Data Profiling in SQL Server

Sandesh Nagaraj

Sandesh.mys@gmail.com

Session Level: Beginner

About Me

- 11 years of IT experience
- Worked as Programmer/Database Developer/BI Developer/Technical Specialist/BI Architect
- MCITP- BI SQL Server 2005/2008
- MCSE- BI SQL Server 2012 (currently busy)
- Certified Data Architect
- LinkedIn: https://za.linkedin.com/in/sandeshnagaraj

Agenda

- Need for Data Profiling
- Definition
- Attribute Analysis
- Relationship Analysis
- SQL Server Offering
- Demo

Need for Data Profiling

- When you work with data on a daily basis, it's very common to explore it.
- To work effectively with the data, you need to understand its profile—Both at the detail level and from a higher aggregated level.
- In data warehousing, it is common to profile the data in advance to identify patterns in the data and determine if there are an data quality issues.

Definition

- **Data profiling** is a necessary precursor for designing any kind of system that uses data.
- **Data profiling** is like running a process to return metrics from a data set.
- Employs analytic methods for looking at data for the purpose of developing a thorough understanding of the content, structure and quality of the data.
- A good data profiling tools can process very large amounts of data and with skills of the analyst, uncover all sorts of issues that need to be addressed

Attribute Analysis

It consists of looking at the data in an individual column and abstracting out a high-level view of the following:

- **Range / Summation**
  - Minimum / Maximum
  - Mean
  - Median
- **Completeness**
  - Portion of the records populated with data
  - Portion of records that are blank
  - Portion of records that are null
- **Uniqueness**
  - Cardinality – counts the number of unique values for a given attribute
  - Frequency distribution – provides a count for each unique value in a given attribute
  - Duplicate data identification
- **Format**
  - Inferred pattern identification – identifies the different unique data formats
  - Inferred pattern frequency – provides a count of each of the different formats
- **Inferred type** – alpha, numeric, date, binary, etc.

Relationship Analysis

It consists of looking at the data and how columns and records relate to one another

- **Structural Integrity**
  - Unique primary keys
  - Foreign keys
  - Foreign key parents
  - Normalized or denormalized table structure
- **Discovery**
  - Functional dependencies
  - Potential primary keys
  - Potential foreign keys
  - % agreement
  - Duplicate data columns
  - Orphan analysis

SQL Server offering

- The data profiling tools in SQL Server include a **Data Profiling task** (in SSIS) and a **Data Profile Viewer**.
- The **Data Profiling task** is a new task that was introduced in SSIS 2008. It helps in understanding large sets of data by offering a set of commonly needed data profiling options.
- The **Data Profile Viewer** is an application that can be used to review the output of the Data Profiling task.

Data Profiling Task

- The Data Profiling task was introduced in SSIS 2008. It has the capability to create multiple types of data profiles across any target tables/Views you specify.
- Using it can be as simple as adding it to a package, selecting a SQL Server table/view, and picking one or more profiles to run against that table/view.
- Each profile returns a different set of information about the data in the target table/view.

Types of profiles

The Data Profiling task supports eight different profiles.

- Five of these looks at individual columns in a table.
  - THE COLUMN LENGTH DISTRIBUTION PROFILE
  - THE COLUMN NULL RATIO PROFILE
  - THE COLUMN PATTERN PROFILE
  - THE COLUMN VALUE DISTRIBUTION PROFILE
  - THE COLUMN STATISTICS PROFILE
- Two of them looks at multiple columns in the same table.
  - THE CANDIDATE KEY PROFILE
  - THE FUNCTIONAL DEPENDENCY PROFILE
- One looks at columns across two tables
  - THE VALUE INCLUSION PROFILE

Profiles that analyze individual columns

Profiles	Description	Valid Data Types*	Example
Column Length Distribution Profile	Reports all the distinct lengths of string values in the selected column and the percentage of rows in the table that each length represents.	char type	you profile a column of United States state codes that should be two characters and discover values longer than two characters.
Column Null Ratio Profile	Reports the percentage of null values in the selected column.	All columns**	you profile a Zip Code/Postal Code column and discover an unacceptably high percentage of missing codes.
Column Pattern Profile	Reports a set of regular expressions that cover the specified percentage of values in a string column.	char type	This profile can also suggest regular expressions that can be used in the future to validate new values. For example, a pattern profile of a United States Zip Code column might produce the regular expressions: \d{5}-\d{4}, \d{5}, and \d{9}.
Column Statistics Profile	Reports statistics, such as minimum, maximum, average, and standard deviation for numeric columns, and minimum and maximum for datetime columns.	Numeric or datetime types (no mean and stddev for datetime column)	you profile a column of historical dates and discover a maximum date that is in the future.
Column Value Distribution Profile	Reports all the distinct values in the selected column and the percentage of rows in the table that each value represents. Can also report values that represent more than a specified percentage of rows in the table.	integer, char and datetime types	you profile a column that is supposed to contain states in the United States and discover more than 50 distinct values.

Profiles that analyze multiple columns

Profiles	Description	Valid Data Types*	Example
Candidate Key Profile	Reports whether a column or set of columns is a key, or an approximate key, for the selected table.	integer, char and datetime types	duplicate values in a potential key column.
Functional Dependency Profile	Reports the extent to which the values in one column (the dependent column) depend on the values in another column or set of columns (the determinant column).	integer, char and datetime types	you profile the dependency between a column that contains United States Zip Codes and a column that contains states in the United States. The same Zip Code should always have the same state, but the profile discovers violations of this dependency.
Value Inclusion Profile	Computes the overlap in the values between two columns or sets of columns. This profile can determine whether a column or set of columns is appropriate to serve as a foreign key between the selected tables.	integer, char and datetime types	you profile the ProductID column of a Sales table and discover that the column contains values that are not found in the ProductID column of the Products table.

Demo

Limitations

- Data Profiling task requires that the data to be profiled be in SQL Server 2000 or later. i.e. can not directly profile data inside Oracle, Access, Excel, or flat files.
- The Data Profiling task also requires that you use an ADO.NET connection manager. Typically, in SSIS, OLE DB connection managers are used, as they tend to perform better. This may mean creating two connection managers to the same database, if you need to both profile data and import it in the same package.
- Data Profile Viewer does require a SQL Server installation, because the viewer is not packaged or licensed as a redistributable component. It is possible to transform the XML output into a more user-friendly format by using XSL Transformations (XSLT) to translate it into HTML, or to write your own viewer for the information.
- The task's performance can vary greatly, depending both on the volume of data you are profiling and on the types of profiles you have requested. Some profiles, such as the Column Pattern profile, are resource intensive and can take quite a while on a large table.

Resource and References

- Microsoft Sites:
  - TechNet: Books Online for SQL Server
- Forums
  - MSDN: SQL Server Integration Services.
- Books:
  - SQL Server MVP Deep Dives - **Chapter 56**
  - SQL Server 2012 Data Integration Recipes: Solutions for Integration Services and Other ETL Tools - **chapter 10**

Q & A

Thank You