Data Profiling

in

ETL

Table of Contents

[**Data Profiling** 3](#_Toc126868716)

[**What is Data Profiling in ETL?** 3](#_Toc126868717)

[***Types of Data Profiling:*** 4](#_Toc126868718)

[***Techniques of Data Profiling*** 5](#_Toc126868719)

[Sample example 6](#_Toc126868720)

[Sample Run/Output 7](#_Toc126868721)

# **Data Profiling**

* Data Profiling can be defined as the process of examining and analysing data to create valuable summaries of it. The process yields a high-level overview that aids in
* discovering data quality issues,
* risks, and
* overall trends.

# **What is Data Profiling in ETL?**

* Data Profiling in ETL is a detailed analysis of source data. It tries to understand the structure, quality and content of source data and its relationships with other data.
* It takes place during the Extract, Transform and Load (ETL) process and helps organizations find the right data for projects.

Organizations use data profiling in ETL in the following scenarios:

* A requirement to automate the data profiling process.
* Reducing human errors during data profiling.
* The need for high-quality and consistent data.
* Identifying problems during data profiling.

Data profiling in ETL is a prerequisite to data analysis. As this process improves the structure, quality, and content of source data, users execute better data analysis and generate valuable intelligence about their organization.

When data profiling in ETL, organizations discover whether data is:

* Unique
* Incomplete
* Corrupted
* Duplicated

Organizations then identify patterns and correlations in data and start to generate insights.

## ***Types of Data Profiling:***

* Structure discovery helps determine whether your data is consistent and formatted correctly. It uses basic statistics to provide information about the validity of data (e.g., sum, maximum). Structure discovery helps understand how well data is structured — for example, what percentage of pin code does not have the correct number of digits.
* Relationship discovery identifies connections between different data sets. Discovering how parts of the data are interrelated.  For example, if a telephone number is incorrectly formatted it could mean that certain customers can’t be reached.
* Content discovery focuses on data quality. Data needs to be formatted, standardized, and properly integrated with existing data in a timely and efficient manner. For example, key relationships between database tables, references between cells or tables in a spreadsheet. Understanding relationships is crucial to reusing data. Related data sources should be imported by preserving relationships.

## 

## ***Techniques of Data Profiling***

1. Column profiling

Column profiling scans through a table and counts the number of times each value shows up within each column. This method can be useful to find frequency distribution and patterns within a column of data.

2. Cross-column profiling

Cross-column profiling is made up of two processes: key analysis and dependency analysis. Key analysis examines collections of attribute values by scouting for a possible primary key. Dependency analysis is a more complex process that determines whether there are relationships or structures embedded in a data set. Both techniques help analyse dependencies among data attributes within the same table.

3. Cross-table profiling

Cross-table profiling uses foreign key analysis, which is the identification of orphaned records and determination of semantic and syntactic differences, to examine the relationships of column sets in different tables. This can help cut down on redundancy but also identify data value sets that could be mapped together.

4. Data rule validation

Data rule validation uses data profiling in a proactive manner to verify that data instances and data sets conform to predefined rules. This process can be achieved either through batch validation or ongoing validation service.

# Sample example

**Overview**

The objective of this utility is to provide a pluggable solution in PySpark to easily profile your data while measuring its quality.

**Process Flow**

**Timeline

Description automatically generated**

# Sample Run/Output

**Incoming Data Format**

Table

Description automatically generated

**NULL/Empty Value Checks**

Table

Description automatically generated

**Summary of Numeric Columns**

Table

Description automatically generated

**Distinct Values for Columns of Interest**

**A picture containing timeline

Description automatically generated**

**Distribution for Aggregate Columns**

**Table

Description automatically generated with medium confidence**

**Data Quality Mismatch Percentage**

**Table

Description automatically generated with medium confidence**

**Source Code**

The entire source code for the above implementation is available at

GitHub: [<https://github.com/hari328373/data_profiling>](https://github.com/hari328373/data_profiling)

**Reference Links:**

1.[Data Analytics | Data Profiling | Use case study: Investment Data | by Manoj Bidadi Raju | Towards Data Science](https://towardsdatascience.com/data-analytics-data-profiling-use-case-study-investment-data-adf872152db6)

2. <https://www.integrate.io/glossary/what-is-data-profiling-in-etl/#:~:text=Data%20profiling%20in%20ETL%20is,the%20right%20data%20for%20projects>.