melissa

A **MELISSA** WHITEPAPER

The Importance of Data Profiling

How Profiling Helps You Make Knowledgeable Business Decisions

By Bud Walker and Joseph Vertido

TABLE OF CONTENTS

INTRODUCTION

Data profiling is a commonly used term in the discipline of data management, yet the perception is that it is elusive, vague, and mostly unappealing to all but the most technical. In this whitepaper, you will rediscover the importance of profiling and explore interesting and useful forms of metadata that the profiling process generates. You will also uncover advanced techniques to ascertain the quality of your data, as well as the ability to automate the consolidation of records from tables having different structures, all through the use of profiled metadata.

Why is data profiling important? The case for data profiling is simple. In order to do any critical work on data, whether it is a new business intelligence or analytics project, a large scale data migration, incremental data quality improvements, or even a full-fledged master data management initiative, it is vital to understand the full extent of the work that needs to be performed. Budgets can be estimated, resources allocated, approval sought from management and the right tools purchased or built, all directly from information derived from the data and not a copy book or human memory. Profiling the columns and tables in the database sets up these complex data projects for success, and is highly recommended before spending potentially millions of dollars on grand BI projects where problems can creep in that could introduce delays, or even skew outcomes.

THE BASICS OF PROFILING

There are many available tools on the market today that provide powerful data profiling capabilities and generate an extensive collection of metadata. A majority of these solutions share common output types of metadata including:

COLUMN STRUCTURE

Maximum/Minimum Lengths and Inferred Data Type –
These types of metadata provide information on proper table formatting for a target database. It is considered problematic, for example, when an incoming table has values that exceed the maximum allowed length.

MISSING INFORMATION

NULLs and Blanks –
Missing data can be synonymous with bad data. This applies, for example, where an Address Line is Blank or Null, which in most cases is considered a required element.

DUPLICATION

Unique and Distinct Counts –
This allows for the indication of duplicate records. De-duplication is a standard practice in Data Quality and is commonly considered problematic. Ideally, there should only be a single golden record representation for each entity in the data.

RANGES

Date, Time, String, and Number Ranges –
Certain fields may accept only a certain maximum or minimum range. Data that is beyond the allowed possible thresholds should be subject to investigation or correction.

SPACES

Leading/Training Spaces and Max Spaces between Words –
Unnecessary spaces must be considered for standardization purposes. Spacing discrepancies may prove problematic, for example, when doing table lookups for exact matches.

CASING AND CHARACTER SETS

Upper/Lower Casing and Foreign, Alpha Numeric, Non UTF-8 Characters –
Casing is also an important piece for Data Quality Standardization. For Canadian Postal Codes, for example, it is recommended for the data to be in upper case. Invalid character set usage is usually an indication for bad data.

THE BASICS OF PROFILING

PATTERNS

Regular Expression Representation and Value Patterns –
RegEx representations allow for identifying deviations from formatting rules. Dates that are in different formats can easily be determined through Regular Expressions. Similarly, value patterns allow for more exact identification of proper formatting. U.S. State Abbreviations, for instance, must follow the pattern AA, which means two capitalized alpha characters.

STATISTICS

Maximum/Minimum Value, Average Value, Quartiles and Standard Deviation –
These statistics allow for gaining a basic understanding on the trends of data such as gathering statistics for income ranges provides knowledge of average income as well as how spread apart they are.

FREQUENCIES

Value and Length Frequencies –
Frequencies give us an overview of the distribution of records. It allows us to view the most and least frequently occurring values and lengths. This can be used to determine trends in the data to generate reports like gathering the value frequency for age to provide information on common age-range demographics from your customer data.

Such forms of metadata are considered as generic types. Generic metadata types provide information for generic data to accommodate any and all forms of data. What this really means is that whether you are profiling a name field, a customer number field, or a social security field, the type of data involved becomes irrelevant as the profiler treats each field of data as if it were the same. It certainly allows for creating useful statistics and assumptions about data, but at a very high level.

DATA-DRIVEN PROFILING METADATA

In order to make data profiling more relevant, new kinds of metadata need to be produced. The use of generic metadata information is useful for gathering a very broad overview of your data, such as how many blanks there are, or the number of repeating values. However, these kinds of metadata don't produce essential information that is relevant to specific domains like contact data. In order to derive a more meaningful analysis of data, the metadata you need to produce must also be specific to each type of domain. In other words, you want to move away from generic metadata types, and move toward using specific kinds of metadata that is relevant to the actual domain involved.

This specialized form of profiling can only be achieved using a data-driven approach. Take for example U.S. ZIP™ codes. If you were to assess, in this case, how many invalid ZIP codes are in the data, the only way this can be done is by making use of a reference data of all possible and valid U.S. ZIP codes. This technique applies to other domains such as phone numbers, social security, city names, etc. In order to assess the quality of the actual contents of data, it calls for a reference data-driven approach.

Through this data-driven approach, you can produce very unique types of metadata that are not simply based on rules, norms, and proper formatting. Here are a few examples:

Postal Code Validity – Number of valid and invalid U.S. / Canadian Postal Codes.

State and Postal Code Mismatch – Number of records where the Postal Code does not exist in the given state.

Country Validity – Count of invalid country names, Alpha-2, Alpha-3, or numeric country abbreviations.

Email Syntax – Number of email addresses with syntax errors, Mobile Domains, Spam Trap Domains, and Disposable Domains.

Phone Validity – Number of invalid U.S. and Canadian phone numbers.

Names Formatting – Number of inconsistently ordered full names as well as suspicious and profane names.

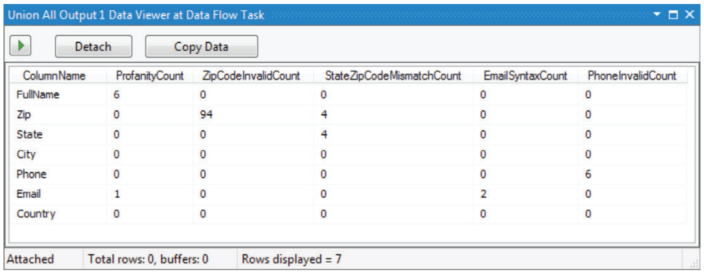
A PRACTICAL APPLICATION OF PROFILING

I. A PRECURSOR TO DATA QUALITY

The purpose of profiling and gathering metadata ultimately, is to gain knowledge, allowing you to ascertain the quality of data and the issues involved in the data assets of your organization. It is a necessary precursor to data cleansing, because in order to fix problems, you first need to know what they are. Data profiling is often a technique used to increase organizational awareness of the need for data quality right from the outset, and to pitch the necessity of data quality tools and sound governance practices to an organization's management team, mainly to win stakeholders that are necessary to free up money for the purchase of off-the-shelf tools.

II. DATA QUALITY STATISTICS

Data profiling not only reveals the problems, but also, how prevalent they are. The metadata produced, in essence, show the statistical occurrences of found issues. Take for example:



ColumnName	ProfanityCount	ZipCodeInvalidCount	StateZipCodeMismatchCount	EmailSyntaxCount	PhoneInvalidCount
FullName	6	0	0	0	0
Zip	0	94	4	0	0
State	0	0	4	0	0
City	0	0	0	0	0
Phone	0	0	0	0	6
Email	1	0	0	2	0
Country	0	0	0	0	0

Attached | Total rows: 0, buffers: 0 | Rows displayed = 7

The numbers generated for the different metadata show the frequency of occurring problems in the data.

III. MONITORING

Monitoring in a general sense can be defined as profiling over time. It is not enough to assume that data has improved after performing the necessary data quality routines. In order to gather quantifiable results, it is also necessary to perform profiling afterwards.

By using the same profiling techniques, it is possible to reassess the current state of the quality of data, how much it has improved, and determine if there are more problems that need to be addressed. In order to assure the quality of data over time, the profiling and monitoring cycle should continuously be employed.

Lastly, profiling is not limited to a passive discovery process. Data profiling with a good tool can also be employed as an active monitoring solution, where data stewards armed with deep knowledge of organizational data can set up rules that will alert them when a field somehow exceeds a predefined rule threshold or is now coming back with a different pattern as data is being streamed by backend jobs. Active monitoring is something that can be employed to safeguard collected data as it is being processed in a change data capture scenario or during delta loads into the data warehouse. Active monitoring is not part of the initial data discovery process but something that is employed by the data steward so that they may keep a finger on the pulse as data is moved around the organization, ensuring that newly introduced problems in the upstream capture process can be caught and remedied proactively.

IV. REPORTING

The only way to produce reports on the quality of data is through profiling. Analytics is an essential part of data quality and the recorded metadata produce trends that show the rate at which issues occur, and how prevalent they are. By analyzing the metadata, you become aware of which areas require improvement for more rigid data quality procedures.

Producing the needed reports over time provides an effective way for visualizing and ensuring that the quality of data is maintained.

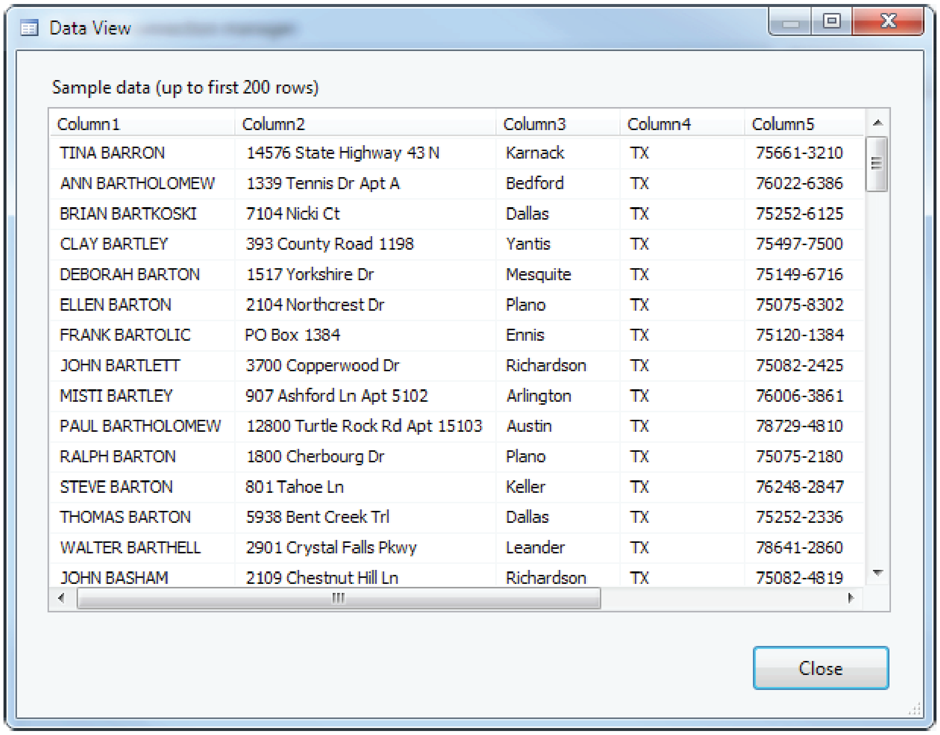
INFERRED DATA TYPES AND AUTOMATED DATA CONSOLIDATION

The knowledge you get from profiling and metadata also extends to other very useful applications such as the automation of importing and consolidating data.

Record consolidation is a very common scenario that many Database Administrators have to deal with. It is also very common that the data sourced from multiple files are of varying formats. This includes different header names, different header order, or even headers that are not common between all the sources. On top of this, new sources may arrive that introduce yet another new file format. This is an extremely problematic and time consuming scenario in file consolidation that must be resolved manually.

In order to provide some form of automation for consolidating data, Profiling could also be employed. Again, profiling by definition provides you with the metadata for understanding and gaining knowledge about your data – this includes information on the actual data type of each column. By utilizing techniques that involve both logic and reference data, you can accurately assume the data type contained in each column – this is also known as Inferred Data Types.

Take for example the following data:



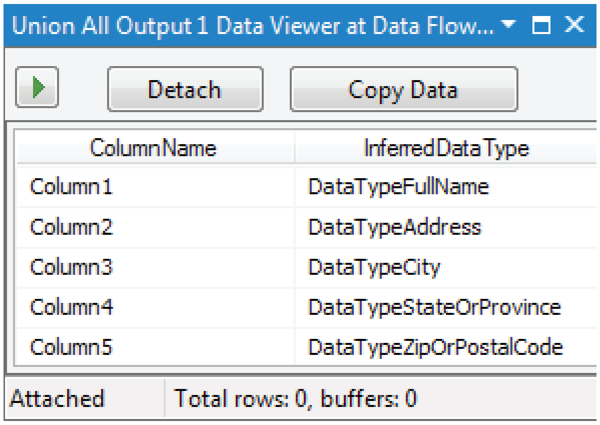
Sample data (up to first 200 rows)

Column1	Column2	Column3	Column4	Column5
TINA BARRON	14576 State Highway 43 N	Karnack	TX	75661-3210
ANN BARTHOLOMEW	1339 Tennis Dr Apt A	Bedford	TX	76022-6386
BRIAN BARTKOSKI	7104 Nicki Ct	Dallas	TX	75252-6125
CLAY BARTLEY	393 County Road 1198	Yantis	TX	75497-7500
DEBORAH BARTON	1517 Yorkshire Dr	Mesquite	TX	75149-6716
ELLEN BARTON	2104 Northcrest Dr	Plano	TX	75075-8302
FRANK BARTOLIC	PO Box 1384	Ennis	TX	75120-1384
JOHN BARTLETT	3700 Copperwood Dr	Richardson	TX	75082-2425
MISTI BARTLEY	907 Ashford Ln Apt 5102	Arlington	TX	76006-3861
PAUL BARTHOLOMEW	12800 Turtle Rock Rd Apt 15103	Austin	TX	78729-4810
RALPH BARTON	1800 Cherbourg Dr	Plano	TX	75075-2180
STEVE BARTON	801 Tahoe Ln	Keller	TX	76248-2847
THOMAS BARTON	5938 Bent Creek Trl	Dallas	TX	75252-2336
WALTER BARTHELL	2901 Crystal Falls Pkwy	Leander	TX	78641-2860
JOHN BASHAM	2109 Chestnut Hill Ln	Richardson	TX	75082-4819

INFERRED DATA TYPES AND AUTOMATED DATA CONSOLIDATION

Upon initial analysis, you can see that the header names used give no indication on the type of data within each column. It can also be true that the order of columns, without proper header names and format, be ordered arbitrarily. In order to understand this data, the only solution is to manually investigate the contents of each column. This is necessary to successfully import and consolidate with other data.

To automate this, you must also be able to automatically detect the contents within each column and the metadata produced are the Inferred Data Types. If you were to infer the 5 columns in the sample table, it would yield the following results:



ColumnName	InferredDataType
Column1	DataTypeFullName
Column2	DataTypeAddress
Column3	DataTypeCity
Column4	DataTypeStateOrProvince
Column5	DataTypeZipOrPostalCode

Attached   Total rows: 0, buffers: 0

It is possible to successfully automate data inference by creating the rules and structures that define each domain of interest. For example, in order to infer U.S. states, you may check whether a column contains 2–character capitalized strings. However, that can be taken further by using a reference data of all possible U.S. states and their abbreviations, providing accurate knowledge as to whether a column contains U.S. states. This technique can be applied to other domains. Social Security, Product Codes, Dates, and so on, all have a set of rules, structure, and content that can be defined. It is through these definitions that you gain the necessary profiling knowledge in order to provide accurate metadata for producing inferred data types.

Automating the import process would just be a matter of inferring the data type for each column, and importing the columns needed to be consolidated.

Conclusion

Profiling is a very broad subject that involves many different aspects of data. However, what is important to remember is that the metadata and knowledge generated through profiling can be leveraged in many ways. It is through this knowledge of your data that you are able to assess its quality, what needs to be improved, and how much you are able to improve it over time. It also provides other capabilities, such as the automation of data consolidation by inferring data types, and rule-based data quality, all through the information gained through profiling metadata.

melissa

www.melissa.com

About Melissa

Since 1985, Melissa has specialized in global intelligence solutions to help organizations unlock accurate data for a more compelling customer view. Our breadth of data and flexible API technology integrates with numerous third-party platforms, so it works for you and makes sense for your business. More than 10,000 clients worldwide in key industries like insurance, finance, healthcare, retail, education, and government, rely on Melissa for full spectrum data quality and identity verification software, including data profiling, cleansing, matching, and enhancement services, to gain critical insight and drive meaningful customer relationships.

US

22382 Avenida Empresa
Rancho Santa Margarita, CA 92688-2112

800.MELISSA (635.4772)

UK

+44 (0)20 7718 0070

GERMANY

+49 (0) 221 97 58 92 40

INDIA

+91 (0)80 4854 0142

AUSTRALIA

+61 02 8091 6000