The Practitioner's Guide to Data Profiling

A DataFlux White Paper

Prepared by David Loshin

DataFlux
A SAS COMPANY

YOUR DATA.
YOUR BUSINESS.
ONE SOLUTION.

DataFlux
A SAS COMPANY

Introduction

Data profiling is everywhere – its importance as a key technology is difficult to understate, as it has become the *de facto* starting point for practically any data-oriented project or program. As data profiling has emerged as a critical technical commodity, it underpins a variety of information management programs, including data quality assessment, data quality validation, metadata management, data integration and Extraction, Transformation, and Loading (ETL) processing, data migrations, and modernization projects.

Data profiling enables a suite of analysis and assessment algorithms that, when applied in the proper context, provides hard evidence for potential issues that exist within a data set. This paper considers the techniques used by data profiling tools, the ways the analyses are performed, and how those analyses can yield value in a number of application contexts, including reverse engineering, assessment for potential anomalies, business rule validation, and validation of metadata and data models.

The power of data profiling lies in the ability to summarize details about large data sets from many different angles. Yet the lure of undirected data analysis can drive the analyst on an almost never-ending quest to find yet one more undiscovered piece of knowledge. The value of data profiling rests in the ability to couple the capabilities of the technical tool with the knowledge of how that technology is applied in support of a program's goals. In turn, understanding how these techniques work and how they are used provides additional insight into ways that profiling can be applied.

In this paper, we consider four methods used by data profiling tools for analysis:

- Column profiling
- Cross-column profiling
- Cross-table profiling
- Data rule validation

Given these techniques, analysts apply profiling to a number of operational processes and application contexts. In the paper we will look at these:

- Reverse engineering
- Anomaly analysis
- Metadata compliance

Finally, the paper considers some key characteristics of data profiling products useful for supporting business objectives.[1]

Data Profiling Methods

In general, there are four methods employed by data profiling tools:

- Column profiling, which provides statistical measurements associated with the frequency distribution of data values (and patterns) within a single column (or data attribute)
- Cross-column profiling, which analyzes dependencies among sets of data attributes within the same table
- Cross-table profiling, which evaluates relationships and intersections between tables
- Data rule validation, which verifies the conformance of data instances and data sets with predefined rules

Column Profiling

From a computational perspective, column profiling is relatively straightforward. The process typically scans through the entire table and counts the number of occurrences of each value within each column. Technically, this can be implemented using hash tables that map each value that appears in the column to the number of times that value appears.

Despite its simplicity, the frequency distribution of column values exposes some insightful characteristics, and enables a number of interesting analyses, including those shown in Table 1. The frequency analysis provides summarization/aggregate values as well as descriptive characteristics.

Analysis Method	*Description*
Range Analysis	Scans values to determine whether they are subject to a total ordering, and determines whether the values are constrained within a well-defined range
Sparseness	Evaluates the percentage of the elements that are not populated
Cardinality	Analyzes the number of distinct values that appear

[1] This paper is based on material from David Loshin's book, "The Practitioner's Guide to Data Quality Improvement," Morgan Kaufmann, 2010.

	within the column
Uniqueness	Indicates if each of the values assigned to the attribute is unique
Value Distribution	Presents an ordering of the relative frequency (count and percentage) of the assignment of distinct values
Value Absence	Identifies the appearance and count of occurrences of null values
Type Determination	Characterizes data type and size
Abstract Type Recognition	Refines the semantic data type association with a specific attribute, often depending on pattern analysis
Overloading	Attempts to determine if an attribute is being used for multiple purposes
Format Evaluation	Attempts to resolve unrecognized data into defined formats
Minimum value	Provide the minimum value based on the ordering properties of the data set
Maximum value	Provide the maximum value based on the ordering properties of the data set
Mean	Provide the average value (for numeric data);
Median	Provide the middle value (if such a thing is defined);
Standard deviation	Mostly relevant only for numeric value sets

Table 1: Column profiling methods

Frequency analysis can suggest potential areas for further investigation. For example, low frequency values ("outliers") are potential data quality violations, and a small number of very high frequency values may indicate a flag or code attribute.

Cross-Column Profiling

Cross-column analysis is much more computationally intensive than column analysis. There are two main aspects to cross-column profiling:

- Key analysis – Older data systems are often susceptible to the absence of an artificially created primary key. However, there is still a need for unique differentiation among each of the entities stored within a table. Key analysis examines collections of attribute values across each record to determine candidate primary keys.

- Dependency analysis – Functional dependency analysis provides an ability to determine if there are embedded relationships or embedded structures within a data set. A functional dependency between column X and column Y says that, given any two records, if the corresponding values of column X are the same, then the corresponding values of column Y will be the same. This implies that the value of column X *determines* the value of column Y, or that column Y is said to be *dependent* on column X. This abstraction can be extended to sets of columns as well. Functional dependency analysis can be used for identification of redundant data, mapped values, and helps to suggest opportunities for data normalization.

The algorithms that profiling tools use for cross-column analysis require multiple scans through the table, with intermediate results managed within complex data structures. The results of each algorithmic iteration must be cached to support each subsequent processing stage.

Cross-Table Profiling

Cross-table profiling iteratively reviews how the values within column sets in different tables potentially intersect and overlap. Reviewing the cardinality of the columns and the degree to which column data sets overlap across tables suggests dependencies across tables, relationships, and redundant storage, as well as opportunities for identification of data value sets that are mapped together. Some key capabilities for cross-table profiling include:

- Foreign key analysis, which seeks to map key relationships between tables

- Identification of orphaned records, indicative of a foreign key relationship that is violated because a child entry exists where a parent record does not

- Determination of semantic and syntactic differences, such as when differently named columns hold the same values, or same-named columns hold different values

The redundancy analysis provided by cross-table profiling is straightforward, in that the values in each of a pair of columns respectively selected from different tables are evaluated for set intersection. Conclusions drawn from the analyst's review provide the insight: depending on the size of the sets and the degree to which the sets overlap, two columns might be tagged as foreign keys (if the value set is large and mostly disparate) or perhaps as both taking their value from the same reference domain (if the value set is small).

Data Rule Validation

The first three analytical methods are often used for the undirected discovery of anomalies in data sets. A fourth type of analysis uses data profiling in a proactive manner to validate defined assertions. Most data profiling tools allow the analyst to define expression-based data rules. For example, a data rule for a customer table might assert that the *birth date* field must not be missing a value.

These rules may exhibit assertions associated with a high degree of precision by looking at specific data attributes, a medium degree of precision by focusing on assertions and expressions at the data instance or record level, or define rules that must apply across a collection of records or over an entire tables, or across multiple tables.

Data rule validation can be performed in two different ways:

- Batch validation, in which an entire data set can be subjected to validation of a collection of rules at the same time
- Validation service, in which specific data instances can be submitted by the application to the service for validation

The main difference is in supporting the different levels of granularity. In the batch approach, the entire data set to be evaluated is presented to the profiler, and therefore rules of all levels of granularity can be applied. In the service mode, the types of rules to be validated are limited by the scope of the set of data instances provided. Since the entire data set is usually not provided, the rules are limited to column and data instance rules.

Application Contexts

With an understanding of the capabilities of a data profiling tools and aspects of the underlying algorithms, we can examine the ways that data profiling is used in specific application contexts.

Reverse Engineering

The absence of documented knowledge about a data set (which drives the need for anomaly analysis) also drives the desire for deeper knowledge of the business terms, related data elements, their definitions, the reference data sets used, and structure of the attributes in the data set – in other words, its metadata. Reverse engineering in the data world is used to review the structure of a data set for which there is little or no existing metadata, or for which there are reasons to suspect the completeness or quality of existing metadata. The results of the review help to discover, document and organize the "ground-truth" regarding the data set's metadata.

Here, the results of data profiling are used to incrementally capture a knowledge base associated with data element structure, semantics and use. Column values are analyzed to:

- Determine if there are commonly-used value domains
- Determine if discovered data domains map to known conceptual value domains
- Review the size and types of each data element
- Identify any embedded pattern structures associated with any data element
- Identify keys and how those keys are used to refer to other data entities

The results of this reverse engineering process can populate a metadata repository. The discovered metadata supplements development activities such as business process renovation, enterprise data architecture, master data management, services-oriented architecture, data warehouse development or data migration.

Every structured data collection has an underlying data model, whether it is explicitly defined or not. Many older systems built prior to the use of modern data modeling tools have an underlying model, even though it may not conform to today's expectations for comprehensive entity-relationship diagrams. At some point, either when the data is to be aggregated into a data warehouse or when the application is migrated to an updated architecture, the analysts will need a better understanding of that model. The goals of data reverse engineering include identifying the system's information components, reviewing the relationships between those components, and identifying any embedded object hierarchies with the intention of identifying opportunities for model improvement.

Domain Discovery and Analysis

This process helps the analyst identify sets of values that logically belong together with respect to a well-defined business meaning. Some domain discovery activities include:

- Identifying enumerated domains and reference data - This is the process of identifying a set of values that are the source data set for one or more table attributes. Enumerated domains are typically character string value sets with a limited number of entries. Examples of enumerated domains are codes like status codes, department codes, etc.
- Analyzing range constraints - This process identifies attributes whose values lie between two well-defined endpoints. While this often applies to integral values, this can also apply for character string values and dates, as well as other abstract data types.
- Identifying mapped values associated with conceptual domains - As a combination of range constraint analysis and enumerated domain analysis, this process evaluates integer domains to determine if the values are mapped to another set of values that can provide some business context. The issue here is that often the same integral value sets may be used for

many different attributes, even though the numbers relate to different conceptual domains, and therefore the analyst's goal is to identify the mapping between numbers and the values or concepts they actually represent.

- Abstract data type analysis - Often there are specific patterns that value sets exhibit that suggest a more complex data type that may have some semantic meaning. This process reviews patterns and helps the analyst to suggest potential abstract data type definitions used within the data set.

Unraveling conceptual and value domains and their associated mappings is a critical step in any data migration program, whether that is driven by data warehousing or the need for master data management (MDM).

Embedded Structure Analysis

From the reverse engineering standpoint, understanding the intended structure of a data set will help in clarifying its definition, semantics, and the data quality and data transformations associated with a renovated data architecture or a data migration. Some examples include:

- Identifying opportunities for normalization – Functional dependencies discovered through cross-column profiling may indicate redundant data elements that are actually embedded tables. Identifying these dependencies helps in data migration efforts to identify inherent relational structure embedded within relatively flat data.

- Identifying opportunities for consolidation - Identifying the same embedded data used in more than one data set suggests an opportunity for consolidating that data as reference data.

- Understanding relational structure - This process looks for embedded or exposed foreign key relationships, and helps to document a replica version of the existing data model.

- Data Definition Language (DDL) generation - As prospective relational and structural suggestions are made, eventually the DDL for a target representation of the data can be generated as part of a migration strategy.

- Syntactic consistency - Knowing that two data set attributes (or columns) are intended to represent the same set of values, this is a process to ensure that the value sets share the same format specification (i.e., are syntactically compatible).

- Semantic consistency - This is the process of determining that two (or more) columns that share the same value set and are discovered to refer to the same concept have the same name.

Anomaly Analysis

Within a well-controlled data management framework, one might expect that the data analysts will have an understanding of typical data issues and errors that occur in the different data set. But even environments with well-defined processes for data management, there is often little visibility into data peculiarities in relation to existing data dependencies. This opacity is magnified as secondary reuse of data for alternate purposes (such as data warehousing).

To monitor data set usability with respect to meeting data quality expectations of all downstream users, there must be a process to establish a baseline measure of the quality of the data set that is distinct from specific downstream application uses. This objective review "level sets" the expectations for data set acceptability. Anomaly analysis is a process for empirically analyzing the values in a data set to look for unexpected behavior to provide that initial baseline review. Essentially, anomaly analysis:

- Executes a statistical review of the data values stored in all the data elements in the data set
- Examines value frequency distributions
- Examines the variance of values
- Logs the percentage of data attributes populated
- Explores relationships between columns
- Explores relationships across data sets

These capabilities reveal potential errors in data values, data elements or records. Discovered errors can be logged and then brought to the attention of the business data consumers to determine the degree to which each flaw leads to critical business impacts.

The goal of anomaly analysis is to empirically review all the data elements in the data set, examine their frequency distribution and explore relationships between columns to reveal potential flawed data values. This section looks at some common anomalies that can be identified as a result of data profiling.

Column Anomalies

The types of anomalies or potential errors that can be discovered through column profiling mirror the types of analyses performed. Common column anomalies include:

- Sparseness, which identifies columns that are infrequently populated, even when they are expected to have values

- **Unused columns,** indicated either by being largely unpopulated or populated with the same value in all records
- **Nulls,** which examines the percentage of absent values and to identify abstract null representations (e.g., "N/A" or "999-99-9999")
- **Overloaded attributes**, suggesting that columns are used for storing more than one conceptual data element
- **Unexpected value frequencies**, presenting those columns whose values are expected to reflect certain unexpected frequency distribution patterns, or who do not comply with the expected patterns
- **Outlier review**, a process for those columns whose values do not reflect the expected frequency distribution to identify and explore those values whose frequencies are much greater than expected and those values whose frequencies are much lower than expected
- **Range analysis**, which occurs when a column's values do not fall within one (or more) constrained value ranges. This may involve single range analysis or more complex value clustering
- **Format and/or pattern analysis**, which inspects representational alphanumeric patterns and formats of values and highlights when the frequency of each to determine value patterns that are not reasonable
- **Value domain noncompliance**, showing those columns whose values do not comply with known data domains
- **Composed value analysis**, a process that looks for sets of values that appear to be composed of two or more separable values. As an example, consider a product code that is composed of a unique value appended to a code representing the manufacturing site, such as NY-123, the "NY" representing the manufacturing site

Cross-Column Anomalies

Some types of cross-column anomalies include:

- **Derived values**, which looks for columns whose values should be computed as functions of other values within the same record, but are not valid. As an example, a purchase order line total should equal the quantity of items ordered multiplied by the unit cost of the items
- **Non-uniqueness**, which reviews all records in the data set (table) to determine when exact duplicates exist

- Primary key validity, which, when given a set of attributes expected to form a primary key for the data set, shows when expected unique keys are not unique across all records in the table

- Functional dependency invalidity, which is intended to discover records that do not conform to discovered or known functional dependencies

Cross-Table Anomalies

Examples of cross-table anomalies include:

- Referential consistency noncompliance - Referential integrity asserts that for any foreign key relationship, every foreign key in table B that refers to a primary key in table A must exist in table A. Assuming that we know the existence of a foreign key relationship, this process checks its compliance. This process can check for both orphans (child rows without parents) and childless parents if desired.

- Syntactic inconsistency - This process evaluates columns that are expected to take values from the same data domain and checks consistency with the domain rules. This can expose when there is not an agreement of syntactic form among common attributes.

- Semantic inconsistency/synonym analysis - This is the process of determining that two (or more) columns that share the same value set and are discovered to refer to the same concept have the same name, or mapped to the same conceptual data element. Conversely, this process can explore value sets that are intended to represent the same concept but have values that occur in one column but not others (non-overlapping values).

Metadata Compliance

The results of profiling can also be used as a way of determining the degree to which the data actually observes any already existent metadata, ranging from data element specifications, validity rules associated with table consistency (such as uniqueness of a primary key), as well as demonstrating that referential integrity constraints are enforced properly in the data set.

Metadata compliance is a relatively straightforward concept. Every structured data set has some associated descriptive structure metadata, although in some instances the metadata is not explicitly documented. But when documentation exists, either in the form of entity relationship diagrams with a detailed physical or logical model, or in DDL, COBOL copy books, other data model descriptions, Excel spreadsheets, or text documents, the data profiling tool can help the analyst validate the defined metadata against the existing data artifacts.

This structure metadata usually contains information about data tables and about the attributes within each table, including attribute name, data type, length, along with

other constraints, such as whether the field may be null. In addition, appropriate data domains are often documented along with each data element.

There are typically three areas of metadata that can be reviewed for conformance using a data profiling tool:

- Type compliance, in which the data type of values are compared with the expected data types;
- Domain compliance, in which the set of valid values are reviewed for compliance with documented value sets;
- Constraint compliance, in which defined constraints are reviewed for compliance with documented constraints.

In each of these situations, the data quality practitioner should be alerted if any discrepancies are identified:

- Type compliance – A data profiling tool can scan the values in each column and propose the data type and length of the values in the column. These proposed data types and lengths for each column can be compared with documented column metadata.
- Domain compliance – Because profiling provides information about the sets of values that appear within a column, it can be used to compare to any lists of valid values. In this situation, the profiler can use cross-table analysis (actually, the need is for comparing two column sets) to verify that the set of values used in the column are a subset of the values in the enumerated value domain.
- Constraint compliance – The cross-column and cross-table analyses can be used to determine when known constraints are violated. This uses the data rule validation capabilities of the profiler.
- Data model integrity – A data model contains inherent specifications regarding attribute-level constraints, cross-table relationships, and cardinality. For example, a data model may specify that there is a one-to-one relationship between two tables. Profiling can be used to verify that the actual data corresponds to the constraints, as well as more sophisticated approaches to automatically analyzing relational structure and creating a proposed model. This proposed model can then be compared to the actual model as a review of differences between the logical view and what actually is in the data.

In addition to the other metadata aspects discussed in prior sections, it would be worthwhile to capture the discovered data model for tracking purposes to determine if there are specific data model issues that have been highlighted as a result of profiling and whether those model issues have been addressed.

Summary

Data profiling techniques can help the analyst understand the empirical truth associated with enterprise data, and the breadth of analysis techniques can help the analyst draw many conclusions. However, without directing the analyst processes that use profiling technologies, there is a risk of continued review, drill-down, and ultimate "analysis paralysis" in which the specific business objectives are never achieved.

Before starting a data profiling initiative, it's vital to understand the tie between the information gleaned from the reports and the proposed business outcomes of the data-driven initiatives. By clarifying specific analysis processes and techniques (such as those presented in this paper) that use data profiling technology, the analyst team can establish a well-defined scope with specific goals (such as documenting metadata or identifying potential data errors) that can be achieved in a timely and predictable manner.

To learn more about data quality, visit:
dataflux.com/knowledgecenter/dq

YOUR DATA.
YOUR BUSINESS.
ONE SOLUTION.

www.dataflux.com

Corporate Headquarters
DataFlux Corporation
940 NW Cary Parkway
Suite 201
Cary, NC 27513-2792
USA
877 846 3589 (USA & Canada)
919 447 3000 (Direct)
info.us@dataflux.com

DataFlux United Kingdom
Enterprise House
1-2 Hatfields
London
SE1 9PG
+44 (0)20 3176 0025
info.uk@dataflux.com

DataFlux Germany
In der Neckarhelle 162
69118 Heidelberg
Germany
+49 (0) 69 66 55 42 04
info.de@dataflux.com

DataFlux France
Immeuble Danica B
21, avenue Georges Pompidou
Lyon Cedex 03
69486 Lyon
France
+33 (0) 4 72 91 31 42
info.fr@dataflux.com

DataFlux Australia
300 Burns Bay Road
Lane Cove, NSW 2066
Australia
+61 2 9428 0553
info.au@dataflux.com