the relationships based on a sample of the source data. Enter a value from 0.01% to 100.00%. Default is 80%. The Intersource Structure Analysis function uses this setting.
Analyze String Datatype Columns of Precision up to	Column precision threshold for columns of string datatypes. For example, if you set the threshold to 20, the Integration Service does not profile String datatype columns with a precision greater than 20. Enter a value from 1 to 200. Default is 20. The following functions use this setting: - Redundancy Analysis - Candidate Key - Functional Dependencies Analysis - Intersource Structure Analysis
Analyze Integer Datatype Columns of Precision up to	Column precision threshold for columns of integer datatypes. For example, if you set the threshold to 28, the Integration Service does not return Integer datatype columns with a precision greater than 28. Enter a value from 1 to 38. Default is 28. The following functions use this setting: - Redundancy Analysis - Candidate Key - Functional Dependencies Analysis - Intersource Structure Analysis

Configuring Auto Profile Default Functions Options

You can define which functions the Profile Wizard includes by default when you create auto profiles. You can override the defaults when you create new auto profiles. Select at least one default auto profile function.

You define default column-level and source-level functions to include in new auto profiles in the Auto Profile Default Functions tab of the Options dialog box.

The following table describes the default auto profile functions:

Option	Description
Aggregate Functions	Calculates an aggregate value for numeric or string values in a source column. Default is selected.
Domain Inference	Reads all values in the column and infers a pattern that fits the data. Default is selected.

Option	Description
Distinct Value Count	Returns the number of distinct values for the column. Default is selected.
Row Count	Counts the number of rows read from a source. It can also report the number of rows in each group. Default is selected.
Candidate Key Evaluation	Calculates the number and percentage of unique values in one or more source columns. Default is cleared. This function may increase profile session processing time.
Redundancy Evaluation	Calculates the number of duplicate values in one or more source columns. Default is cleared. This function may increase profile session processing time.
Functional Dependency Analysis	Determines exact and approximate dependencies between columns and column sets within a source. Default is cleared. This function may increase profile session processing time.

For more information about auto profile Data Profiling functions, see “Source-Level Functions” on page 27 and “Column-Level Functions” on page 31.

Creating an Auto Profile

Create an auto profile to learn more about source data or mapplet output data during mapping development. When you create an auto profile, the Designer creates a data profile with the following functions:

- **Aggregate functions**. Calculates an aggregate value for numeric or string values in a column. Use aggregate functions to count null values, determine average values, determine minimum or maximum values, and minimum and maximum length for strings values.
- **Candidate Key Evaluation**. Calculates the number and percentage of unique values in one or more source columns.
- **Distinct Value Count**. Returns the number of distinct values for the column. You can configure the auto profile to load verbose data to the Data Profiling warehouse.
- **Domain Inference**. Reads all values in a column and infers a pattern that fits the data. You can configure the Profile Wizard to filter the Domain Inference results.
- **Functional Dependency Analysis.** Determines exact and approximate dependencies between columns and column sets within a source.
- **Redundancy Evaluation**. Calculates the number of duplicate values in one or more source columns.
- **Row Count**. Counts the number of rows read from the source during the profile session. When you create a data profile that uses the Row Count function with data samples, the Row Count function estimates the total row count.

Auto Profile Naming Conventions

The Designer uses the following naming conventions for an auto profile:

```
AP_<source/mapplet name>
```

For example, if you generate an auto profile for the source CustomerData, the Designer names the auto profile AP_CustomerData.

After you create the auto profile, the Designer generates a mapping based on the profile functions. The Designer uses the following default naming convention when it saves the profile mapping to the repository:

```
m_DP_AP_<source/mapplet name>
```

For example, if you create an auto profile called AP_CustomerData, the profile mapping name is m_DP_AP_CustomerData.

Tip: You can rename an auto profile in the Profile Manager. Click Description on the Auto Profile Column Selection page to change the name or description of the profile. Or, you can change the naming convention for profile mappings in the default data profile options.

If you have an auto profile for a source, and you generate another auto profile for the same source, the Designer does not overwrite the existing auto profile. It creates a new auto profile using the following naming convention:

```
AP_<source/mapplet name>N
```

where N is the latest version number of the previous auto profile plus 1. For example, if you have an auto profile AP_CustomerData, and you generate a new auto profile for the source CustomerData, the auto profile name is AP_CustomerData1.

The Designer generates a mapping for the new auto profile that uses the following default naming convention:

```
m_DP_AP_<source/mapplet name>N
```

where N is the latest version number of the auto profile plus 1.

Tip: As source data changes, you may need to create new auto profiles for the source. However, you can preserve existing auto profiles and their corresponding mappings in the Data Profiling warehouse for tracking or auditing purposes.

Steps to Create an Auto Profile

When you create an auto profile, you can profile groups or columns in the source. Or, you can profile the entire source. Auto profiling large sources impacts performance.

To create an auto profile:

1. Select the source definition in the Source Analyzer or mapplet in the Mapplet Designer you want to profile.
2. Launch the Profile Wizard from the following Designer tools:
   - **Source Analyzer.** Click Sources > Profiling > Create Auto Profile.
   - **Mapplet Designer.** Click Mapplets > Profiling > Create Auto Profile.

   Tip: You can right-click a source in the Navigator and select Profiling > Create Auto Profile from any Designer tool.

   The Auto Profile Column Selection page of the Profile Wizard appears in the following cases:
   - You set the default data profile options to open the Auto Profile Column Selection dialog box when you create an auto profile.
   - The source definition contains 25 or more columns.

   If the Auto Profile Column Selection page does not display, the Designer generates an auto profile and profile mapping based on the profile functions. Go to step 13.

   Note: If you skip this page, you cannot configure verbose data loading settings or domain or structure inference tuning settings.
3. Optionally, click Description to add a description for the data profile. Click OK.

   Enter a description up to 200 characters.
4. Optionally, select the groups or columns in the source that you want to profile.

   By default, all columns or groups are selected.
5. Select Load Verbose Data if you want the Integration Service to write verbose data to the Data Profiling warehouse during the profile session.

   By default, Load Verbose Data option is disabled.

   Loading verbose data for large sources may impact system performance.

Note: If you load verbose data for columns with a precision greater than 1,000 characters, the Integration Service writes truncated data to the Data Profiling warehouse during the profile session.

6. Click Next.

7. Select additional functions to include in the auto profile. You can also clear functions you do not want to include.

   The Profile Wizard selects the profile functions you specified in the default data profile options.

8. Optionally, click Save As Default to create new default functions based on the functions selected here.

9. Optionally, click Profile Settings to enter settings for domain inference and structure inference tuning.

   The Profile Settings dialog box displays the default domain inference tuning and structure inference settings.

10. Optionally, modify the default profile settings and click OK.

    The following table describes the domain and structure inference profile settings you can modify for auto profiles:

Option	Description
Maximum Number of Patterns	Integration Service returns the most frequently occurring patterns up to the number of patterns you specify. Enter a value between 1 to 1000. Default is 3.
Minimum Pattern Frequency	Integration Service returns patterns that occur at or above the frequency you specify. Enter a value between 0.01% to 100%. Default is 30%.
Maximum Size of Column-Set to Analyze	Analyzes column sets from one to seven columns.
Maximum Allowable Error Candidate Key (%)	Defines the threshold for the percentage of unique values in one or more source columns.
Maximum Allowable Error in Redundancy Analysis (%)	Defines the threshold for the percentage of redundant values in one or more source columns.
Maximum Allowable Error in Functional Dependencies (%)	Defines the threshold for the percentage of rows with functional dependencies between one or more source columns.

11. Click Configure Session to configure the session properties after you create the data profile.

12. Click Next if you selected Configure Session, or click Finish if you disabled Configure Session.

    The Designer generates a data profile and profile mapping based on the profile functions.

13. Configure the Profile Run options and click Next.

    The Session Setup page appears.

14. Configure the Session Setup options.

15. Click Finish.

    If you selected Run Session, the Profile Manager starts the session. If you cleared the Run Session option, the Profile Wizard saves the session properties you configured and closes.

Creating a Custom Profile

You can create a custom profile from the following Designer tools:

- **Source Analyzer.** Click Sources > Profiling > Create Custom Profile.

- **Mapplet Designer.** Click Mapplets > Profiling > Create Custom Profile.
- **Profile Manager.** Click Profile > Create Custom.

Note: You can right-click a source in the Navigator and select Profiling > Create Custom Profile from any Designer tool. If you create a custom profile this way, you can only profile that source. If you need to include multiple sources in the profile, or if you want to create an intersource function, use the Designer menu commands.

You can also edit or delete a data profile.

To create a custom profile, complete the following steps:

1. **Enter a data profile name and optionally add a description.** For more information, see "Step 1. Enter a Data Profile Name and Description" on page 15.
2. **Add sources to the data profile.** For more information, see "Step 2. Add Sources to the Data Profile" on page 15.
3. **Add, edit, or delete a profile function and enable session configuration.** For more information, see "Step 3. Add Functions and Enable Session Configuration" on page 16.
4. **Configure profile functions.** For more information, see "Step 4. Configure Profile Functions" on page 16.
5. **Configure the profile session if you enable session configuration.** For more information, see "Step 5. Configure the Profile Session" on page 18.

Step 1. Enter a Data Profile Name and Description

When you start the Profile Wizard, the General Properties page prompts you to enter a name for the data profile and add a description.

Data profile names must start with a letter and cannot contain the following characters:

```
.+-=~`!%^&*()[]{}'\";:/?,<>\\|\t\r\n @
```

When you are finished, click Next.

Step 2. Add Sources to the Data Profile

The Profile Sources page prompts you to select the sources you want to profile. If you selected the source definitions or mapplets you want to profile before you launched the Profile Wizard, this page does not display unless you selected a multi-group source definition.

If you select a multi-group source definition, the Profile Wizard adds all groups in the source definition to the data profile. You can add source definitions, mapplets, and groups in a multi-group source definition or output from multiple groups in mapplets to the data profile. When you profile a multi-group source, such as an XML source, you can select the groups in the source definition you want to profile or the entire source definition. For more information about eligible sources for profiling, see "Profiling Sources and Mapplet Output Data" on page 8.

To add sources to the data profile, select a source definition, a group in a source definition, or a mapplet and click the Add Source button. To remove a source, select the source definition, group, or mapplet and click the Remove Source button.

If you want to profile multiple sources, you can create a mapplet that combines multiple sources and create a data profile based on the mapplet output data.

Note: If you use a source as a lookup source within a data profile, it cannot be used as a non-lookup source within the same data profile. For example, when you create a Domain Validation function using a Column Lookup domain, the source you use for the column lookup cannot be a profiled source in the same data profile. If two profile sources attempt to validate data against each other, the Designer creates an invalid mapping.

Step 3. Add Functions and Enable Session Configuration

After you add sources to the data profile, use the Function-Level Operations page to complete the following tasks:

- **Add functions.** When you add functions to the data profile, the Profile Wizard opens the Function Details page for you to configure details about the functions.
- **Edit functions.** You can edit existing functions for the data profile.
- **Delete functions.** You can remove functions from the data profile.
- **Organize functions.** Use the Up and Down arrows to organize the functions in a data profile. The order of the functions does not affect the data profile results.
- **Configure Domain Inference function parameters.** You can widen or narrow the scope for the results from the Domain Inference function depending on whether you want to view the primary domains or exception data. Click Tuning Parameters to configure the number of patterns and pattern frequency for Domain Inference.
- **Configure Structure Inference function parameters.** You can adjust any of the following default parameters for structure inference functions:
  - Maximum column-set size to profile
  - Error percentage threshold for Candidate Key Analysis, Redundancy Analysis, and Functional Dependencies Analysis functions
  - Confidence threshold for primary key-foreign key or primary key-primary key relationships

  Click Tuning Parameters to configure the Structure Inference settings.
- **Select columns to load in verbose mode.** When you configure a function, you can select the columns to load in verbose mode.
- **Enable session configuration.** When you enable session configuration, the Profile Wizard prompts you to configure the profile session for the mapping. If you configured the default data profile options to always run profile sessions interactively, this option is selected by default.

If you finish adding functions to the data profile and you have not enabled session configuration, click Finish. The Profile Wizard generates the profile mapping.

If you finish adding functions to the data profile and you enabled session configuration, click Next. The Profile Wizard prompts you to configure the profile session.

Step 4. Configure Profile Functions

When you add a function to the data profile, the Function Details page prompts you to complete the following tasks:

- **Name the function.** Function names are not case sensitive and cannot contain spaces.

  Function names must start with a letter and cannot contain the following characters:

  ```
  .+-=~`!%^&*()[]{}'\";:/?,<>\\|\t\r\n @
  ```
- **Enter a description of the function.** Optionally, enter text to describe the function.
- **Select the type of function.** You can select source-level, column-level, or intersource functions.
- **Select a function.** The functions you can configure depends on the type of function you choose.

If you added multiple sources to the data profile, you must select the source you want to apply the function to.

If you select an intersource function, you must select at least two sources or two groups from different sources to apply the function to.

After you select the function type and function, click Next. The Profile Wizard prompts you to specify the function details for the function. The Function Details window and available options change depending on the type of function you select.

Each function type has a subset of functionality you can configure to perform calculations on the source data.

When you finish configuring the function, the Profile Wizard returns to the Function Level Operations page. From the Function Level Operations page, you can continue to add and configure functions for the data profile.

Configuring a Function with Group By Columns

Some functions let you generate profile data in a profile session run by group. When you configure a function, you can determine the column by which you want to group the data.

To select a group by column:

1. Configure a source-level function or column-level function.
2. Select Generate Profile Data By Group.
3. Click Group By.

   The Group By Columns dialog box appears.
4. Select the columns you want to group by.

   You can select up to three columns. You cannot group by the column for which you created the function. For example, if you created a Business Rule Validation function for the column Agreement_Status, you cannot select this column to group by.
5. Click OK.

Configuring a Function for Verbose Mode

When you configure a function for verbose mode, the Integration Service writes verbose data to the Data Profiling warehouse during a profile session.

You can configure verbose mode for the following functions:

- Source-level Business Rule Validation
- Column-level Business Rule Validation
- Domain Validation
- Distinct Value Count
- Row Uniqueness

The type of verbose data the Integration Service can load to the target depends on the function for which you configure verbose mode.

For most functions, you can load the following types of verbose data:

- **Rows that meet the business rule.** The Integration Service writes rows to the Data Profiling warehouse that meet the business rule in the function. For example, for a Domain Validation function, you might load the values that match the specified domain pattern.
- **Rows that do not meet the business rule.** The Integration Service writes rows to the Data Profiling warehouse that do not meet the business rule in the function. For example, when you create a Row Uniqueness function, you might load only duplicate rows.
- **All rows.** The Integration Service writes all verbose data rows to the Data Profiling warehouse. For example, when you create a Domain Validation function, you might load verbose data for values that match the domain pattern and those that do not.

Note: For Distinct Value Count and Row Uniqueness, you load duplicate rows or all rows as verbose data.

The following table describes the types of verbose data you can load with each function:

Function	Verbose Data Type Load Options
Source-level Business Rule Validation	No Rows, Invalid Rows Only, Valid Rows Only, All Rows
Column-level Business Rule Validation	No Rows, Invalid Rows Only, Valid Rows Only, All Rows

Function	Verbose Data Type Load Options
Domain Validation	No Rows, Invalid Rows Only, Valid Rows Only, All Rows
Distinct Value Count	No Rows, Duplicate Rows Only, All Rows You can select Load Only Selected Column as Verbose Data to load only the data for the column selected.
Row Uniqueness	No Rows, Duplicate Rows Only, All Rows

When you configure a function, you can select the columns to profile in verbose mode. The maximum precision is 1,000 characters. If the precision for the data in the column you select exceeds 1,000 characters, the Integration Service writes truncated data to the Data Profiling warehouse.

To configure a function for verbose mode:

1. Configure a function that can output data in verbose mode.
2. Select the type of verbose data to load on the Function Role Details page.
3. Click Finish.
4. Click Verbose Columns on the Function-Level Operations page.

   The Columns for Verbose Data dialog box appears. By default, all columns in the source are selected.
5. Clear the columns you do not want to profile in verbose mode.
6. Click OK.

Step 5. Configure the Profile Session

If you enabled session configuration on the Function-Level Operations page, the Profile Wizard opens the Profile Run page. You can configure and run a profile session or save the session configuration and run the profile session at another time.

Generating the Profile Mapping

After you create a data profile, the Designer generates a mapping based on the data profile metadata. You must save changes to store the new data profile and profile mapping in the repository. The Designer saves the data profile and profile mapping in the repository folder that stores the source or mapplet output you profiled.

You can view profile mappings in the Designer. The Designer denotes profile mappings in the Repository Navigator with a Profile Mappings icon.

The profile mapping name is based on the data profile name. By default, the mapping name contains the prefix m_DP_. For example, if you name the data profile SalaryValidation, the mapping name for the data profile is m_DP_SalaryValidation.

You can change the naming convention for profile mappings in the default data profile options.

Editing a Data Profile

You can edit a data profile to change any of the properties you configured. When you edit a data profile, the Designer regenerates the profile mapping.

You can edit an auto profile to change the data profile name or description and to add and delete functions. However, if you add a source to an auto profile, it becomes a custom profile.

When you delete a source for which you have defined functions in a data profile, the Designer marks the functions as invalid. If a function is invalid, the data profile is invalid. You can edit the functions to use a valid source or sources.

Similarly, if you delete a column, change the column datatype, or rename the column, all functions using this column are invalid. You can edit the functions to use the modified version of the column or another column.

To edit a data profile:

1. Launch the Profile Manager.

   You can launch the Profile Manager from the following Designer tools:

   - **Source Analyzer.** Click Sources > Profiling > Launch Profile Manager.
   - **Mapplet Designer.** Click Mapplets > Profiling > Launch Profile Manager.

   The Profile Manager appears.

2. From the Profile View, select the profile you want to edit.
3. Click Profile > Edit.

   The Profile Wizard appears.

4. Use the Profile Wizard to change any of the data profile properties.

Deleting a Data Profile

You can delete a data profile from the repository. When you delete a data profile, you can also delete the associated profile mapping from the repository. However, if you delete a profile mapping, the Designer does not delete the data profile associated with the mapping. You must delete the data profile manually.

Tip: If you delete a profile mapping, but do not want to delete the data profile, you can save the data profile and regenerate the profile mapping in the Profile Manager.

To delete a data profile:

1. Launch the Profile Manager.

   You can launch the Profile Manager from the following Designer tools:

   - **Source Analyzer.** Click Sources > Profiling > Launch Profile Manager.
   - **Mapplet Designer.** Click Mapplets > Profiling > Launch Profile Manager.

   The Profile Manager appears.

2. From the Profile View, select the profile you want to delete.
3. Click Profile > Delete.

   The Profile Manager prompts you to delete the selected data profile.

4. Click Yes to delete the data profile and the associated profile mapping.
5. Click OK.

Purging the Data Profiling Warehouse

The repository retains the related metadata and profile session results from deleted data profiles in the Data Profiling warehouse. You can purge metadata and profile session results from the Data Profiling warehouse for deleted data profiles. You can also purge profile session results that are no longer associated with data profile metadata.

Working with Profile Mappings

When you create a data profile, the Designer generates mappings to create the metadata for the data profile. You can copy or deploy profile mappings to other folders or other repositories. Or, you can combine profile mappings with other mappings in the Designer.

Copying Data Profiling Objects

After you create a data profile and the Designer generates the profile mapping, you can copy the mapping to another repository or repository folder. When you copy a profile mapping, the Designer does not copy the data profile associated with the mapping. To copy all the objects associated with a data profile, you must copy or deploy the entire folder to another repository or repository folder.

You cannot copy a profile mapping with a reusable domain. If you copy a mapping with a reusable domain and run a session for the mapping, the session fails.

Combining Data Profile Mappings with Other Mappings

You can combine profile mappings with other mappings. This allows the Integration Service to read the source data once to perform mapping logic and profile the data. To combine mappings, make a copy of the profile mapping and combine the copy with another mapping. Do not modify the original profile mapping. If you modify the original profile mapping, you may lose the changes you added because the Designer regenerates the mapping each time you edit it from the Profile Manager. When the Designer regenerates the mapping, it overwrites any changes.

Using Mapplets to Extend Data Profiling Functions

A function can operate on a column, source, or multiple sources. Sometimes, you need to combine data from multiple sources or multiple columns to use a particular function with it. Or, you may need to aggregate data to get the profiling results you want. For example, you want to create a Business Rule Validation function that operates on aggregate values from a source. You need to aggregate the values before you can profile the data using the Business Rule Validation function.

Use a mapplet when you want to profile the following information:

- Aggregate data from a single source
- Data from two or more sources with one or more matching ports
- Data from two sources with all matching ports

Extending Data Profiling Functionality with Mapplets

Complete the following steps to extend data profiling functionality with mapplets:

1. **Create the mapplet**. Create a mapplet to aggregate data or join or merge sources.
2. **Create a data profile using the mapplet output data as a source.** Create an auto profile based on the mapplet output data. Or, create a custom profile based on the mapplet output data and add functions to the data profile.
3. **Run the data profile.** Run the profile session. When you run the profile session from the Profile Manager, it processes the mapplet data as it would if you were running a workflow. You do not need to run a workflow to aggregate or join the data.
4. **View the Data Profiling report**. Open the Data Profiling report in the Profile Manager to view the results.

Profiling Aggregate Data

When you want to profile aggregate data, use a mapplet to aggregate the data before you create the data profile.

Use the mapplet to aggregate data when you want to use a column-level function on aggregate data. For example, you have an Employee Expenses flat file source that provides information on employee expenditures.

The following example shows the data from the Employee Expenses flat file source:

EID	Spending Date	Amount	Reason
12	12/3/2003	123.22	Acquired new books.
19	4/09/2004	600.21	Purchased ticket to Boston.
213	6/29/2004	215.61	Purchased new software.
12	6/12/2004	921.56	Acquired new books.
19	6/16/2004	740.21	Purchased ticket to New York.
21	7/21/2004	712.88	Purchased a new computer.

To test data consistency, you want to create a data profile that shows employees who spent over $1,000 in the last six months. To get this information, aggregate and filter the spending amounts before you create a data profile.

Creating the Mapplet

Create a mapplet to aggregate and filter the data. Later, use the Profile Manager to create a data profile for the mapplet.

When you create the mapplet, add a Filter transformation to filter out purchases older than six months. Add the following condition to the Filter transformation:

```
DATE_DIFF (SYSDATE,SPENDING_DATE,'MM')<6
```

After you filter the data, add an Aggregator transformation to aggregate the cumulative spending for each employee.

In the Aggregator transformation, add the cumulative_emp_spending output port to aggregate the total amount of money spent by each employee. Group the results by employee ID (EID), to see the total for each employee. Connect the EID port and the cumulative_emp_spending port to the Output transformation. You can then use the Profile Manager to profile the mapplet output data.

Note: You do not need to run a session to generate the correct mapplet output data. When you run a profile session, the Profile Manager processes the mapplet before it profiles the mapplet output data.

Creating the Data Profile

After you create the mapplet to filter and aggregate the data, you can profile the mapplet output data. From the Profile Manager, create a custom profile.

The custom profile locates the employees who spent more than $1,000 in the last six months using the source-level Business Rule Validation function. Create a Business Rule Validation function using the following expression:

```
cumulative_emp_spending > 1000
```

When you specify the type of verbose data to load to the Data Profiling warehouse, select valid rows only. This ensures that you can view the verbose data for the employees who spent over $1,000 in the last six months.

After you create the Data Profile, you can run a profile session.

Viewing the Data Profiling Report

After you run a profile session, you can view the profile results in a PowerCenter Data Profiling report. For example, when you view a PowerCenter Data Profiling report, you can view the verbose data for the rows that do not satisfy the Business Validation Rule. You can see that the employee with employee ID 19 has spent $1,340.42. All other employees have spent under $1,000, and therefore do not appear in the Verbose Report Summary.

Profiling Multiple Sources with One Matching Port

To profile two related sources with one or more matching ports, you can create a mapplet with a Joiner transformation to join the sources. Then, profile the mapplet output data.

Creating the Mapplet

For example, you have an Items relational source and a Manufacturers relational source. The Items relational source contains information about items, such as item description, wholesale cost, and item price. The Manufacturers relational source contains information about the manufacturers who manufacture the items. You want to find the manufacturers whose items sell with a markup that is greater than 50 percent of the wholesale cost.

You need to join the two sources before you can create a data profile. Create a mapplet using a Joiner transformation to join the two sources.

Use the following join condition:

```
MANUFACTURER_ID1 = MANUFACTURER_ID
```

Creating the Data Profile

After you join the sources, you can profile the mapplet output data to find the manufacturers whose items sell with a markup that is greater than 50 percent of the wholesale cost.

You create a custom profile with a source-level Business Rule Validation function and enter the following expression in the Rule Editor:

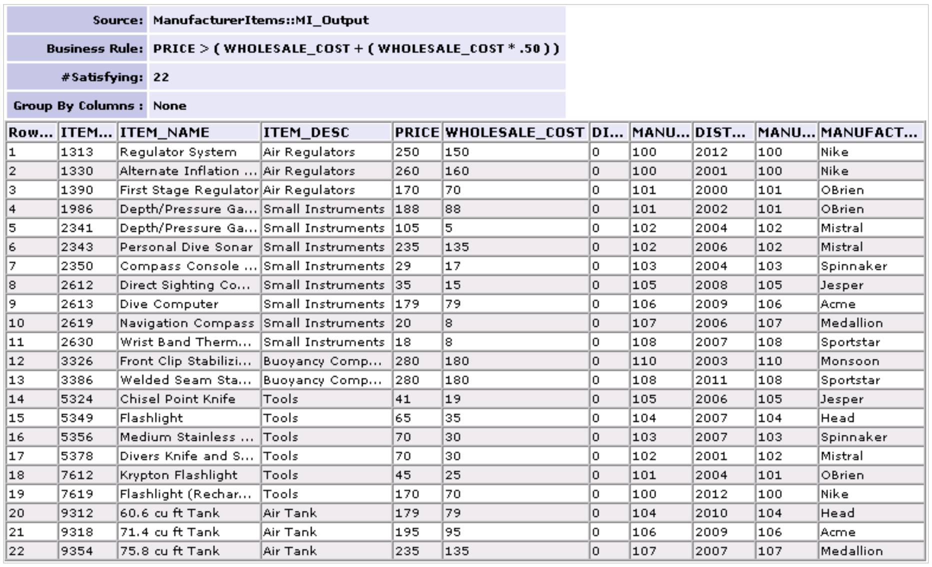
```
PRICE > (WHOLESALE_COST +(WHOLESALE_COST * .50))
```

When you specify the type of verbose data to load to the Data Profiling warehouse, select valid rows only. You can see the verbose data for the rows that meet the business rule.

Viewing the Data Profiling Report

Once you create a data profile, you can run a profile session. After you run a profile session, you can view the profile results in a PowerCenter Data Profiling report. For example, when you view a PowerCenter Data Profiling report, you can see the information for the companies whose items sell with a markup greater than 50 percent of the wholesale cost.

The following figure shows the Data Profiling report for the mapplet output data:



Source:	ManufacturerItems::MI_Output
Business Rule:	PRICE > (WHOLESALE_COST + (WHOLESALE_COST * .50))
#Satisfying:	22
Group By Columns :	None

Row...	ITEM...	ITEM_NAME	ITEM_DESC	PRICE	WHOLESALE_COST	DI...	MANU...	DIST...	MANU...	MANUFACT...
1	1313	Regulator System	Air Regulators	250	150	0	100	2012	100	Nike
2	1330	Alternate Inflation ...	Air Regulators	260	160	0	100	2001	100	Nike
3	1390	First Stage Regulator	Air Regulators	170	70	0	101	2000	101	OBrien
4	1986	Depth/Pressure Ga...	Small Instruments	188	88	0	101	2002	101	OBrien
5	2341	Depth/Pressure Ga...	Small Instruments	105	5	0	102	2004	102	Mistral
6	2343	Personal Dive Sonar	Small Instruments	235	135	0	102	2006	102	Mistral
7	2350	Compass Console ...	Small Instruments	29	17	0	103	2004	103	Spinnaker
8	2612	Direct Sighting Co...	Small Instruments	35	15	0	105	2008	105	Jesper
9	2613	Dive Computer	Small Instruments	179	79	0	106	2009	106	Acme
10	2619	Navigation Compass	Small Instruments	20	8	0	107	2006	107	Medallion
11	2630	Wrist Band Therm...	Small Instruments	18	8	0	108	2007	108	Sportstar
12	3326	Front Clip Stabilizi...	Buoyancy Comp...	280	180	0	110	2003	110	Monsoon
13	3386	Welded Seam Sta...	Buoyancy Comp...	280	180	0	108	2011	108	Sportstar
14	5324	Chisel Point Knife	Tools	41	19	0	105	2006	105	Jesper
15	5349	Flashlight	Tools	65	35	0	104	2007	104	Head
16	5356	Medium Stainless ...	Tools	70	30	0	103	2007	103	Spinnaker
17	5378	Divers Knife and S...	Tools	70	30	0	102	2001	102	Mistral
18	7612	Krypton Flashlight	Tools	45	25	0	101	2004	101	OBrien
19	7619	Flashlight (Rechar...	Tools	170	70	0	100	2012	100	Nike
20	9312	60.6 cu ft Tank	Air Tank	179	79	0	104	2010	104	Head
21	9318	71.4 cu ft Tank	Air Tank	195	95	0	106	2009	106	Acme
22	9354	75.8 cu ft Tank	Air Tank	235	135	0	107	2007	107	Medallion

Profiling Sources that Use All Matching Ports

When you want to profile data from two sources that use the same ports, use a mapplet to merge the sources and then profile the mapplet output data. For example, you have the order1 and order2 flat file sources that contain order information from two different stores. The sources use the same structure, but are stored in two different files. Both sources contain the following ports:

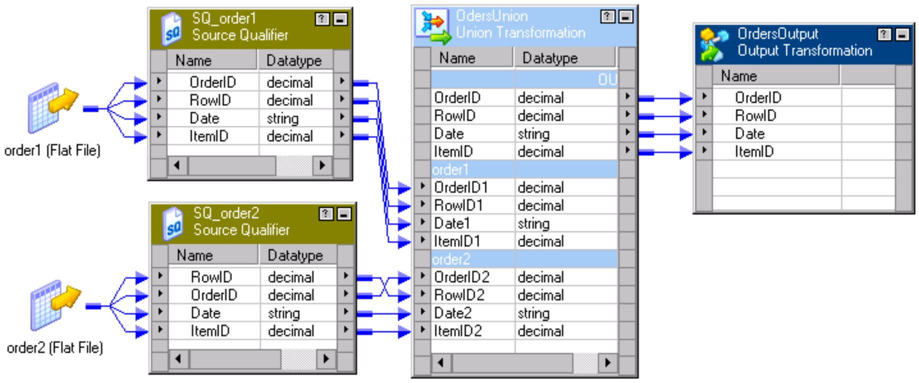
- OrderID
- RowID
- Date
- ItemID

You want to create an auto profile that displays profile data from both sources. You also want to ensure that the order IDs are distinct for order items in both stores.

Creating the Mapplet

To create an auto profile that displays data from both sources, create a mapplet that uses a Union transformation to merge the sources.

The following figure shows a mapplet that uses a Union transformation to merge source data:



After you merge the data using the Union transformation, you can profile the mapplet output data using the Profile Manager.

Creating the Data Profile

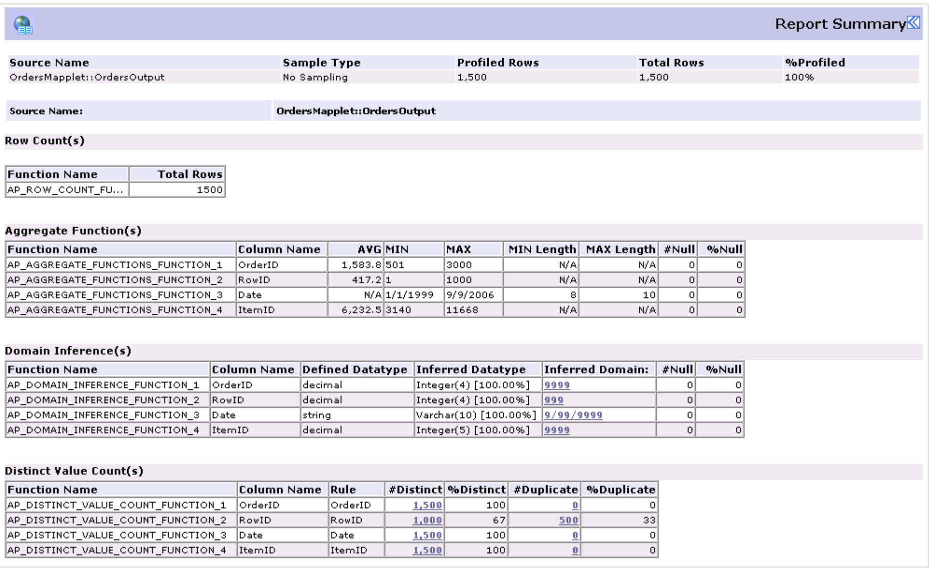
Create an auto profile based on the merged source data. This ensures that functions, such as the Row Uniqueness function, can generate reports based on all the data rather than only a part of the source data.

When you specify the type of verbose data to load to the Data Profiling warehouse, select all rows to see both valid and exception data.

Viewing the Data Profiling Report

After you create the data profile, you can run a profile session. After you run a profile session, you can view the profile results in a PowerCenter Data Profiling report. For example, when you view a PowerCenter Data Profiling report, you view data from both branches of the store. The percentage of duplicate values for the OrderID column is zero percent. Therefore, the Order ID for each order item is unique.

The following figure shows the Data Profiling report for the Union mapplet:



Report Summary

Source Name	Sample Type	Profiled Rows	Total Rows	%Profiled
OrdersMapplet::OrdersOutput	No Sampling	1,500	1,500	100%

Source Name:	OrdersMapplet::OrdersOutput

Row Count(s)

Function Name	Total Rows
AP_ROW_COUNT_FU...	1500

Aggregate Function(s)

Function Name	Column Name	AVG	MIN	MAX	MIN Length	MAX Length	#Null	%Null
AP_AGGREGATE_FUNCTIONS_FUNCTION_1	OrderID	1,583.8	501	3000	N/A	N/A	0	0
AP_AGGREGATE_FUNCTIONS_FUNCTION_2	RowID	417.2	1	1000	N/A	N/A	0	0
AP_AGGREGATE_FUNCTIONS_FUNCTION_3	Date	N/A	1/1/1999	9/9/2006	8	10	0	0
AP_AGGREGATE_FUNCTIONS_FUNCTION_4	ItemID	6,232.5	3140	11668	N/A	N/A	0	0

Domain Inference(s)

Function Name	Column Name	Defined Datatype	Inferred Datatype	Inferred Domain:	#Null	%Null
AP_DOMAIN_INFERENCE_FUNCTION_1	OrderID	decimal	Integer(4) [100.00%]	9999	0	0
AP_DOMAIN_INFERENCE_FUNCTION_2	RowID	decimal	Integer(4) [100.00%]	999	0	0
AP_DOMAIN_INFERENCE_FUNCTION_3	Date	string	Varchar(10) [100.00%]	9/99/9999	0	0
AP_DOMAIN_INFERENCE_FUNCTION_4	ItemID	decimal	Integer(5) [100.00%]	9999	0	0

Distinct Value Count(s)

Function Name	Column Name	Rule	#Distinct	%Distinct	#Duplicate	%Duplicate
AP_DISTINCT_VALUE_COUNT_FUNCTION_1	OrderID	OrderID	1,500	100	0	0
AP_DISTINCT_VALUE_COUNT_FUNCTION_2	RowID	RowID	1,000	67	500	33
AP_DISTINCT_VALUE_COUNT_FUNCTION_3	Date	Date	1,500	100	0	0
AP_DISTINCT_VALUE_COUNT_FUNCTION_4	ItemID	ItemID	1,500	100	0	0

Data Profiling Performance

Profiling data can consume available system resources and reduce performance. The following factors can impact performance:

- Data characteristics
- Data profile configuration
- Data profile and session settings
- Hardware

Performance Issues

The following factors may impact performance:

- **Source data size.** Profiling large amounts of data may cause sessions to run for a long time. As you profile data, consider source sizes and create data profiles to run efficiently. For example, you can profile a sample of source data or use a custom profile to only profile certain characteristics of the source data.
- **Datatypes.** Auto profile sessions for numeric data take longer to run than sessions that profile string data.

- **Verbose mode profile sessions.** Verbose data provides more details about the data that results from a profile function. However, a session might take longer to run in verbose mode. Run a session in verbose mode only if you expect the Integration Service to return a small number of rows.
- **Auto profiling.** Auto profiling large sources may increase session run time. You can limit the number of source columns in the data profile, select functions to include in a data profile, and adjust the profile settings to improve performance.
- **Custom profiling.** Including more than five columns in the Domain Validation function in a custom profile increases the processing time significantly.
- **Source column numbers.** The number of columns in a source directly impacts performance. For example, sources with 200 or more columns can cause sessions to run slowly. When you configure a data profile, you can select the columns you want to analyze. The fewer columns you select, the better the performance may be. You can also use mapplets to improve performance.

Improving Performance

You can optimize performance by taking the time to assess your data profiling requirements and desired results. Use the following tips to improve performance.

Use Custom Profiles or Edit Auto Profiles

Auto profile sessions may take a long time to complete. However, they are useful to learn basic information about your source data. After you use an auto profile to determine the information you would like to know about your source data, you can complete one of the following tasks to reduce session run time:

- **Edit the auto profile.** You can reduce the number of columns to profile, add and remove functions, or modify the profile settings.
- **Create a custom profile.** Configure only the functions you want to run against the source data and set the profile settings. This is more efficient than running all of the functions in an auto profile against the source data.

Modify Profile Settings

You can modify the default data profile settings for a data profile to improve performance. For example, in auto profiles you can reduce the maximum column set size to be analyzed.

Modifying structure inference settings can also improve performance. For example, you can increase the error margin of the Candidate Key Evaluation function.

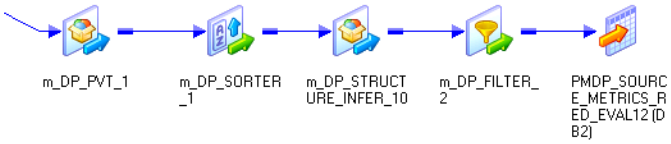
Increase the Aggregator Transformation Cache Size

If you know the approximate size of the source data, you can increase the Aggregate_2 and Aggregate_4 transformation cache size to match the source data size to improve performance. Aggregate_2 and Aggregate_4 transformations remove duplicate rows based on the number of rows. The Aggregate_2 transformation appears in all auto profiles. The Aggregate_4 transformation appears in custom profiles depending on the number of sources and which functions you included in the data profile.

Set the Aggregator transformation cache size on the Properties tab of the Edit Transformations dialog box for the transformation.

Editing an Auto Profile Mapping

Performance improves if you remove the following sequence of transformations from an auto profile mapping:



Removing these transformations disables the Candidate Key Evaluation, Redundancy Evaluation, and Functional Dependency Analysis functions in an auto profile. The Candidate Key Evaluation in particular can slow down performance.

Note: You can also select which functions to include in an auto profile and then edit the profile to remove these functions.

Automatic Random Sampling

You can improve performance by using automatic random sampling against relational sources rather than using manual sampling. Automatic random sampling reduces the load on the Integration Service.

Using Mapplets

Throughput decreases as column numbers increase. To more efficiently profile a source with many columns, split the source into mapplets and run the sessions in parallel.

Partition mapplets to improve performance. To gain maximum performance across the system resources, base the number of mapplets on the number of available CPUs.

Pipeline Partitioning

Depending on the source or target database, you can increase the number of partitions in a pipeline to improve session performance. When you increase the number of partitions, the Integration Service creates multiple connections to sources and targets and processes partitions of data concurrently. You can also use partitioning with mapplets within a data profile.

Use the following rules and guidelines when you use partitioning with a mapping that contains a structure inference function:

- Use only hash auto-keys partitioning in the Sorter transformation.
- Use only pass-through partitioning in the structure inference Custom transformation.
- Use only key range partitioning in the Source Qualifier transformation.

Troubleshooting

I received a message indicating that the data profile was successfully updated. The message also indicated that the data profile is not valid and must be run in interactive mode.

When you edit a data profile, the Designer validates the profile mapping. After validating the mapping, the Designer displays a message if the mapping includes blocking transformations. Mappings that include blocking transformations may cause the associated session to hang during the run.

CHAPTER 3

Working with Functions

This chapter includes the following topics:

Overview

You include functions in a data profile to perform calculations on sources during a profile session. When you create an auto profile, the Designer adds a predefined set of functions to the data profile. When you create a custom profile, you create functions that meet your business needs, and add them to the data profile. You can add the following types of functions to a data profile:

- **Source-level functions.** Perform calculations on two or more source columns, source group, or mapplet group. For more information, see "Source-Level Functions" on page 27.
- **Column-level functions.** Perform calculations on one column in a source. For more information, see "Column-Level Functions" on page 31.
- **Intersource functions.** Perform calculations on two or more sources, source groups, or mapplet groups. For more information, see "Intersource Functions" on page 35.

For many profile functions, you can write verbose data to the Data Profiling warehouse during a profile session.

Use Data Profiling reports to view information about the verbose data.

Source-Level Functions

Source-level functions perform calculations on two or more columns of a source, source group, or mapplet group.

You can add the following source-level functions to a data profile:

- **Row Count.** Counts the number of rows read from the source during the profile session. If you enable data sampling, the Row Count function returns an estimated row count based on the rows sampled. For more information, see "Row Count" on page 28.

- **Business Rule Validation.** Calculates the number of rows for one or more source columns that satisfy a specified business rule, and evaluates those rows that do satisfy the business rule. For more information, see “Business Rule Validation” on page 28.
- **Candidate Key Evaluation.** Calculates the number and percentage of unique values in one or more source columns. This helps you identify source columns you might use as a primary key. For more information, see “Candidate Key Evaluation” on page 29.
- **Redundancy Evaluation.** Calculates the number of duplicate values in one or more source columns. This helps you identify columns to normalize into separate tables. For more information, see “Redundancy Evaluation” on page 29.
- **Row Uniqueness**. Calculates the number of unique and duplicate values in the source based on the columns selected. You can profile all columns in the source row or choose individual columns to profile. For more information, see “Row Uniqueness” on page 30.
- **Functional Dependencies Analysis.** Determines exact and approximate dependencies between columns and column sets in a source. For more information, see “Functional Dependencies Analysis” on page 30.

When you specify a source-level function on the Function Details page of the Profile Wizard, the Profile Wizard prompts you to configure the function on the Function Role Details page. The options available on the Function Role Details page for source-level functions depend on the function you select.

Row Count

The Row Count function returns the number of rows in a source. It can also report the number of rows in each group. If you configure a session for random manual sampling or automatic manual sampling, the Row Count function returns an estimate of the total source rows. If you configure a session to sample First N Rows, the Row Count function returns the number of rows read during the session.

The following table describes the properties of the Row Count function:

Property	Description
Generate Profile Data by Group	Select to group source rows in a particular column. You can view a result for each group in the Data Profiling report.
Group by Columns	If you selected Generate Profile Data by Group, select a column to group by. You can select a column of any datatype except Binary. If you select a column of a numeric datatype, the precision must be between 1 and 28 digits. If you select a column of the String datatype, the precision must be between 1 to 200 characters.

Business Rule Validation

The source-level Business Rule Validation function calculates the number of rows for one or more columns in a source that satisfy a business rule and the number of rows that do not. Then, the function evaluates the rows that satisfy the business rule. A business rule is a valid Boolean expression that you create with the Rule Editor.

For example, you want to profile source data during mapping development to learn how many rows in the AVG_PRICE column in the source contain values greater than $50.00 and how many rows in the AVG_PROFIT column contain values greater than $25.00 for a specific manufacturer. You can define a business rule in a source-level Business Rule Validation function that evaluates each source row in the selected columns to see how many values satisfy the business rule.

The following table describes the properties of the source-level Business Rule Validation function:

Property	Description
Rule Summary	Click Rule Editor to enter a business rule. Once you enter a business rule, the rule appears in the Rule Summary dialog box. Use a valid Boolean expression. You can enter Business Rule Validation functions in the Business Rule Editor. If you enter other functions available through Business Rule Validation, such as Date or String functions, the session may generate unexpected results.
Generate Profile Data by Group	Select to group source rows in a particular column. You can view a result for each group in the Data Profiling report.
Group by Columns	If you selected Generate Profile Data by Group, select a column to group by. You can select a column of any datatype except Binary. If you select a column of a numeric datatype, the precision must be between 1 and 28 digits. If you select a column of the String datatype, the precision must be between 1 to 200 characters.
Specify the Type of Verbose Data to Load into the Warehouse	Select one of the following types of verbose data to load to the Data Profiling warehouse: - No Rows - Valid rows only - Invalid rows only - All Rows The character limit is 1000 bytes/K, where K is the maximum number of bytes for each character in the Data Profiling warehouse code page. If the column exceeds this limit, the Integration Service writes truncated data to the Data Profiling warehouse.

Candidate Key Evaluation

The Candidate Key Evaluation function calculates the number and percentage of unique values for one or more columns in a source. Use this function to determine the column in a source to use as a primary key. You can use the column with the highest percentage of unique values as the primary key.

You can define the maximum size of the column-set you want the Integration Service to analyze. For example, if you specify 3, the Integration Service does not consider keys like (A,B,C,D) since they have four columns. You can analyze column sets of up to seven columns.

You set the default column precision threshold in the Options dialog box in the Profile Manager. You can edit this value when you create a data profile.

The Candidate Key Evaluation function uses the following profile options:

- Maximum size of column-set to analyze
- Maximum allowable error Candidate Key (%)
- Analyze String datatype columns of precision up to
- Analyze Integer datatype columns of precision up to

The following table describes the properties of the Candidate Key Evaluation function:

Property	Description
Select Column(s)	Columns to profile. By default, the Designer selects all columns of numeric datatype with a precision between 1 and 28 digits or String datatype with a precision of 1 to 20 characters.

Redundancy Evaluation

The Redundancy Evaluation function calculates the number of duplicate values in one or more columns of a source. Use this function to identify columns to normalize into separate tables. You can normalize the columns that have the highest percentage of redundant values.

You can define the maximum size of the column-set you want the Integration Service to analyze. For example, if you specify 3, the Integration Service does not consider redundant column-sets like (A,B,C,D) since they have four columns. You can analyze column sets of up to seven columns.

You set the default column precision threshold in the Options dialog box in the Profile Manager. You can edit this value when you create a data profile.

The Redundancy Evaluation function uses the following profile options:

- Maximum allowable error in Redundancy Analysis
- Maximum size of column-set to analyze
- Analyze String datatype columns of precision up to
- Analyze Integer datatype columns of precision up to

The following table describes the properties of the Redundancy Evaluation function:

Property	Description
Select Column(s)	Columns you want to profile using the Redundancy Evaluation function. By default, the Designer selects all columns of numeric datatype with a precision between 1 and 28 digits or String datatype with a precision of 1 to 20 characters.

Row Uniqueness

The Row Uniqueness function calculates the number of unique and duplicate values based on the columns selected. You can profile all columns in the source row or choose columns to profile. This helps you identify columns to normalize into a separate table. You can also use this function to test for distinct rows.

You can use this function to analyze flat files, which have no internal validation tools, such as primary key constraints or unique indexes. For example, if you have a flat file source that uses unique employee ID values to identify each row, you can select all columns to test for duplicate rows. Because you have no primary key constraints, you can verify that you did not create duplicate entries for employees.

The following table describes the properties of the Row Uniqueness function:

Property	Description
Select Column(s)	Columns you want to profile using the Row Uniqueness function. By default, the Designer selects all columns of numeric datatype with a precision between 1 and 28 digits or String datatype with a precision of 1 to 10 characters.
Generate Profile Data by Group	Groups source rows in a particular column. You can view a result for each group in the Data Profiling report.
Group by Columns	If you selected Generate Profile Data by Group, select a column to group by. You can select a column of any datatype except Binary. If you select a column of a numeric datatype, the precision must be between 1 and 28 digits. If you select a column of the String datatype, the precision must be between 1 to 200 characters.
Specify the Type of Verbose Data to Load into the Warehouse	Select for the Integration Service to write verbose data to the Data Profiling warehouse. You can load the following types of verbose data: - All rows - No rows - Duplicate rows only The character limit is 1000 bytes/K, where K is the maximum number of bytes for each character in the Data Profiling warehouse code page. If the column exceeds this limit, the Integration Service writes truncated data to the Data Profiling warehouse.

Functional Dependencies Analysis

The Functional Dependencies Analysis function determines if the values in a column are dependent on the values of another column or set of columns. A column is functionally dependent on another column or set of columns when you can uniquely determine a value in one column based on the values in another column or column set. For example, postal codes are functionally dependent on city names and street addresses.

This function determines two types of dependencies:

- **Exact.** All values in a column show a functional dependency on values in another column or column set in the source data.
- **Approximate.** The values between one column and another column or column set do not satisfy functional dependency criterion in a defined percentage of rows. For example, a first name can often indicate gender. Therefore, gender has an approximate functional dependency on first name.

Both exact and approximate dependencies appear in Data Profiling reports. An exact dependency appears with 100% of the values matching. An approximate dependency appears with the percentage of values satisfying the function parameters.

Note: Each column in a source is functionally dependent on the primary keys of that source.

When you configure this function, you determine the minimum percentage of rows that must satisfy the function parameters. The Integration Service returns all rows within the limit you set.

You set a default percentage and column precision threshold in the Options dialog box in the Profile Manager. You can edit this value when you create a data profile.

The Functional Dependencies Analysis function uses the following profile options:

- Maximum size of column-set to analyze
- Maximum allowable error in Functional Dependencies (%)
- Analyze String datatype columns of precision up to
- Analyze Integer datatype columns of precision up to

When you create a Functional Dependencies Analysis function, select columns for which you want to infer a dependency. Click Finish.

Column-Level Functions

Column-level functions perform a calculation on one column in a source. You can add the following column-level functions to a data profile:

- **Business Rule Validation**. Calculates the number of rows in a single source column that satisfy and do not satisfy a specified business rule, and evaluates those rows that do satisfy the business rule. For more information, see “Business Rule Validation” on page 32.
- **Domain Validation**. Calculates the number of values in the profile source column that fall within a specified domain and the number of values that do not. When you create a Domain Validation function, you include domains in the function. For more information, see “Domain Validation” on page 32.
- **Domain Inference**. Reads all values in the column and infers a pattern that fits the data. For more information, see “Domain Inference” on page 33.
- **Aggregate Functions**. Calculates an aggregate value for a numeric or string value in a column. For more information, see “Aggregate Functions” on page 33.
- **Distinct Value Count**. Returns the number of distinct values for the column. For more information, see “Distinct Value Count” on page 34.

When you specify a column-level function on the Function Details page of the Profile Wizard, the Profile Wizard prompts you to configure the function. The options available on the Function Role Details page for column-level functions depend on the function you select.

Note: If you configure the profile session to write verbose data to the Data Profiling warehouse, the Integration Service loads all selected columns to the Data Profiling warehouse except those with the datatype Raw. You must select at least one column to write verbose data to the Data Profiling warehouse.

Business Rule Validation

The column-level Business Rule Validation function calculates the number of rows in a single source column that satisfy a business rule and the number of rows that do not. A business rule is a valid Boolean expression that you create with the Rule Editor.

The following table describes the properties of the column-level Business Rule Validation function:

Property	Description
Selected Column	Column you want to apply the business rule to.
Generate Profile Data by Group	Groups source rows in a particular column. You can view a result for each group.
Group by Columns	If you selected Generate Profile Data by Group, select a column to group by. You can select a column of any datatype except Binary. If you select a column of a numeric datatype, the precision must be between 1 and 28 digits. If you select a column of the String datatype, the precision must be between 1 to 200 characters.
Rule Summary	Click the Rule Editor button to enter a business rule. After you enter a business rule, the rule appears in the Rule Summary dialog box. Use a valid Boolean expression. You can enter Business Rule Validation functions in the Business Rule Editor. If you enter other functions available through Business Rule Validation, such as Date or String functions, the session may generate unexpected results.
Specify the Type of Verbose Data to Load into the Warehouse	Select to have the Integration Service write verbose data to the Data Profiling warehouse. You can load the following types of verbose data: - No Rows - Valid rows only - Invalid rows only - All Rows The character limit is 1000 bytes/K, where K is the maximum number of bytes for each character in the Data Profiling warehouse code page. If the column exceeds this limit, the Integration Service writes truncated data to the Data Profiling warehouse.

Domain Validation

The Domain Validation function calculates the number of values in a source column that fall within a specified domain and the number of values that do not. A domain is the set of all possible valid values for a column. For example, a domain might include a list of state abbreviations or a list of ZIP code patterns.

The following table describes the properties of the Domain Validation function:

Property	Description
Selected Column	Column you want to evaluate against the domain.
Domain Summary	Click the Domains button to select a reusable domain or create a non-reusable domain. After you enter a domain, it appears in the Domain Summary box.
Specify the Type of Verbose Data to Load into the Warehouse	Select to have the Integration Service write verbose data to the Data Profiling warehouse. You can load the following types of verbose data: - No Rows - Valid rows only - Invalid rows only - All Rows The character limit is 1000 bytes/K, where K is the maximum number of bytes for each character in the Data Profiling warehouse code page. If the column exceeds this limit, the Integration Service writes truncated data to the Data Profiling warehouse.

When you click the Domains button to add a domain, the Domain Browser appears.

Select the domain you want to use for the domain validation function. Click Close to return to the Domain Validation Function Role Details page.

If you validate the domain against a List of Values domain or a Domain Definition File Name domain, the list must use a code page that is two-way compatible with the Integration Service.

Domain Inference

The Domain Inference function reads all values in the column and infers a pattern that fits the data. The function determines if the source values fit a list of values derived from the domain column values. Use this function to determine data quality.

For example, if the source contains a column with social security numbers, use the Domain Inference function to determine the pattern of numbers in the column. The Domain Inference function can also infer a pattern of 'STRING WITH ONLY SPACES' for columns containing non-null blank space data.

This function can infer domains for columns with a numeric datatype with a precision of 28 digits or less or a String datatype with a precision of 200 characters or less. This function can also infer domains for columns of the Date/Time datatype.

When you create a Domain Inference function, select a column for which you want to infer a domain. Click Finish.

Configuring Domain Inference Settings

When you work with the Domain Inference function, you can configure the Profile Wizard to filter the Domain Inference results. You can narrow the scope of patterns returned to view the primary domains. Or, you can widen the scope of patterns returned to view exception data. You configure these settings from the Profile Settings dialog box when you create data profiles.

When you use a Domain Inference function in a data profile, you can configure the following settings in the Profile Settings dialog box:

- **Maximum number of patterns.** The Integration Service returns the most frequently occurring patterns up to the number of patterns you specify. For example, you create an auto profile for a source, and you set the maximum number of patterns to 20. The Integration Service returns the top 20 patterns.
- **Minimum pattern frequency.** The Integration Service returns patterns that occur at or above the frequency you specify. For example, if you set the minimum pattern frequency to 30 percent, the Integration Service returns patterns that occur 30 percent of the time or more and filters the rest.

For example, a source contains customers with Canadian and U.S. ZIP codes. Most of the customers are in the United States, but you also have Canadian customers. You want to view exception data and see what percentage of the ZIP codes are Canadian. Configure the data profile with a maximum of 300 patterns and a minimum pattern frequency of one percent.

When you run the profile session, the Integration Service returns a maximum of 300 patterns. You can view a wide range of exception data. The data profile also infers a domain for Canadian ZIP codes, which occurs in a small percentage of the data.

Note: When you configure minimum pattern frequency, the Integration Service ignores null data and calculates the percentage of patterns to return based on non-null data. Therefore, if you have null data in the source row, the percentage of patterns returned may not represent a percentage of the total data.

Aggregate Functions

An Aggregate function calculates an aggregate value for a numeric or string value applied to one column of a profile source.

You can add the following aggregate functions to a data profile:

- **NULL Value Count.** The number of rows with NULL values in the source column.
- **Average Value.** The average value of the rows in the source column.
- **Minimum Value.** The minimum value of the rows in the source column.
- **Maximum Value.** The maximum value of the rows in the source column.

The Aggregate function you can add to a source column depends on the datatype of the column.

The following table describes the Aggregate functions you can add based on the datatype of the source column:

Aggregate Function	Allowed Datatypes	Precision
NULL Value Count	Binary, Date/Time, Numeric, String	4,000 or less
Minimum Value	Date/Time, Numeric, String	200 or less
Maximum Value	Date/Time, Numeric, String	200 or less
Average Value	Numeric	1,000 or less

The following table describes the properties of the Aggregate Functions:

Property	Description
Selected Column	Column you want to apply the aggregation to.
Aggregate Functions	Functions you want to apply to the source column.
Generate Profile Data by Group	Groups source rows in a particular column. You can view a result for each group.
Group by Columns	If you selected Generate Profile Data by Group, select a column to group by. You can select a column of any datatype except Binary. If you select a column of a numeric datatype, the precision must be between 1 and 28 digits. If you select a column of the String datatype, the precision must be between 1 to 200 characters.

Distinct Value Count

The Distinct Value Count function returns the number of distinct values for the column. You can also enter a calculated expression to return the number of distinct values for the column based on the result of the expression.

For example, you want to determine the number of distinct SOUNDEX codes for last names in the table Customer_Data. You add Customer_Data to the data profile and create a Distinct Value Count function. You select the LAST_NAME column in the table to apply the function to and use the Rule Editor to create the following expression:

```
SOUNDEX(LAST_NAME)
```

When the Integration Service runs the profile session, it counts the number of distinct SOUNDEX codes that result from the expression and writes the values to the Data Profiling warehouse.

The following table describes the properties of the Distinct Value Count function:

Property	Description
Selected Column	Column you want to apply the function to.
Rule Summary	Click Rule Editor to enter an expression. Once you enter an expression, it appears in the Rule Summary dialog box. If you enter other functions available through the Rule Editor, such as Date or String functions, the session may generate unexpected results.
Generate Profile Data by Group	Groups source rows in a particular column. You can view a result for each group.

Property	Description
Group by Columns	If you selected Generate Profile Data by Group, select a column to group by. You can select a column of any datatype except Binary. If you select a column of a numeric datatype, the precision must be between 1 and 28 digits. If you select a column of the String datatype, the precision must be between 1 to 200 characters.
Specify the Type of Verbose Data to Load into the Warehouse	Select to have the Integration Service write verbose data to the Data Profiling warehouse. You can load the following types of verbose data: - All rows - No rows - Duplicate rows only The character limit is 1,000 bytes/K, where K is the maximum number of bytes for each character in the Data Profiling warehouse code page. If the column exceeds this limit, the Integration Service writes truncated data to the Data Profiling warehouse.

Intersource Functions

Intersource functions perform calculations on two or more sources, source groups from different sources, or mapplet output groups, and generate information about their relationship. You can add the following intersource functions to a data profile:

- **Referential Integrity Analysis**. Compares the values of columns in two sources to determine orphan values. For more information, see “Referential Integrity Analysis” on page 35.
- **Join Complexity Evaluation**. Measures the columns in multiple sources that satisfy a join condition. For more information, see “Join Complexity Evaluation” on page 36.
- **Intersource Structure Analysis.** Determines the primary key-foreign key relationships between sources. For more information, see “Intersource Structure Analysis” on page 36.

To create a custom profile and include an intersource function, use the Designer menu commands. If you right-click a source to create a custom profile, you cannot include additional sources in the profile.

Note: If you configure the profile session to write verbose data to the Data Profiling warehouse, the Integration Service loads all selected columns to the Data Profiling warehouse except those with the datatype Raw. You must select at least one column to write verbose data to the Data Profiling warehouse.

Referential Integrity Analysis

The Referential Integrity Analysis function compares the values of columns in two sources to determine orphan values. When you create this function, you select columns that you want to analyze. You can specify up to six join conditions for the function. During the profile session, the Integration Service determines the number and percentage of rows that appear in a specified column in the master source but not in the detail source. Use Referential Integrity Analysis to analyze orphan rows.

For example, you have an items table that includes item names, item IDs, and manufacturer IDs. You also have a item summary table that includes manufacturer IDs, average prices, and average profits. On the items table, the Manufacturer ID field corresponds to a manufacturer on the item summary table. You can use Referential Integrity Analysis to confirm the relationships and determine if any rows on the items table are orphans.

This function can evaluate unique values in columns of numeric datatypes with a precision of 28 digits or less or columns of the String datatype with a precision of 200 characters or less. The columns can use any combination of datatypes except Date and Numeric and Non-Raw with Raw. Use Raw with Raw when you disable verbose mode.

The following table describes the properties of the Referential Integrity Analysis function:

Property	Description
Port Name	Select columns from the list that you want to profile for each source. The Profile Wizard lists mapplets according to the mapplet name. It lists sources using the following syntax: <DBD name>:<source name> or FlatFile:<source name>.
Datatype	Datatype for the columns you want to profile in the corresponding source. The columns can use any combination of datatypes except Date and Numeric and Non-Raw with Raw. Use Raw with Raw when you disable verbose mode.
Specify the Type of Verbose Data to Load into the Warehouse	Select for the Integration Service to write all unmatched or orphaned rows to the Data Profiling warehouse. You can load the following types of verbose data: - No rows - Orphan rows The character limit is 1000 bytes/K, where K is the maximum number of bytes for each character in the Data Profiling warehouse code page. If the column exceeds this limit, the Integration Service writes truncated data to the Data Profiling warehouse.

Join Complexity Evaluation

The Join Complexity Evaluation function determines column values that satisfy join conditions. This function provides information to help you analyze join complexity. This is useful for designing and optimizing queries. You can select up to five sources and specify up to six join conditions for the Join Complexity Evaluation function.

For example, you have an items table that includes prices, manufacturers, and distributors. You also have an item summary table that includes maximum and minimum prices, manufacturer names, and manufacturer ID numbers. You want to determine the possible join conditions between the tables based on the data in the columns. Define the columns for the join conditions.

The Data Profiling report returns the Cartesian products for the two columns being profiled. It shows the number of rows containing each of the corresponding values in the profiled columns. You can use this information to develop queries to determine more specific information about the sources.

The following table lists sample output for the Join Complexity Evaluation function:

Column Value	Cartesian Product Value
100	4
101	3
102	4
103	3

When you configure the Join Complexity Evaluation function, select the column in each source that you want to profile. The columns can use any datatype except Binary. The columns can use any combination of datatypes except Date and Numeric, Raw with Raw, and Non-Raw with Raw.

Intersource Structure Analysis

The Intersource Structure Analysis function determines primary key-foreign key relationships between sources. It applies a confidence measure and identifies only those relationships that have confidence equal to or greater than the confidence measure. Use this function to analyze up to 15 sources. After you select the sources for the Intersource Structure Analysis function, select the columns you want to profile.

You define a confidence measure when you define the default options. A confidence measure is an acceptable percentage of accuracy. Data Profiling uses this confidence measure to filter the relationships based on a sample of the source data. The Intersource Structure Analysis function identifies the defined minimum percentage of rows that satisfy the primary key-foreign key inference confidence measure.

Define the minimum confidence value in the Options dialog box for all data profiles. Click the Tuning Parameters button to access the Profile Settings to modify this value when creating custom profiles.

The Intersource Structure Analysis function uses the following default data profile options:

- Maximum size of column-set to analyze
- Maximum allowable error Candidate Key (%)
- Minimum confidence required for PK-FK or PK-PK relationship (%)
- Analyze String datatype columns of precision up to
- Analyze Integer datatype columns of precision up to

After you run a profile session, the PowerCenter Data Profiling report displays the relative confidence measure to show the estimated accuracy of the relationships.

For example, you have the following tables:

- **Items.** Includes data about manufacturers, distributors, and item codes.
- **Item Summary.** Includes item codes, distributor data, and customer orders.
- **Manufacturers.** Includes manufacturer names and IDs.
- **Distributors.** Includes distributor names and IDs, manufacturer IDs, and item information.

You want to learn about relationships between the fields in each table. Use the Intersource Structure Analysis function to determine the primary key-foreign key relationships between the tables. The Data Profiling report shows the columns you selected for the function and the percentage of rows, based on a sample, that exhibit a primary key-foreign key relationship.

By default, no columns are selected. You can select the source to select all columns for a source. Or, you can expand the source to select columns from the source.

Tip: Place the pointer over a column to display the datatype, precision, and scale of the column.

The following table lists sample output for the Intersource Structure Analysis function:

Source Name	Primary Key Fields	Related Source Name	Related Fields	Relationship Type	Relative Confidence (%)
Manufacturers	Manufacturer_ID	Items	Manufacturer_ID	PK-FK	99
Manufacturers	Manufacturer_ID	Item Summary	Manufacturer_ID	PK-PK	99
Manufacturers	Manufacturer_Name	Item Summary	Manufacturer_Name	PK-PK	99
Distributors	Distributor_ID	Items	Distributor _ID	PK-FK	95
Item Summary	Manufacturer_ID	Items	Manufacturer_ID	PK-FK	99

CHAPTER 4

Working with Domains

This chapter includes the following topics:

Overview

Domains are sets of all valid values for a column. When you create a custom profile, you can create domains or use domains that Informatica provides. Some domains contain a list of all valid values that the source column can contain. Some domains contain a regular expression that describes a range or pattern of values that the source column can contain.

Use the following domains to profile source data:

- **Prepackaged domains.** Domains that Informatica provides.
- **Custom domains.** Domains that you create.

You can create reusable and non-reusable custom domains. Apply a reusable domain to multiple Domain Validation functions. Apply a non-reusable domain to one Domain Validation function. For more information about configuring a Domain Validation function, see “Column-Level Functions” on page 31.

You can create a domain from the Profile Manager or when you configure a function. Any domain you create from the Profile Manager is a reusable domain. Most custom domains can be reusable or non-reusable. Column Lookup domains are always non-reusable.

You can view reusable and prepackaged domains from the Domain Browser in the Profile Manager. You can view non-reusable domains from the Domain Browser when you define a function to which the non-reusable domain applies.

Prepackaged Domains

Informatica provides a set of prepackaged domains to verify data, such as phone numbers, postal codes, and email addresses. Use prepackaged domains when you create a custom profile with a Domain Validation function.

You can view prepackaged domains in the Domain Browser. Prepackaged domains install the first time you open the Domain Browser. It may take a few moments for the Domain Browser to open for the first time. Prepackaged domains are reusable.

From the Domain Browser, you can edit or delete prepackaged domains.

The following table describes the prepackaged domains and their descriptions:

Domain Name	Description
Country-Codes	Contains a list of country abbreviations.
Country-Names	Contains a list of country names.
Email-Address	Validates source against email patterns.
URL	Validates source against accepted URL patterns. Validates HTTP, HTTPS, and FTP URLs.
North-American-Phone-Number	Validates source against North American phone number patterns.
Dollar-Currency	Displays source currency in U.S. currency values.
US-Zip-Codes-Pattern (faster)	Validates the source against the U.S. ZIP code pattern.
US-Zip-Codes-List (more-accurate)	Contains a list of United States Postal Service ZIP codes.
US-State-Codes (50-states)	Contains a list of U.S. state abbreviations.
US-State-Codes (extended)	Contains a list of U.S. state abbreviations with additional values for territories and outlying areas served by the United States Postal Service.
US-State-Names (50-states)	Contains a list of the names of all U.S. states.
US-State-Names (extended)	Contains a list of the names of all U.S. states with additional values for territories and outlying areas.
US-Social-Security-Number	Validates the source against the U.S. social security number pattern.
Canadian-State-Codes	Contains a list of Canadian province abbreviations.
Canadian-State-Names	Contains a list of Canadian province names.
Canadian-Zip-Codes	Validates the source against the Canadian ZIP code pattern.
UK-Postal-Codes	Validates the source against the U.K. postal code pattern.
North-American-Industry-Classification-System (NAICS)-codes	Contains a list of the North American Industry Classification System codes.

Custom Domains

You include custom domains in a Domain Validation function when you configure a custom profile. During a profile session, the Domain Validation function uses the domains you specify to validate source values or to help you infer patterns from source data.

You can create the following types of domains:

- **List of Values.** Domains defined by a list of values.
- **Regular Expression.** Domains defined by a range of values in an expression.
- **Domain Definition File Name.** Domains defined by an external file containing values.
- **Column Lookup.** Domains defined by a column of a flatfile or relational source.

After you create a domain, you can edit or delete the domain.

List of Values Domains

You can create a List of Values domain from the Profile Manager or when you configure a function. When you create a List of Values domain, you add the values you want to apply to the domain. For example, to use a domain that lists area codes in a region, create a List of Values domain. You can also use a List of Values domain to evaluate data for non-null white spaces.

To create a List of Values domain:

1. To create a List of Values domain from the Profile Manager, click Tools > Domains.

   To create a domain when you define a function, click the Domains button on the Profile Function Details page.

   The Domain Browser dialog box appears.

2. Click New to create a new domain.

   The Domain Details dialog box appears.

3. Enter a name for the domain.

   The domain name cannot contain spaces.

4. Clear Reusable Domain if you do not want to be able to use this domain in other Domain Validation functions.

   If you configure a domain from the Profile Manager, the domain is reusable by default. You cannot make the domain non-reusable.

5. Select List of Values as the domain type.
6. In the Static field, enter a new domain value to manually add a value to the list of values.

   If you want to add a file with a list of values, go to step 9.

7. Click Add to add the domain value you entered to the list of values.

   Enter one value at a time and click Add to create the list. When you enter a domain value, the Designer ignores spaces before and after the value.

8. Repeat steps 6 to 7 for each domain value you want to add.
9. Optionally, click Values File to add a list of values.

   If you want to add domain values manually, go to step 6.

10. Navigate to the file, and select the file to use.

    Use a flat file or a relational source column as a domain.

11. Select the appropriate code page from the list.

    The code page you specify must be a subset of the code page for the operating system that hosts the PowerCenter Client. You can specify localization and code page information in the file list. If you do not specify the localization information, the Integration Service uses default values.

12. If you want to remove a domain value, select the value from the list of values and click Remove.
13. Click OK to save the domain.
14. Click Close.

Regular Expression Domains

Create a Regular Expression domain from the Profile Manager or when you configure a function. When you create a Regular Expression domain, you create a range of values you want to apply to the domain. For example, if you want to use a domain that includes values for 5-digit employee ID numbers, create a Regular Expression Domain.

Using perl Compatible Regular Expression Syntax

You must use perl compatible regular expression syntax when you create a Regular Expression domain.

The following table describes perl compatible regular expression syntax guidelines to create a regular expression:

Syntax	Description
. (a period)	Matches any one character.
[a-z]	Matches one instance of a letter. For example, [a-z][a-z] can match ab or CA.
\d	Matches one instance of any digit from 0-9.
\d	Matches one instance of any digit from 0-9.
\s	Matches a whitespace character.
()	Groups an expression. For example, the parentheses in (\d-\d-\d\d) groups the expression \d\d-\d\d, which finds any two numbers followed by a hyphen and any two numbers, as in 12-34.
{}	Matches the number of characters. For example, \d{3} matches any three numbers, such as 650 or 510. Or, [a-z]{2} matches any two letters, such as CA or NY.
?	Matches the preceding character or group of characters zero or one time. For example, \d{3}(-{d{4})? matches any three numbers, which can be followed by a hyphen and any four numbers.
* (an asterisk)	Matches zero or more instances of the values that follow the asterisk. For example, *0 is any value that precedes a 0.
+	Matches one or more instances of the values that follow the plus sign. For example, \w+ is any value that follows an alphanumeric character.

For example, to create a regular expression for U.S. ZIP codes, you can enter the following perl syntax:

```
\d{5}(-\d{4})?
```

This expression lets you find 5-digit U.S. ZIP codes, such as 93930, and 9-digit ZIP codes, such as 93930-5407.

In this example, \d{5} refers to any five numbers, such as 93930. The parentheses surrounding -\d{4} group this segment of the expression. The hyphen represents the hyphen of a 9-digit ZIP code, as in 93930-5407. \d{4} refers to any four numbers, such as 5407. The question mark states that the hyphen and last four digits are optional or can appear one time.

When you enter a regular expression, you can validate the regular expression with test data. Use the test data instead of the source data in the repository. The test data should represent the source data you plan to use the Regular Expression domain against.

Converting COBOL Syntax to perl Compatible Regular Expression Syntax

If you are familiar with COBOL syntax, use the following information to help you write perl compatible regular expressions for a Regular Expression Domain.

The following table describes examples of COBOL syntax and their perl compatible equivalents:

COBOL Syntax	perl Syntax	Description
9	\d	Matches one instance of any digit from 0-9.
9999	\d\d\d\d or \d{4}	Matches any four digits from 0-9, as in 1234 or 5936.
x	[a-z]	Matches one instance of a letter.
9xx9	\d[a-z][a-z]\d	Matches any number followed by two letters and another number, as in 1ab2.

Converting SQL Syntax to perl Compatible Regular Expression Syntax

If you are familiar with SQL syntax, use the following information to help you write perl compatible regular expressions for a Regular Expression domain.

The following table describes examples of SQL syntax and their perl compatible equivalents:

SQL Syntax	perl Syntax	Description
%	. *	Matches any string.
A%	A.*	Matches the letter “A” followed by any string, as in Area.
_	. (a period)	Matches any one character.
A_	A.	Matches “A” followed by any one character, such as AZ.

Creating a Regular Expression Domain

Use the following procedure to create a Regular Expression domain.

To create a Regular Expression domain:

1. To create a domain from the Profile Manager, click Tools > Domains.

   To create a domain when you define a function, click the Domains button on the Profile Function Details page.

   The Domain Browser dialog box appears.

2. Click New to create a new domain.

   The Domain Details dialog box appears.

3. In the Domain Details dialog box, enter a name for the domain.

   The domain name cannot contain spaces.

4. Clear Reusable Domain if you do not want to be able to use this domain in other Domain Validation functions.

   If you configure a domain from the Profile Manager, the domain is reusable by default. You cannot make the domain non-reusable.

5. In the Expression box, enter a new domain value.

6. To validate the regular expression, click Test Data.

   The Test Data dialog box appears.

7. Enter the test data you want to test the regular expression against.

8. Click Test to validate the regular expression.

   You can view the results of the test in the Result Box at the bottom of the Test Data Dialog box.

9. When you finish validating the regular expression, click Done.

10. Click OK to save the domain.

11. Click Close.

Domain Definition File Name Domains

Create a Domain Definition File Name domain from the Profile Manager or when you configure a function. You create a Domain Definition File Name domain when you have a text file you want to use to validate source data. For example, you have a file that contains a list of area codes in a region. Create a Domain Definition File Name domain for the Integration Service to use the file to validate the source data according to this list.

When you create a Domain Definition File Name domain, select a file name with the domain values you want to use. Store the file on the same machine as the Integration Service on which you want to run the session. Use the $PMSourceDir process variable when you specify the file name path.

To create a Domain Definition File Name domain:

1. To create a domain from the Profile Manager, click Tools > Domains.

   To create a domain when you define a function, click the Domains button on the Profile Function Role Details page.

   The Domain Browser dialog box appears.

2. Click New to create a new domain.

   The Domain Details dialog box appears.

3. Enter a name for the domain.

   The domain name cannot contain spaces.

4. Clear Reusable Domain if you do not want to be able to use this domain in other Domain Validation functions.

   If you configure a domain from the Profile Manager, the domain is reusable by default. You cannot make the domain non-reusable.

5. In the Static box, enter a new domain value.

   Use process variables, such as $PMSourceFileDir and $PMRootDir, when you specify a file name and path for the domain value.

   Note: The file you specify must use a code page that is a subset of the Integration Service code page. You can specify a code page by entering valid syntax on the first line of the file.

6. Click OK to save the domain.

7. Click Close.

Column Lookup Domains

Create a Column Lookup domain when you configure a Domain Validation function. When you create a Column Lookup domain, you select the column of a flat file or relational source containing the values you want to apply to the domain. For example, to use a domain that lists distributor IDs, create a Column Lookup domain using a source file with distributor IDs as a column. Column Lookup domains are not reusable.

To create a Column Lookup domain:

1. To create a Column Lookup domain when you define a function, click the Domains button on the Profile Function Details page.

   The Domain Browser dialog box appears.

2. Click New to create a new domain.

The Domain Details dialog box appears.

3. Enter a name for the domain.

   The domain name cannot contain spaces.

4. Select Column Lookup as the domain type.

   Column Lookup domains are not reusable.

5. Navigate to the file, and select the file and column to use.

   Use a flat file or a relational source column as a domain.

6. Click OK to add the selected column and save the domain.

Specifying Localization and Code Page Information

When you import a list of values or a domain definition file name, the file can contain localization information.

Use the following guidelines to specify localization information:

- Enter localization information in the first line of the file.
- Enter localization information using 7-bit ASCII.
- Use the following syntax:

  ```
  locale=<language>_<territory>.<codepage>@<sort>
  ```

  where language, territory, code page, and sort represent the following information:

Parameter	Description
Language	Language to use for month names and weekday names and other territory independent items.
Territory	Country or territory name to reference country dependent information such as currency symbols, numeric and monetary formatting rules, and Date/Time formats.
Codepage	Character encoding to use. The code page you specify must be a subset of the code page for the operating system that hosts the PowerCenter Client.
Sort	Collation sequence to use. For example, use Binary.

For example, you can specify the following localization information for a U.S. English file:

```
locale=English_UnitedStates.US-ASCII@binary
```

For a Japanese file, you can specify the following localization information:

```
locale=Japanese_Japan.JapanEUC@binary
```

Editing a Domain

You can edit a reusable domain to change the domain name and domain value, expression, or file name. When you define a function, you can edit a non-reusable domain for the function to change the domain name, domain type, and domain value, expression, or file name. You can edit prepackaged domains to change the domain value, expression, or file name.

To edit a domain:

1. To edit a reusable or prepackaged domain from the Profile Manager, click Tools > Domains.

   To edit a domain when you define a function, click the Domains button on the Profile Function Role Details page.

The Domain Browser dialog box appears.

2. Select the domain you want to edit, and click Edit.

   The Domain Details dialog box appears.

3. If you are editing a custom domain, optionally change the domain name.

   If you are editing a List of Values domain, add or remove domain values.

   If you are editing a Regular Expression domain, modify the domain expression.

   If you are editing a Domain Definition File Name domain, modify the file name that contains the domain values.

4. Click OK.

5. Click Close.

Deleting a Domain

You can delete a domain if you no longer want to apply it to Domain Validation functions. If you delete a domain, the Designer invalidates all of the data profiles and related profile mappings that reference the domain.

To delete a domain:

1. To delete a domain from the Profile Manager, click Tools > Domains.

   To delete a domain when you define a function, click the Domains button on the Profile Function Role Details page.

   The Domain Browser dialog box appears.

2. Select the domain you want to delete, and click Delete.

3. Click Close.

CHAPTER 5

Running Profile Sessions

This chapter includes the following topics:

Overview

To generate information about source data from a data profile, create and run a profile session. You can create and run profile sessions from the following tools:

- **Profile Manager.** You can run sessions from the Profile Manager immediately after you create or edit a data profile to quickly obtain profile results from a source. You can also run a session from the Profile Manager when you want to profile a sample of source data instead of the entire source. When you create sessions from the Profile Manager, the Profile Manager creates a workflow and associates it with the session.
- **Workflow Manager.** If you want to monitor ongoing data quality issues, you can create a persistent session and workflow for the profile mapping in the Workflow Manager and add a scheduling task. You can perform a time-dimensional analysis of data quality issues. You can also edit and run persistent sessions that you create in the Profile Manager from the Workflow Manager.

You can run the following types of sessions:

- **Temporary.** Temporary sessions run on demand and are not stored in the repository. You can only run temporary sessions from the Profile Manager. You cannot run temporary sessions from the Workflow Manager.
- **Persistent.** Persistent sessions can run on demand and are stored in the repository. You can run persistent sessions from the Profile Manager and the Workflow Manager.

After you run a profile session, you can monitor the session from the Profile Manager or from the Workflow Monitor.

Working with Data Samples

When you run a session from the Profile Manager, you can create a data profile based on a sample of data rather than the entire source. You use data sampling when you want to understand general trends within the data or view some exception data.

Running Sessions from the Profile Manager

You can run a profile session from the Profile Manager to quickly profile sources during mapping development. A session you run from the Profile Manager is called an interactive session.

You can run sessions from the Profile Manager in the following cases:

- Immediately after you create a data profile
- At any time for an existing data profile

You can run temporary or persistent sessions from the Profile Manager. When you run a temporary session, the Integration Service uses normal mode to load data to the target.

Running a Session when You Create a Data Profile

You can configure and run profile sessions when you create a data profile. To do this, select Configure Session on the Auto Profile Column Selection page of the Profile Wizard when you create an auto profile. Or, select Configure Session on the Function-Level Operations page of the Profile Wizard when you create a custom profile. After the Profile Wizard generates the profile mapping, it prompts you to configure and run the session.

If you want the Integration Service to run a session immediately after you create a data profile by default, configure the Always Run Profile Interactively option in the default data profile options.

Running a Session for an Existing Data Profile

If you want the Integration Service to run a session for an existing data profile, select the data profile in the Profile Manager and click Profile > Run. The Profile Wizard prompts you to configure the session before running it.

Configuring a Session in the Profile Wizard

You must configure a data profiling session before you can run it. When you configure a session to run from the Profile Manager, you create and configure the session in the Profile Wizard.

To configure a data profiling session:

1. From the Profile Wizard, select Configure Session on the Function-Level Operations page or the Auto Profile Function Selection page and click Next.

   From the Profile Manager, select the data profile in the Profile Navigator and click Run.

   The Profile Run page of the Profile Wizard appears.

2. Configure the following properties:

Profile Run Properties	Description
Server	Select an Integration Service to run the profile session.
Profile Run	Select one of the following options: - Create a persistent session for this data profile. A persistent session is stored in the repository. - Create a temporary session for this data profile. A temporary session is not stored in the repository.
Sampling	Select one of the following options: - No Sampling. Select to disable profile sampling. When you disable sampling, the Designer generates a data profile based on all selected source data. - Sample First N Rows. Select to read the first N rows from the source up to the number of rows you specify. - Automatic Random Sampling. Select to run a data profile for a random sample of source data and allow the Integration Service to determine the percentage of data to sample. The percent sampled represents a percentage of the total row count. The Integration Service scales the percent of data to sample based on the total row count. - Manual Random Sampling. Select to specify a percentage of data to sample. You can specify 1 to 100 percent. The Integration Service selects random row sets from all parts of the source data. Note: If you select Sample First N Rows, the total rows displayed in the Data Profiling report header will be the same as the number of rows selected for sampling.

3. Click Next.

   The Session Setup page of the Profile Wizard appears.

4. Configure the following session properties:

Session Setup Properties	Description
Source Properties	Configure source connection properties on the Connections tab. Configure source properties on the Properties tab. Configure reader properties on the Reader tab. The Source properties are the same as those in the session properties you configure in the Workflow Manager.
Target Connections	Relational database connection to the Data Profiling warehouse database. This is the relational database connection you configured for the Data Profiling warehouse in the Workflow Manager.
Reject File Directory	Directory for session reject files. Default reject file directory is $PMBadFileDir\.
Run Session	Runs the session immediately. Otherwise, the Profile Manager saves the session configuration information and exits.

Monitoring Interactive Sessions

When you run an interactive session, you can monitor the session from the Profile Manager. The profile node icon in the Navigator Window changes depending on the status of the session. The session status appears in the Profile Session Status window.

You can also monitor the profile session from the Workflow Monitor. The Integration Service creates a workflow for profile sessions.

If a persistent interactive session fails, the Integration Service writes a session log. You can review the session log, correct any errors, and restart the session. To view the session log, click View > Session Log in the Profile Manager.

When the session successfully finishes, you can view reports that contain the profile results.

Creating a Session in the Workflow Manager

You can create persistent profile sessions and workflows in the Workflow Manager. You can create and configure a profile session and workflow from the Workflow Manager to monitor ongoing data quality issues, or if you want to have more control over the session and workflow properties. Also, you can add workflow tasks, such as a scheduler that lets you schedule the workflow to run regularly. This lets you perform a time-dimensional analysis of data quality issues.

When you create a persistent session and workflow in the Workflow Manager, the following guidelines apply:

- You must run the session from the Workflow Manager.
- You must view the session status in the Workflow Monitor.
- If you create a session with the test load option enabled in the workflow session properties, the rows will not be rolled back at the end of the profile session.
- You can run the session in real time.

You can view Data Profiling reports that you generate from the Workflow Manager in the Profile Manager.

Profiling Data Samples

When you run a session from the Profile Manager, you can create a data profile based on a sample of data rather than the entire source. Use data sampling when you want to understand general trends within the data or view some exception data.

For example, you have a large data source and you want to quickly verify that a business rule you created in a custom profile returns the data you expect. To test this business rule, run a profile session based on a sample of data to verify that the business rule returns the data you expect. You can quickly test the data without running a profile session on a large data source.

When you run a profile session with data sampling, the Data Profiling report shows the estimated total number of rows in the source, the number of rows sampled, the mode of sampling used, and the percentage of the total source sampled.

When you sample relational data, the Integration Service can sample data by delegating sampling to the database or by sampling as the Integration Service reads data from the database. You may be able to improve data sampling performance and accuracy by enabling the Integration Service to delegate sampling to the database.

Complete the following steps to profile samples of data:

1. Select a function type.
2. Select a data sampling mode.

Step 1. Select a Function Type

When you create data profiles, some functions can return useful data using data samples. Other functions need to perform calculations on all source data to return useful information.

The following table describes function behavior with data samples:

Function Type	Behavior when Used with Data Samples
Source-level functions	Data profiles created with source functions and data samples can display general patterns within the data.
Column-level functions	Data profiles created with the following column-level functions and data samples can display general patterns within the data: - Domain Inference - Business Rule Validation - NULL Count and Average Value Aggregate functions Minimum and Maximum Value Aggregate functions can have inconsistent results because a column can have unusually high maximum values or unusually low minimum values. Distinct Value Count function returns values from the sample records.
Intersource functions	You cannot use data samples with intersource functions.

Step 2. Select a Data Sampling Mode

When you create a profile session with data sampling, you can use a different sampling mode depending on the percentage of data you want to sample. Choose from the following data sampling modes:

- **Automatic Random Sampling.** Allows the Integration Service to determine the percentage of data to sample. The Integration Service scales the percentage of data to sample based on the size of the source data. If you do not know the size of the source data, you can enable the Integration Service to determine the percentage of data to sample.
- **Manual Random Sampling**. You can specify a percentage of data to sample from 1 to 100. The Integration Service selects random data from all parts of the source. Use this option if you want to control the percentage of data to sample.
- **Sample First N Rows**. Select to read the first N rows from the source up to the number of rows you specify. Use First N Row sampling when you cannot use manual random sampling for the profile function. You can also use First N Row sampling when you want to specify the number of rows to profile. For example, if you have a very large source, you can sample the first 100 rows to understand some basic information about the source.

Profiling Relational Data Samples

When you create a profile session based on relational data samples, the Integration Service can obtain data samples in the following ways:

- **By sampling when the Integration Service reads data**. The Integration Service performs a sampling algorithm on the source data when it reads the data from the database.
- **By delegating sampling to the database.** The Integration Service delegates the sampling operation to the database, and the Integration Service reads the sampled data after the sampling operation has taken place on the database.

When you run a profile session configured to sample data, the Integration Service first attempts to delegate sampling to the database. If it cannot delegate sampling to the database, it runs an algorithm on the data as it reads the data from the database.

When you run a profile session, the Integration Service can delegate sampling to the Oracle and DB2 databases.

Often, you can optimize performance and accuracy by enabling the Integration Service to delegate sampling to the database.

To configure the Profile Manager to allow the Integration Service to delegate sampling to the database, ensure that you use native database drivers to connect to the database. You do not need to complete any other steps to configure the Integration Service to delegate sampling.

Performing Sampling as the Integration Service Reads Data

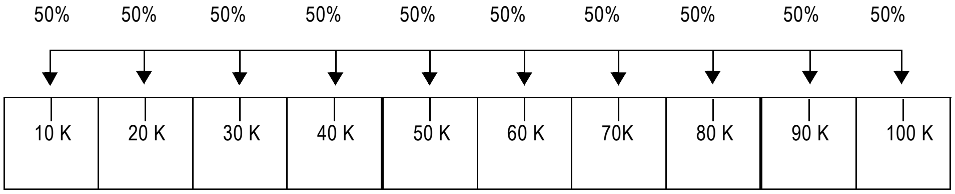
When the Integration Service performs sampling while it reads data, it selects the data samples using the C rand function. You can specify that the Integration Service samples data in the following ways:

- **Manual random sampling**. You specify a percentage of data to sample, from 1 to 100. The Integration Service selects random data from the source up to the percentage you specify.
- **Automatic random sampling**. The Integration Service determines the percentage of data to sample based on the size of the source data.

Using Manual Random Sampling

When you use manual random sampling, you specify a percentage of data for the Integration Service to sample. The Integration Service samples that percentage of data as it reads rows. For example, if you have 100,000 rows and you sample 50 percent of the data, the Integration Service reads 50 percent of the first 10,000 rows, and 50 percent of the next 10,000 rows, until it samples 50 percent of the data across the data set.

The following figure shows a manual random sample:

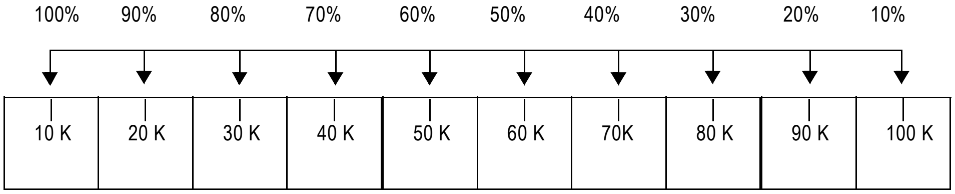


Using Automatic Random Sampling

When the Integration Service performs automatic random sampling, it determines the percentage of data to sample as it reads data from the database. Because you do not specify the percentage of data to sample, the Integration Service begins by sampling 100 percent of the data and gradually samples progressively less data as it reads from the database. The Integration Service does this to ensure that it samples sufficient data for accuracy.

For example, you have 100,000 rows of data to sample. When the Integration Service reads the data, it samples 100 percent of the first 10,000 rows, 90 percent of the next 10,000 rows, 80 percent of the next 10,000 rows, and progressively less data for each group of 10,000 rows.

The following figure shows how data is sampled using automatic random sampling:

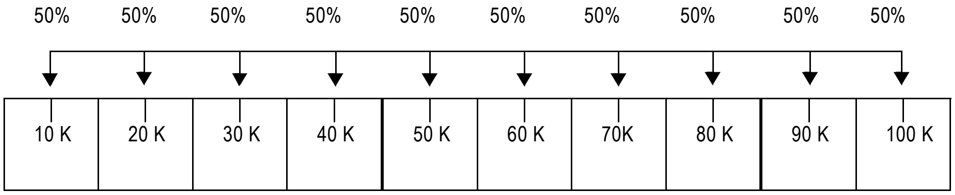


Delegating Sampling to the Database

When the Integration Service delegates sampling to the database, the database uses native algorithms to sample data. The database samples data in the same way for both automatic and manual random sampling because in both cases the database uses the same algorithm to perform the sampling function.

When the Integration Service delegates sampling to the database, the database performs a sampling operation. Once the database has sampled the data, the Integration Service reads the sampled data. For example, you have 100,000 rows of data and you sample 50 percent of the data, the database samples 50 percent of the first 10,000 rows, and 50 percent of the next 10,000 rows. The database selects an equal percentage of data samples from each part of the data source. This means that while the database selects random samples, it selects the samples equally from across the data set.

The following figure shows sampling delegated to the database:



Note: To enable database sampling, use native connections to connect to the repository database.

Improving Data Sampling Accuracy and Performance

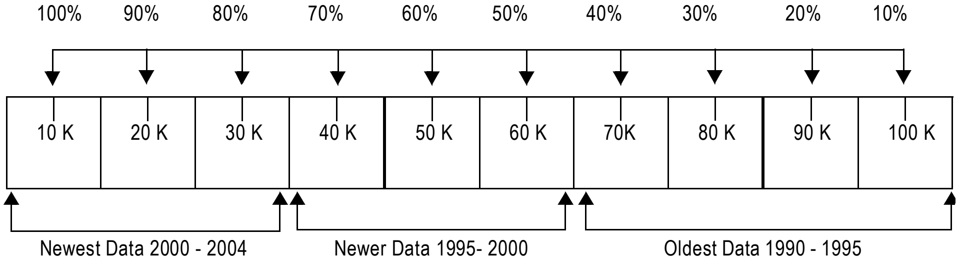
You can improve data sampling accuracy and performance when you work with the following data sources:

- Data sources with historical data
- Large data sources

Working with Historical Data

If you have historical data, you can improve data accuracy by enabling the Integration Service to delegate sampling to the database. For example, you have 100,000 rows of data, and the latest data is stored in the first 30,000 rows, the newer data is stored in the next 30,000 rows, and the oldest data is stored in the last 40,000 rows. When you select automatic random sampling and the Integration Service samples as it reads data, it samples the greatest percentage of the data from the latest data, and the least percentage of data from the oldest data. If there are great differences between the historical data and the latest data, the resulting data profile reflects the latest data more accurately than the older data.

The following table shows how Data Profiling processes an historical sample of data:



To improve data accuracy, you can ensure the Integration Service delegates sampling to the database. When you enable database sampling, the database samples data equally from across the data set.

When the Integration Service cannot delegate sampling to the database, you can improve data profiling results for historical data by using manual random sampling. Manual random sampling, like database sampling, samples a fixed percentage of data across the data set.

The following table describes recommended sampling percentages for the most accurate results with manual random sampling:

Data Size	Percentage of Data to Sample
1 million or more rows	1 percent
100,000 or more rows	10 percent
100 - 1,000 rows	Greater than 50 percent

Working with Large Data Sources

You can improve performance when you have a large source by ensuring the Integration Service delegates sampling to the database. For example, you have a data source with 100,000,000 rows, and you select automatic random sampling. When the Integration Service delegates sampling to the database, the database selects a sample of 1,000,000 rows and the Integration Service reads only 1,000,000 rows. However, if the Integration Service performs the sampling algorithm, it must first read 100,000,000 rows and then sample from those rows.

Troubleshooting

An interactive session failed due to an invalid SQL query.

The interactive session failed because an owner name for the Target Name Prefix was not specified in the session properties. However, you cannot specify the owner name for an interactive session. To specify an owner name for the session, configure a persistent session in the Workflow Manager. Specify the owner name for the Target Name Prefix in the session properties.

I tried to run an interactive session for an SAP R/3 source, but the session failed.

When you create a data profile for an SAP R/3 source, the Designer does not generate an ABAP program for the data profile. If you run an interactive session immediately after creating the data profile, the session fails. Create an ABAP program for the profile mapping and then run the session.

An interactive session failed with an error message stating that the buffer block size is too low.

You ran an interactive session for a source with a large number of rows with high precision. Or, you ran an interactive session for a multi-group source with a large number of groups. As a result, the Integration Service could not allocate enough memory blocks to hold the data, and the session failed.

Create a persistent session for the profile mapping in the Workflow Manager. In the session properties, set the value for the buffer block size that the error message in the session log recommends.

I tried to restart a persistent session from the Profile Manager and I received the following error message:

```
ERROR: Failed to fetch Workflow [id = <workflow ID>] from the repository. Please check
the session log.
```

However, when I try to locate the session log, no session log for this session run exists.

You deleted the workflow for the profile session in the Workflow Manager. When you delete a workflow in the Workflow Manager, the Integration Service cannot run the session or generate a session log. To run the profile session from the Profile Manager, reconfigure and save the session information in the Profile Wizard.

I edited session properties for a session from the Workflow Manager. However, the properties I configured in the Workflow Manager do not appear on the Session Setup page of the Profile Wizard.

You edited the profile session from the Workflow Manager while the Profile Wizard was open. To refresh the session properties in the Profile Wizard, close and reopen the Profile Wizard.

I edited session properties for a persistent session from the Profile Wizard, but the session properties I configured were not saved.

You edited the profile session from the Profile Wizard while the session was open in the Workflow Manager. Close the workflow in the Workflow Manager before you edit the session properties from the Profile Wizard.

CHAPTER 6

Viewing Profile Results

This chapter includes the following topics:

Overview

After you run a profile session, you can view the session results in a report. There are two types of Data Profiling reports:

- **PowerCenter Data Profiling reports.** Reports you can view from the Profile Manager after running a profile session. PowerCenter Data Profiling reports display data for the latest session run. Use PowerCenter Data Profiling reports to quickly view profile results during mapping development.
- **Data Analyzer Data Profiling reports.** Reports you can view from Data Analyzer after running a profile session. Data Analyzer reports provide a historic view of data. They also display information about rejected rows in the profile results. Use Data Analyzer Data Profiling reports when you want to monitor data quality during production.

PowerCenter Data Profiling Reports

After you run a profile session, you can view PowerCenter Data Profiling reports from the Profile Manager. The reports provide information based on the last profile session run. The reports display data profile information based on the functions in the data profile. Use PowerCenter Data Profiling reports to view the latest profile information about source data during mapping development.

You can view two types of PowerCenter Data Profiling reports:

- **Auto profile.** Auto profile reports contain information for the predefined functions in an auto profile. Auto profile reports provide an overall view of the profile information in a source.
- **Custom profile.** Custom profile reports contain information for the functions you defined in the custom profile. The amount of information in custom profile reports varies depending on the functions in the data profile.

The report format of PowerCenter Data Profiling reports is static. However, you can resize columns and set a default number of rows for each grid. PowerCenter stores the reports as an XML file in the following directory:

```
<PowerCenter client>\bin\Extensions\DataProfile\ProfileReports
```

PowerCenter Data Profiling reports contain information about the data profile in a report summary. It also contains profile results in the body of the report. Each section in the body of the report corresponds to a different function.

Auto Profile Reports

An auto profile report displays information about source data based on the functions in an auto profile. Depending upon which functions you included in the auto profile, the report displays information for the following functions:

- **Aggregate functions**. Calculates an aggregate value for numeric or string values in a column. Use aggregate functions to count null values, determine average values, and determine minimum or maximum values.
- **Candidate Key Evaluation**. Calculates the number and percentage of unique values in one or more source columns.
- **Distinct Value Count**. Returns the number of distinct values for the column. You can configure the auto profile to load verbose data to the Data Profiling warehouse.
- **Domain Inference**. Reads all values in a column and infers a pattern that fits the data. You can configure the Profile Wizard to filter the Domain Inference results.
- **Functional Dependency Analysis.** Determines exact and approximate dependencies between columns and column sets within a source.
- **Redundancy Evaluation**. Calculates the number of duplicate values in one or more source columns.
- **Row Count**. Counts the number of rows read from the source during the profile session. When you create a data profile that uses the Row Count function with data samples, the Row Count function estimates the total row count.

An auto profile report contains information about the data profile in a report summary. It also contains profile results in the body of the report.

The following table describes the attributes that display in the report summary of an auto profile report:

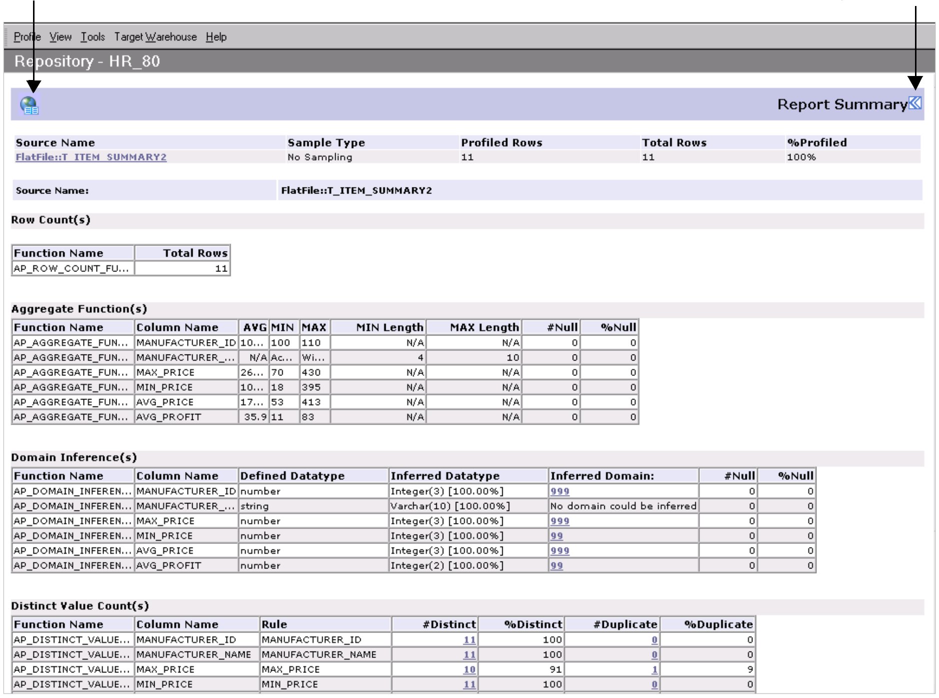
Attribute	Description
Auto Profile Name	Name of the auto profile.
Profile Description	Description of the auto profile.
Profile Run Time	Date and time of the profile session run.
Folder Name	Name of the folder in which the data profile is stored.
Repository Name	Name of the repository in which the data profile is stored.
Source	Name of the source definition on which the auto profile is based in the following formats: - `<database name>::<source definition name>` - `FlatFile::<source definition name>` - `Mapplet::<mapplet name>`
Sample Type	Type of data sampling used in the report.
Profiled Rows	Number of rows applied to the data profile.
Total Rows	Number of rows in the source. Note: If you selected Sample First N Rows, the total rows displayed in the Data Profiling report header will be the same as the number of rows selected for sampling.
% Profiled	Percentage of the total number of source rows profiled.
Groups	Groups in the source definition or mapplet on which the auto profile is based, where applicable.

The body of an auto profile report provides general and detailed information. In an auto profile report, you can click the hypertext links to view information about verbose data for the Distinct Value Count function and Domain Inference function. The verbose report shows the values of the columns defined for the column-level functions.

The following figure shows a sample auto profile report:

Click to launch the report in a browser.

Click to show report summary attributes.



Repository - HR_80

Report Summary

Source Name	Sample Type	Profiled Rows	Total Rows	%Profiled
FlatFile::T_ITEM_SUMMARY2	No Sampling	11	11	100%

Source Name:	FlatFile::T_ITEM_SUMMARY2

Row Count(s)

Function Name	Total Rows
AP_ROW_COUNT_FU...	11

Aggregate Function(s)

Function Name	Column Name	AVG	MIN	MAX	MIN Length	MAX Length	#Null	%Null
AP_AGGREGATE_FUN...	MANUFACTURER_ID	10...	100	110	N/A	N/A	0	0
AP_AGGREGATE_FUN...	MANUFACTURER_...	N/A	Ac...	Wi...	4	10	0	0
AP_AGGREGATE_FUN...	MAX_PRICE	26...	70	430	N/A	N/A	0	0
AP_AGGREGATE_FUN...	MIN_PRICE	10...	18	395	N/A	N/A	0	0
AP_AGGREGATE_FUN...	AVG_PRICE	17...	53	413	N/A	N/A	0	0
AP_AGGREGATE_FUN...	AVG_PROFIT	35.9	11	83	N/A	N/A	0	0

Domain Inference(s)

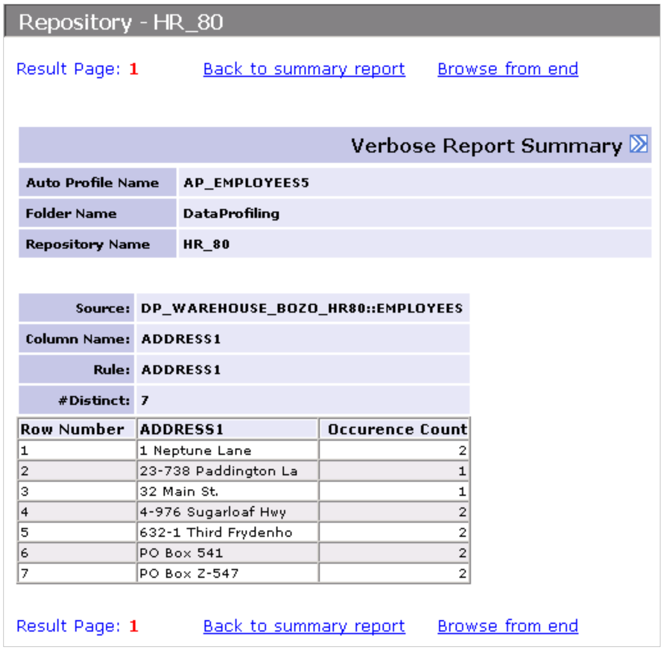
Function Name	Column Name	Defined Datatype	Inferred Datatype	Inferred Domain:	#Null	%Null
AP_DOMAIN_INFEREN...	MANUFACTURER_ID	number	Integer(3) [100.00%]	999	0	0
AP_DOMAIN_INFEREN...	MANUFACTURER_...	string	Varchar(10) [100.00%]	No domain could be inferred	0	0
AP_DOMAIN_INFEREN...	MAX_PRICE	number	Integer(3) [100.00%]	999	0	0
AP_DOMAIN_INFEREN...	MIN_PRICE	number	Integer(3) [100.00%]	99	0	0
AP_DOMAIN_INFEREN...	AVG_PRICE	number	Integer(3) [100.00%]	999	0	0
AP_DOMAIN_INFEREN...	AVG_PROFIT	number	Integer(2) [100.00%]	99	0	0

Distinct Value Count(s)

Function Name	Column Name	Rule	#Distinct	%Distinct	#Duplicate	%Duplicate
AP_DISTINCT_VALUE...	MANUFACTURER_ID	MANUFACTURER_ID	11	100	0	0
AP_DISTINCT_VALUE...	MANUFACTURER_NAME	MANUFACTURER_NAME	11	100	0	0
AP_DISTINCT_VALUE...	MAX_PRICE	MAX_PRICE	10	91	1	9
AP_DISTINCT_VALUE...	MIN_PRICE	MIN_PRICE	11	100	0	0

Click the hypertext link to view verbose summary data.

The following figure shows verbose report results:



Repository - HR_80

Result Page: 1 Back to summary report Browse from end

Verbose Report Summary

Auto Profile Name	AP_EMPLOYEES5
Folder Name	DataProfiling
Repository Name	HR_80

Source:	DP_WAREHOUSE_BOZO_HR80::EMPLOYEES
Column Name:	ADDRESS1
Rule:	ADDRESS1
#Distinct:	7

Row Number	ADDRESS1	Occurence Count
1	1 Neptune Lane	2
2	23-738 Paddington La	1
3	32 Main St.	1
4	4-976 Sugarloaf Hwy	2
5	632-1 Third Frydenho	2
6	PO Box 541	2
7	PO Box Z-547	2

Result Page: 1 Back to summary report Browse from end

Custom Profile Reports

You can view PowerCenter Data Profiling reports for custom profiles. The data the report displays depends on the functions in the data profile. The custom profile report displays data in the order of the functions you defined in the data profile.

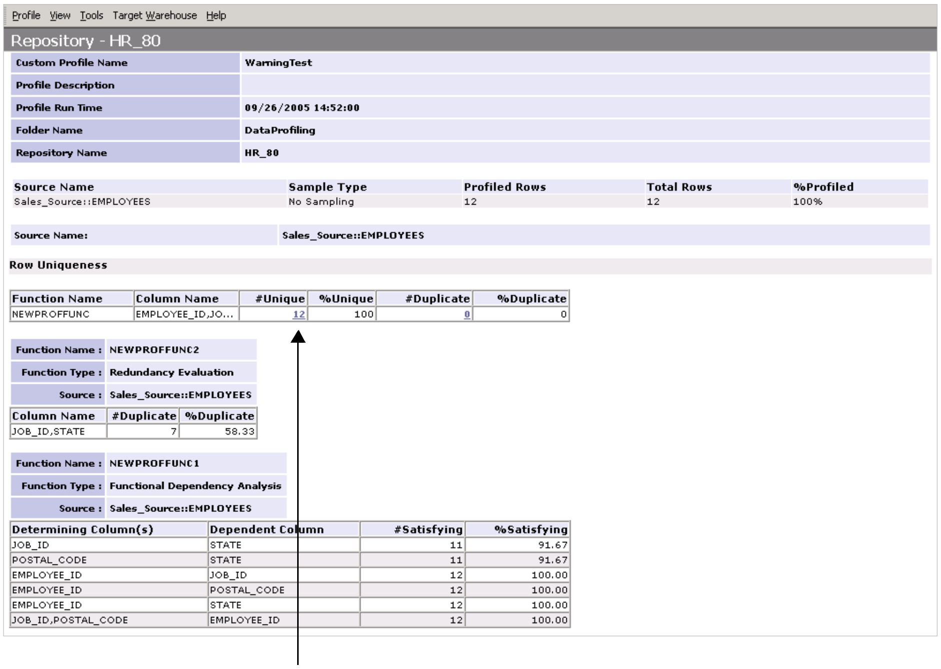
The following table describes the attributes that display in the report summary of a custom profile report:

Attribute	Description
Custom Profile Name	Name of the custom profile.
Profile Description	Description of the custom profile.
Profile Run Time	Date and time of the profile session run.
Folder Name	Name of the folder in which the data profile is stored.
Repository Name	Name of the repository in which the data profile is stored.
Source Name	Type and name of the source definitions upon which the custom profile is based in the following formats: - `<database name>::<source definition name>` - `FlatFile::<source definition name>` - `Mapplet::<mapplet name>.`
Sample Type	Type of data sampling used in the report.
Profiled Rows	Number of rows applied to the data profile.
Total Rows	Number of rows in the source. Note: If you selected Sample First N Rows, the total rows displayed in the Data Profiling report header will be the same as the number of rows selected for sampling.
% Profiled	Percentage of the total number of source rows profiled.
Show report for all sources	Click to show Data Profiling reports for all sources in the custom profile. By default, the report shows all sources. To view a particular source, click the source name on the Sampling Summary page. This filters out results for other sources.

The following table describes the attributes that display for each function in a custom profile report:

Attribute	Description
Function Name	Name of the function in the custom profile.
Function Type	Type of function.
Source	Name of the source definitions in the function.
Group By Columns	Columns selected for grouping by in the report.
Column Names	Name of the column used in the profile function.
Rule	Name of the profile business rule, where applicable.

The following figure shows a sample custom profile report:



Repository - HR_80

Custom Profile Name	WarningTest
Profile Description	
Profile Run Time	09/26/2005 14:52:00
Folder Name	DataProfiling
Repository Name	HR_80

Source Name	Sample Type	Profiled Rows	Total Rows	%Profiled
Sales_Source::EMPLOYEES	No Sampling	12	12	100%

Source Name:	Sales_Source::EMPLOYEES

Row Uniqueness

Function Name	Column Name	#Unique	%Unique	#Duplicate	%Duplicate
NEWPROFFUNC	EMPLOYEE_ID,JO...	12	100	0	0

Function Name :	NEWPROFFUNC2
Function Type :	Redundancy Evaluation
Source :	Sales_Source::EMPLOYEES

Column Name	#Duplicate	%Duplicate
JOB_ID,STATE	7	58.33

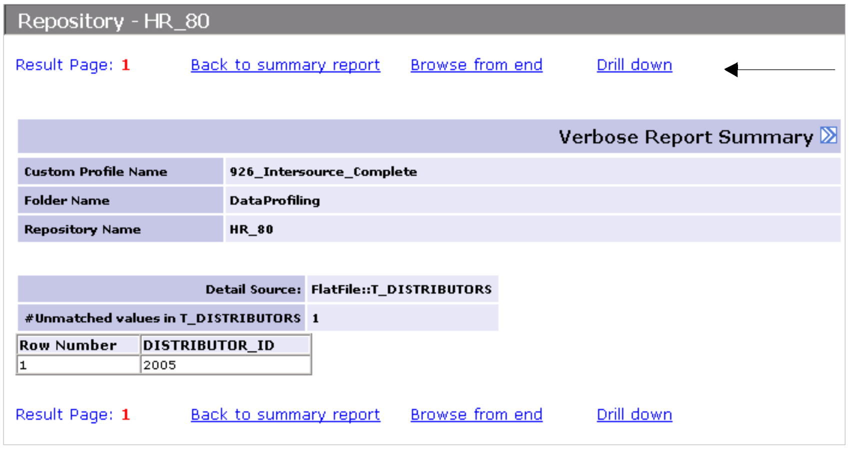
Function Name :	NEWPROFFUNC1
Function Type :	Functional Dependency Analysis
Source :	Sales_Source::EMPLOYEES

Determining Column(s)	Dependent Column	#Satisfying	%Satisfying
JOB_ID	STATE	11	91.67
POSTAL_CODE	STATE	11	91.67
EMPLOYEE_ID	JOB_ID	12	100.00
EMPLOYEE_ID	POSTAL_CODE	12	100.00
EMPLOYEE_ID	STATE	12	100.00
JOB_ID,POSTAL_CODE	EMPLOYEE_ID	12	100.00

Click hypertext link to view verbose data.

Click the hypertext links to view information about verbose data.

The following figure shows verbose report results:



Repository - HR_80

Verbose Report Summary

Custom Profile Name	926_Intersource_Complete
Folder Name	DataProfiling
Repository Name	HR_80

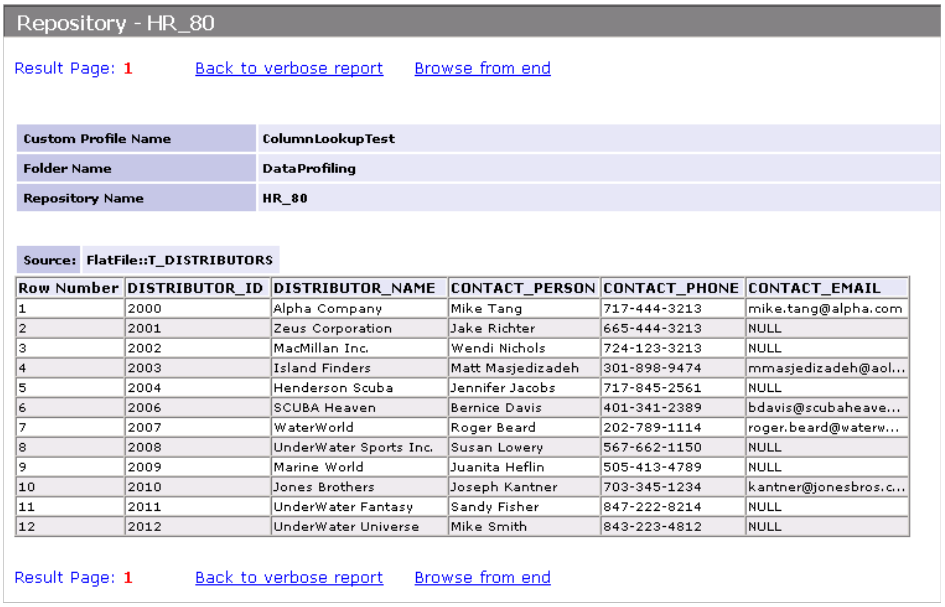
Detail Source:	FlatFile::T_DISTRIBUTORS
#Unmatched values in T_DISTRIBUTORS	1

Row Number	DISTRIBUTOR_ID
1	2005

Click to view the data.

Click Drill Down to view the verbose data that the Integration Service wrote to the Data Profiling warehouse.

The following figure shows the data from a verbose data report:



Repository - HR_80

Result Page: 1 Back to verbose report Browse from end

Custom Profile Name	ColumnLookupTest
Folder Name	DataProfiling
Repository Name	HR_80

Source: FlatFile::T_DISTRIBUTORS

Row Number	DISTRIBUTOR_ID	DISTRIBUTOR_NAME	CONTACT_PERSON	CONTACT_PHONE	CONTACT_EMAIL
1	2000	Alpha Company	Mike Tang	717-444-3213	mike.tang@alpha.com
2	2001	Zeus Corporation	Jake Richter	665-444-3213	NULL
3	2002	MacMillan Inc.	Wendi Nichols	724-123-3213	NULL
4	2003	Island Finders	Matt Masjedizadeh	301-898-9474	mmasjedizadeh@aol...
5	2004	Henderson Scuba	Jennifer Jacobs	717-845-2561	NULL
6	2006	SCUBA Heaven	Bernice Davis	401-341-2389	bdavis@scubaheave...
7	2007	WaterWorld	Roger Beard	202-789-1114	roger.beard@waterw...
8	2008	UnderWater Sports Inc.	Susan Lowery	567-662-1150	NULL
9	2009	Marine World	Juanita Heflin	505-413-4789	NULL
10	2010	Jones Brothers	Joseph Kantner	703-345-1234	kantner@jonesbros.c...
11	2011	UnderWater Fantasy	Sandy Fisher	847-222-8214	NULL
12	2012	UnderWater Universe	Mike Smith	843-223-4812	NULL

Result Page: 1 Back to verbose report Browse from end

Viewing PowerCenter Data Profiling Reports

Use the following procedure to view an auto or custom PowerCenter Data Profiling report.

To view a PowerCenter Data Profiling report:

1. Launch the Profile Manager.

   You can launch the Profile Manager from the following Designer tools:

   - **Source Analyzer.** Click Sources > Profiling > Launch Profile Manager.
   - **Mapplet Designer.** Click Mapplets > Profiling > Launch Profile Manager.

2. Select the data profile for which you want to view a report.

   You can select a data profile from the Navigator window or, you can select a data profile from the Session Status bar.

3. Click View > Report.

   If the Designer requires the connection information for the Data Profiling warehouse, the Connect to an ODBC Data Source dialog box appears. If the Designer does not require the connection information for the Data Profiling warehouse, the report appears.

4. If the Connect to an ODBC Data Source dialog box appears, select the ODBC data source for the Data Profiling warehouse.

5. Enter the user name and password for the Data Profiling warehouse.

6. Click Connect.

   The PowerCenter Data Profiling report appears.

Data Analyzer Data Profiling Reports

You can run Data Analyzer Data Profiling reports to analyze source data from profile sessions. Data Analyzer provides a PowerCenter Data Profiling Dashboard to access the Data Analyzer Data Profiling reports. Data Analyzer provides the following types of reports:

- **Composite reports.** Display the Data Analyzer Data Profiling metadata reports and summary reports. From the composite reports, you can drill down into profile metadata and summary reporting on specific functions.
- **Metadata reports.** Display basic metadata about a data profile and the source-level, column-level, and intersource functions in a data profile. They also provide historic statistics on previous runs of the same data profile.
- **Summary reports.** Display data profile results for specific source-level, column-level, and intersource functions in a data profile.

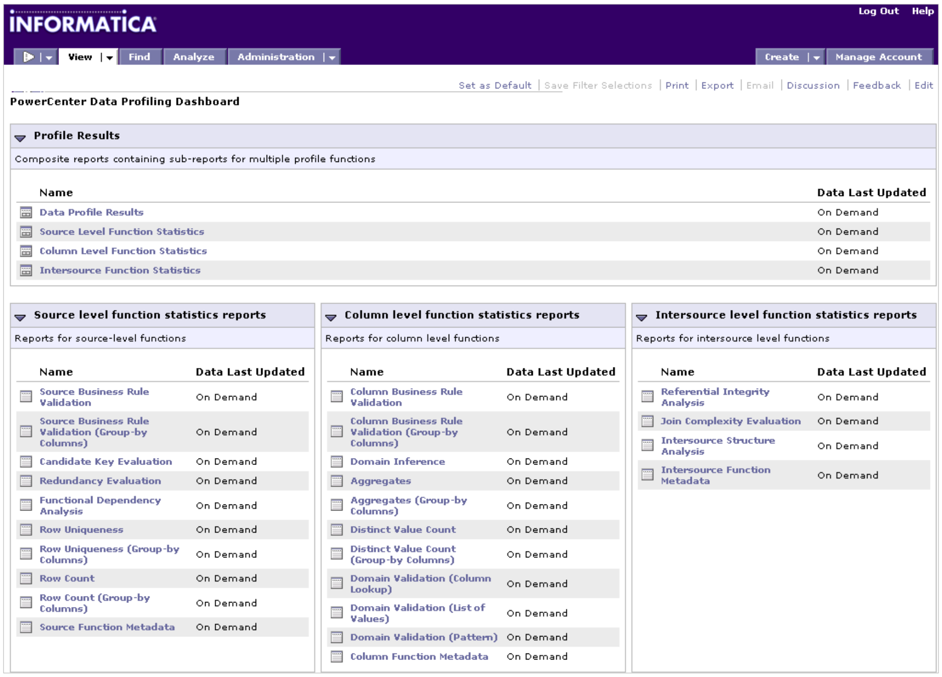
Composite Reports

The PowerCenter Data Profiling Dashboard shows the following composite reports:

- **Data Profile Results.** Displays profile statistics for all source and column-level functions.
- **Column-Level Function Statistics.** Displays profile statistics for all column-level functions.
- **Source-Level Function Statistics.** Displays profile statistics for all source-level functions.
- **Intersource Function Statistics.** Displays profile statistics for all intersource functions.

The composite reports contain the metadata reports and all of the Data Analyzer Data Profiling summary reports listed in Table 6-1 on page 63.

The following figure shows a sample PowerCenter Data Profiling Dashboard in Data Analyzer:



PowerCenter Data Profiling Dashboard

Profile Results

Composite reports containing sub-reports for multiple profile functions

Name	Data Last Updated
Data Profile Results	On Demand
Source Level Function Statistics	On Demand
Column Level Function Statistics	On Demand
Intersource Function Statistics	On Demand

Source level function statistics reports

Reports for source-level functions

Name	Data Last Updated
Source Business Rule Validation	On Demand
Source Business Rule Validation (Group-by Columns)	On Demand
Candidate Key Evaluation	On Demand
Redundancy Evaluation	On Demand
Functional Dependency Analysis	On Demand
Row Uniqueness	On Demand
Row Uniqueness (Group-by Columns)	On Demand
Row Count	On Demand
Row Count (Group-by Columns)	On Demand
Source Function Metadata	On Demand

Column level function statistics reports

Reports for column level functions

Name	Data Last Updated
Column Business Rule Validation	On Demand
Column Business Rule Validation (Group-by Columns)	On Demand
Domain Inference	On Demand
Aggregates	On Demand
Aggregates (Group-by Columns)	On Demand
Distinct Value Count	On Demand
Distinct Value Count (Group-by Columns)	On Demand
Domain Validation (Column Lookup)	On Demand
Domain Validation (List of Values)	On Demand
Domain Validation (Pattern)	On Demand
Column Function Metadata	On Demand

Intersource level function statistics reports

Reports for intersource level functions

Name	Data Last Updated
Referential Integrity Analysis	On Demand
Join Complexity Evaluation	On Demand
Intersource Structure Analysis	On Demand
Intersource Function Metadata	On Demand

From the Find tab, you can access the composite reports in the following directory:

```
Public Folders > PowerCenter Data Profiling
```

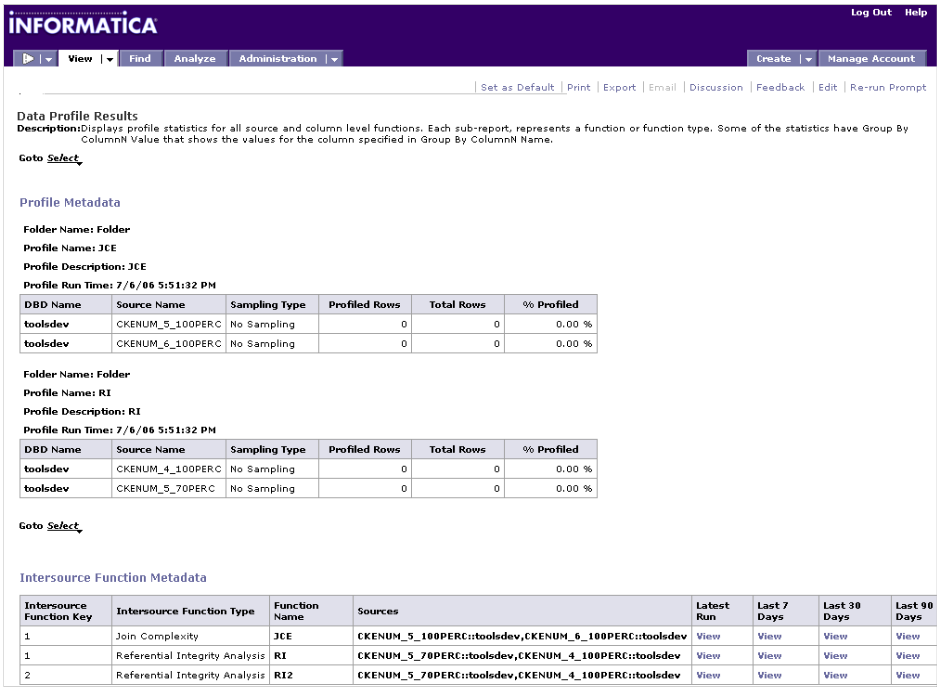
Metadata Reports

The PowerCenter Data Profiling Dashboard shows the following metadata reports:

- **Profile Metadata.** Displays the basic metadata about a data profile, including sampling metrics.
- **Column Function Metadata.** Displays metadata about column-level functions defined in a data profile. You can use this report to determine column-level functions defined in this data profile with links to historic data profile run statistics for those functions.
- **Source Function Metadata.** Displays metadata about source-level functions defined in a data profile. You can use this report to determine source-level functions defined in this data profile with links to historic data profile run statistics for those functions.
- **Intersource Function Metadata.** Displays metadata about intersource functions defined in a data profile. You can use this report to determine intersource functions defined in this profile with links to historic profile run statistics for those functions.

You can view reports for data profile information across session runs for an historic analysis of the data. Data Analyzer Data Profiling reports display profile results for the latest profile run, by default. You can also view reports with data from the most recent profile session run. This is useful during production to monitor data quality. From the metadata report, click View under Last 7 Days, Last 30 Days, or Last 90 Days to see historic data profile run statistics for the selected function type.

The following figure shows a sample Data Analyzer Data Profiling report:



Data Profile Results

Description: Displays profile statistics for all source and column level functions. Each sub-report, represents a function or function type. Some of the statistics have Group By ColumnN Value that shows the values for the column specified in Group By ColumnN Name.

Goto *Select*

Profile Metadata

Folder Name: Folder

Profile Name: JCE

Profile Description: JCE

Profile Run Time: 7/6/06 5:51:32 PM

DBD Name	Source Name	Sampling Type	Profiled Rows	Total Rows	% Profiled
toolsdev	CKENUM_5_100PERC	No Sampling	0	0	0.00 %
toolsdev	CKENUM_6_100PERC	No Sampling	0	0	0.00 %

Folder Name: Folder

Profile Name: RI

Profile Description: RI

Profile Run Time: 7/6/06 5:51:32 PM

DBD Name	Source Name	Sampling Type	Profiled Rows	Total Rows	% Profiled
toolsdev	CKENUM_4_100PERC	No Sampling	0	0	0.00 %
toolsdev	CKENUM_5_70PERC	No Sampling	0	0	0.00 %

Goto *Select*

Intersource Function Metadata

Intersource Function Key	Intersource Function Type	Function Name	Sources	Latest Run	Last 7 Days	Last 30 Days	Last 90 Days
1	Join Complexity	JCE	CKENUM_5_100PERC::toolsdev,CKENUM_6_100PERC::toolsdev	View	View	View	View
1	Referential Integrity Analysis	RI	CKENUM_5_70PERC::toolsdev,CKENUM_4_100PERC::toolsdev	View	View	View	View
2	Referential Integrity Analysis	RI2	CKENUM_5_70PERC::toolsdev,CKENUM_4_100PERC::toolsdev	View	View	View	View

From the Find tab, you can access the historic report templates in the following directory:

You can access the metadata reports in the following directory:

```
Public Folders > PowerCenter Data Profiling > Metadata Reports
```

Summary Reports

When you display a Data Analyzer Data Profiling report, you can drill down to row-level information, drill up for a summary view, and drill across data. You can also view reports for verbose data.

Data Analyzer lets you modify the Data Profiling reports to fit your business needs. You can set alerts for the Data Profiling reports to call attention to fluctuations in the source data. You can also add or change metrics in the reports or build your own reports.

For each summary report, some of the statistics have Group By ColumnN Value that shows the values for the column specified in Group By ColumnN Name. You designate a group-by column in the data profile.

From the Find tab, you can access the summary reports in the following directory:

```
Public Folders > PowerCenter Data Profiling > Summary Reports
```

Table 6-1 describes the Data Analyzer Data Profiling summary reports:

Table 6-1. Data Analyzer Data Profiling Reports

Report Name	Description
Aggregates	Displays aggregate function statistics for the selected column in a source. The Average function only applies to numeric columns. Use this report to see minimum value, maximum value, average value, and null value counts for the selected column in a source.
Aggregates (Group-by Columns)	Displays aggregate function statistics for the selected column in a source. The Average function only applies to numeric columns. The Group By ColumnN Value shows the values for the column specified in Group By ColumnN Name. Use this report to see minimum value, maximum value, average value, and null value counts for the selected column of a source. The report groups Aggregate function values by the selected group-by columns.
Candidate Key Evaluation	Displays the number and percentage of unique values for one or more columns in a source. Use this report to determine the column in a source to use as a primary key. You may want to use the column with the highest percentage of unique values as the primary key.
Column Business Rule Validation	Displays the number of rows in a single source column that satisfy a business rule and the number of rows that do not satisfy a business rule.
Column Business Rule Validation (Group-by Columns)	Displays the number of rows in a single source column that satisfy a business rule and the number of rows that do not satisfy a business rule. The Group By ColumnN Value shows the values for the column specified in Group By ColumnN Name. The report groups statistics by the selected group-by columns.
Distinct Value Count	Displays the number of distinct values and duplicate values for the source column.
Distinct Value Count (Group-by Columns)	Displays the number of distinct values and duplicate values for the source column. The Group By ColumnN Value shows the values for the column specified in Group By ColumnN Name. The report groups statistics by the selected group-by columns.
Domain Inference	Displays the domain that fits the source column data. If the Integration Service cannot infer a domain, the Data Analyzer report displays different information than the Data Profiling report in the following ways: - If the profile contains a single Domain Inference function for which no domain could be inferred, the Data Profiling report displays null count and percent of data where as the Data Analyzer report displays the message, “No domain could be inferred”. - If the profile contains multiple functions, Domain Inference or other functions, the Data Analyzer report will not display the row for the non-inferred domain.
Domain Validation (Column Lookup)	Displays number of values of the source column that fall within the specified lookup column values.
Domain Validation (List of Values)	Displays number of values of the source column that fall within the specified List of Values domain.
Domain Validation (Pattern)	Displays number of values in a single source column that satisfy a pre-defined pattern and the number of values that do not.
Functional Dependency Analysis	Displays dependencies of values of each source column on the values of another column or set of columns in the same source. Use this report to determine if the values in a column are dependent on the values of another column or set of columns.

Table 6-1. Data Analyzer Data Profiling Reports

Report Name	Description
Intersource Structure Analysis	Displays the primary key-foreign key and primary key-primary key relationships and confidence percentage between fields for multiple sources. Use this report to determine relationships between fields across multiple sources.
Join Complexity Evaluation	Displays the join complexity of associated sources. Use this report to determine join complexity between multiple sources.
Redundancy Evaluation	Displays the number of duplicate values in one or more columns in a source. Use this report to identify columns to normalize into separate tables.
Referential Integrity Analysis	Displays the number of unmatched rows after comparing columns of two different sources. Use this report to analyze orphan rows between sources.
Row Count	Displays the number of rows in a source.
Row Count (Group-by Columns)	Displays the number of rows in a source. The Group By ColumnN Value shows the values for the column specified in Group By ColumnN Name. The report groups statistics by the selected group-by columns.
Row Uniqueness	Displays the number of unique and duplicate values based on the columns selected. Use this report to identify columns to normalize into a separate table and also to test for distinct rows.
Row Uniqueness (Group-by Columns)	Displays the number of unique and duplicate values based on the columns selected. The Group By ColumnN Value shows the values for the column specified in Group By ColumnN Name. The report groups statistics by the selected group-by columns.
Source Business Rule Validation	Displays the number of rows for one or more columns in a source that satisfy a business rule and the number of rows that do not.
Source Business Rule Validation (Group-by Columns)	Displays the number of rows in a source that satisfy a business rule and the number of rows that do not. The Group By ColumnN Value shows the values for the column specified in Group By ColumnN Name. The report groups statistics by the selected group-by columns.

Historic Reports Templates

You can access Data Profiling historic report templates through Data Analyzer. The historic report templates folder contains the reports from the Summary Reports folder and provide profile session results for functions for all profile session runs.

From the Find tab, you can access historic report templates in the following directory:

```
Public Folders > PowerCenter Data Profiling > Report Templates > Historic Report
Templates
```

Purging the Data Profiling Warehouse

To manage the Data Profiling warehouse, you can periodically purge session results and metadata you no longer need. When you delete data profiles from a repository, the related metadata and profile session results remain in the Data Profiling warehouse. You can purge metadata and profile session results from the Data Profiling warehouse for deleted data profiles. You can also purge profile session results that are no longer associated with data profile metadata.

When you purge the Data Profiling warehouse, you can purge all metadata and session results, the most recent metadata and session results, or metadata data and session results based on a date. You can purge metadata and session results associated with a particular repository folder or associated with all folders in the repository.

To purge the Data Profiling warehouse:

1. Launch the Profile Manager.

   You can launch the Profile Manager from the following Designer tools:

   - **Source Analyzer.** Click Sources > Profiling > Launch Profile Manager.
   - **Mapplet Designer.** Click Mapplets > Profiling > Launch Profile Manager.

2. Click Target Warehouse > Connect to connect to the Data Profiling warehouse.
3. Enter the Data Profiling warehouse connection information and click Connect.
4. Click Target Warehouse > Purge.

   The Purge Data Profiling Warehouse dialog box appears.

5. Select the folder containing the metadata and/or profile session results you want to purge.

   The folder must be open to purge it. Or, select All Open Folders to purge metadata and profile session results.

6. Select one of the following Purge Types to purge from the Data Profiling warehouse:

   - **Profiles.** Purges metadata and profile session results for all profiles that satisfy the selected criteria.
   - **Orphan Profiles Only.** Purges profile metadata and profile session results that correspond to profiles that have been deleted from the repository.

7. If you select Profiles, select Purge Profile Metadata and Runs or Purge Profile Runs Only.

   If you select Orphan Profiles Only, go to step 9.

8. Select one of the following options:

   - **All.** Purges all profile metadata and session run results for all data profiles.
   - **All Except Latest.** Purges all profile metadata and profile session results, except those corresponding to the latest run of profiles.
   - **Older Than.** Purges profile metadata and profile session results for all profiles that have been run prior to the date you specify.

9. Select Purge Metadata and Runs for Orphan Profiles to delete both metadata and session results that are no longer associated with data profile metadata.
10. Click Purge.

The Data Profiling warehouse may take a few minutes to purge.

APPENDIX A

Code Page Compatibility

This appendix includes the following topic:

Code Page Compatibility

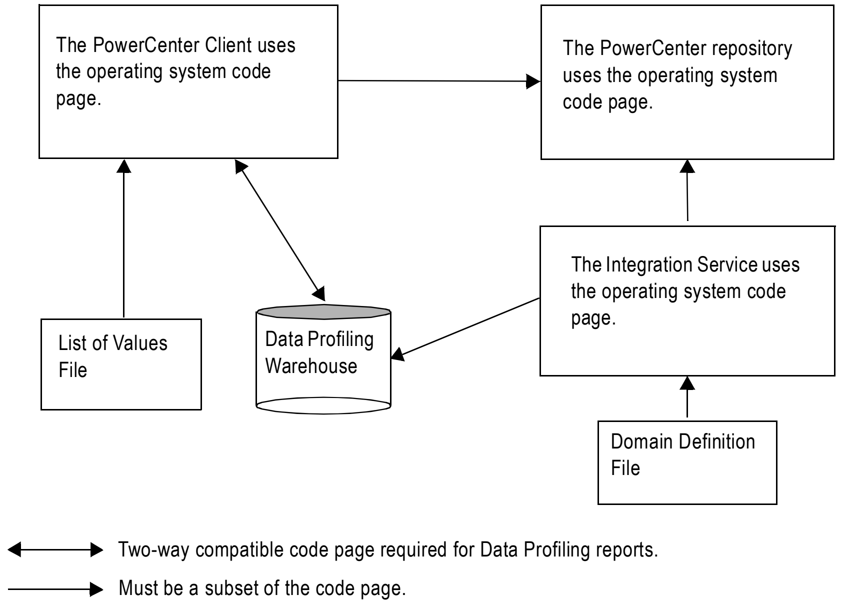
When you use Data Profiling, configure code page compatibility between all PowerCenter components and Data Profiling components and domains.

Follow the instructions in “Understanding Globalization” in the *PowerCenter Administrator Guide* to ensure that the code pages of each PowerCenter component have the correct relationship with each other.

When you work with data profiles, verify that the code pages for the Data Profiling components have the correct relationship with each other:

- The Integration Service must use a code page that is a subset of the Data Profiling warehouse code page.
- The source code page must be subset of the Data Profiling warehouse code page.
- The code page for the PowerCenter Client must be a subset of the repository code page.
- A Domain Definition File Name domain must use a code page that is a subset of the Integration Service code page.
- A List of Values domain must use a code page that is a subset of the code page for the operating system that hosts the PowerCenter Client.
- To view reports from the Profile Manager, the Data Profiling warehouse must use a code page that is one-way compatible with the code page of the operating system that hosts the PowerCenter Client.

The following figure shows code page compatibility requirements for Data Profiling:


The PowerCenter Client uses the operating system code page.
The PowerCenter repository uses the operating system code page.
The Integration Service uses the operating system code page.
List of Values File
Data Profiling Warehouse
Domain Definition File
Two-way compatible code page required for Data Profiling reports.
Must be a subset of the code page.

Appendix B

Glossary

This appendix includes the following topic:

- Glossary of Terms, 69

Glossary of Terms

Aggregate functions

Functions that calculate an aggregate value for a numeric or string value applied to one column of a profile source.

approximate dependency

A functional dependency between one column and another column or column sets where the values satisfy the Functional Dependency Analysis function for all rows except a defined percentage. See also *functional dependency* on page 71.

auto profile

A data profile containing a predefined set of functions for profiling source data.

auto profile report

A PowerCenter Data Profiling report that displays information about source data based on the functions in an auto profile.

Average Value aggregate function

A column-level aggregate function that calculates an aggregate value for the average value of the rows in the source column.

Business Rule Validation column-level function

A column-level function that calculates the number of rows in a single source column that satisfy a business rule and the number of rows that do not.

Business Rule Validation source-level function

A source-level function that calculates the number of rows for one or more columns in a source that satisfy a business rule and the number of rows for one or more columns in a source that do not.

Candidate Key Evaluation function

A source-level function that calculates the number of unique values in one or more columns in a source.

column-level functions

Functions that perform calculations on one column of a source, source group, or mapplet group.

Column Lookup domain

A domain defined by the values in a column of a relational or flat file source. See also *domain* on page 70.

confidence measure

The acceptable percentage of accuracy defined in the Intersource Structure Analysis function. See also *Intersource Structure Analysis function* on page 71.

custom domain

A domain you create to validate source values from source data. You can create a custom domain when you create a Domain Validation function. See also *domain* on page 70, and *Domain Validation function* on page 71.

custom profile

A data profile for which you define functions to profile source data.

custom profile report

A PowerCenter Data Profiling report that displays information about source data based on the functions in a custom profile.

data profile

A profile of source data in PowerCenter. A data profile contains functions that perform calculations on the source data.

Data Profiling views

Views to create metrics and attributes in the Data Profiling schema. PowerCenter Data Profiling reports use these metrics and attributes. You can also use these views to create reports in a business intelligence tool.

Data Profiling warehouse

A relational database that stores data profile results from profile sessions.

Distinct Value Count function

A column-level function that returns the number of distinct values for a column.

domain

A set of all valid values for a source column. A domain can contain a regular expression, list of values, the name of a file that contains a list of values or point to a column of relational or a flat-file source. PowerCenter provides prepackaged domains. You can also create custom domains.

Domain Definition File Name domain

Domains defined by an external file containing a list of values. See also *domain* on page 70.

Domain Inference function

A column-level function that reads all values in the column and infers a pattern that fits the data. The function determines if the values fit a list of values derived from the column values or a pattern that describes the pattern of the source data.

Domain Validation function

A column-level function that calculates the number of values in the data profile source column that fall within a specified domain and the number of values that do not. The Domain Validation function requires that you specify a domain.

exact dependency

A dependency in which all values in a column have a functional dependency on values in another column or column set. See also *functional dependency* on page 71.

functional dependency

A dependency in which you can infer the values in one column by the relationship to values in another column or column set. See also *approximate dependency* on page 69, *exact dependency* on page 71, and *Functional Dependency Analysis function* on page 71.

Functional Dependency Analysis function

A source-level function that determines exact and approximate dependencies between one column and another column or column sets. When functional dependencies exist in a source, you can infer the values in one column by relationships to values in another column or column set. See also *approximate dependency* on page 69 and *exact dependency* on page 71.

group-by columns

Columns by which you want to group data for a custom profile. When you configure a function, you can determine the column by which you want to group the data.

interactive session

A profile session that you run from the Profile Manager.

intersource functions

Functions that perform calculations on two or more sources, source groups, or mapplet groups.

Intersource Structure Analysis function

An intersource function that determines primary key-foreign key and primary key-primary key relationships between sources by applying a confidence measure and identifying only those relationships that have confidence equal to or greater than the defined confidence measure. See also *confidence measure* on page 70.

Join Complexity Evaluation function

An intersource function that determines column values which satisfy join conditions.

List of Values domain

Domains defined by a list of values, either entered manually or selected from a column in a flat file or relational source.

Maximum Value aggregate function

A column-level aggregate function that calculates an aggregate value for the maximum value of rows in the source column.

Minimum Value aggregate function

A column-level aggregate function that calculates an aggregate value for the minimum value of rows in the source column.

non-reusable domain

A domain that applies to one Domain Validation function. See also *reusable domain* on page 72.

NULL Value Count aggregate function

A column-level aggregate function that calculates an aggregate value for the number of rows with NULL values in the source column.

persistent session

A session stored in the repository that you run from the Profile Manager or the Workflow Manager. Use a persistent session to run a profile mapping more than once.

PowerCenter Data Profiling reports

Reports you can view from the Profile Manager after running a profile session. PowerCenter reports provide information based on the last profile session run.

prepackaged domains

Domains PowerCenter provides with Data Profiling, which verify data, such as phone numbers, postal codes, and email addresses. See also *domain* on page 70.

Profile Manager

A tool in the Designer that manages data profiles. Use the Profile Manager to set default data profile options, work with data profiles in the repository, run profile sessions, view profile results, and view sources and mapplets with at least one data profile defined for them.

profile mapping

A mapping the Designer generates when you create a data profile. The PowerCenter repository stores the data profile and the associated mapping.

profile session

A session for a profile mapping that gathers information about source data. The Data Profiling warehouse stores the results of profile sessions. See also *persistent session* on page 72 and *interactive session* on page 71.

Redundancy Evaluation function

A source-level function that calculates the number of duplicate values in one or more columns in the source.

Referential Integrity Analysis function

An intersource function that determines the orphan values in two sources.

Regular Expression domain

A domain defined by a range of values in an expression.

reusable domain

A domain you can apply to multiple Domain Validation functions in one or more data profiles. See also *domain* on page 70.

Row Count function

A source-level function that counts the number of rows read from the source during a profile session.

Row Uniqueness function

A source-level function that calculates the number of unique and duplicate values based on the columns selected. You can profile all columns in the source row or choose individual columns to profile.

source-level function

A function that performs calculations on two or more columns in a source, source group, or mapplet group.

temporary session

A session that is run from the Profile Manager and is not stored to the repository.

verbose mode

An option to view verbose data that the Integration Service writes to the Data Profiling warehouse during a profile session. You can specify the type of verbose data to load when you configure the data profile.

INDEX

NOTICES