269 DZone REFCARDZ

BROUGHT TO YOU IN PARTNERSHIP WITH melissa

Understanding Data Quality

WRITTEN BY SIBANJAN DAS, BUSINESS ANALYTICS AND DATA SCIENCE CONSULTANT

CONTENTS

Data has always been the heart of organizations. It is the keystone for running day-to-day business smoothly and for implementing new strategies in an organization. The ability to analyze data and make data-driven decisions is becoming increasingly important.

Individuals are also highly benefited by the use of data. Be it investing in stocks or finding a suitable house to buy, data provides a wealth of information for us to make decisions. Data is the foundation of decision-making and provides information, helps derive various insights, and helps make predictions required for effective decision-making. There are multiple sources from which data is collected. For example:

- **Internal databases**: These constitute an organization's most relevant and reliable data source. They are usually in a structured format and commonly record data from various internal applications like ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), and HCM (Human Capital Management).

- **Flat files**: Flat files are one of the most used data sources for an organization. Flat files arise from sources that are external to an organization, or when there is no proper mechanism to integrate various internal data sources. For example, a vendor can send periodical flat files that can be uploaded to an organization's internal databases. Also, in cases where is no integration between two or more applications in an organization, flat files serve as a medium to exchange data. Most of the time, the data in a flat file is considered unreliable, and several checks are performed to verify and validate the data.

- **Web services and APIs**: Web services are a highly preferred medium for communication and data exchange between different applications. They provide a standardized way to communicate and exchange data. They are reliable, and data validation can be embedded easily.

Other sources like data from social media, blog posts, audio, and videos are gradually becoming vital sources of information that need to be stored and analyzed.

However, not all data are useful or serve a given need. For instance, let's say I am looking to buy a house. However, I get data that provides historical trends of house purchases from a different area other than where I am considering. This does not fit my need. The data is not going to serve the purpose unless the information is good enough.

Data that is fit for intended use is termed as useful data. Bad data inhibits the process of analysis. Finding a reliable dataset straight away is very difficult. We have to craft and nurture good data. In this Refcard, we will discuss various techniques to manage, monitor,

Is bad data threatening your business?
Call in the
Fabulous 4
It's Clobberin' Time...with Data Verify Tools!
Address!
Phone!
Email!
Name!

Visit Melissa Developer Portal to quickly combine our APIs (address, phone, email and name verification) and enhance ecommerce and mobile apps to prevent bad data from entering your systems.

With our toolsets, you'll be a data hero – preventing fraud, reducing costs, improving data for analytics, and increasing business efficiency.

- **Single Record & Batch Processing**
- **Scalable Pricing**
- **Flexible, Easy to Integrate Web APIs: REST, JSON & XML**
- **Other APIs available: Identity, IP, Property & Business**

Let's Team Up to Fight Bad Data Today!

melissadeveloper.com
1-800-MELISSA

melissa

and improve data quality in an organization. Some of this can also be useful for individuals who rely on data for their activities.

High-quality data has the following properties:

1. Fit for use — correct and complete.
2. Proper representation of the real-world scenario to which it refers.
3. It is usable, consistent, and accessible.

Data quality can be measured based on the following dimensions:

- **Completeness**: Is there any missing or unusable data?
- **Conformity**: Does the data conform to a standard format?
- **Consistency**: Are the data values providing consistent information or giving conflicting information?
- **Accuracy**: Is the data correct or out-of-date?
- **Duplicates**: Are the data records or attributes repeated where they should not be repeated?
- **Integrity**: Is the data referenceable or there are missing constraints?

There are two main characteristics that define data quality.

1. DATA USABILITY

Usablility means the data can contribute relevant content required for a particular task. For example, data on customer age or location might contribute well to a customer retention program for the consumer packaged goods industry. However, data on weather or the soil quality of customer locations might not be usable for this retention activity. However, this weather or soil quality data might be useful for customers targeted for the floral industry. So, data usability correlates with its ability to drive action/insight for particular tasks, and it needs to be an accurate representation relevant to the work. When similar data is present at multiple locations such as different databases and data warehouses, they need to be synchronized to have the same representation of the data.

2. DATA QUANTITY

Data quantity defines the amount of data required for an analysis. Estimating and assessing the data quantity at the beginning of a data quality initiative is crucial for the success of the program. Do we need too little or too much data? What is the number of observations? What are the drawbacks of not having much data? These are questions that can help us decide the tools and techniques required to drive the data quality initiative.

Manually inspecting the data to ensure fit for use is the best way to ensure data quality. This is possible when the data quantity is too small. However, with the volume of data we currently have, it is too high to rely solely on the manual process. To eliminate human error and reduce data inaccuracies, we have to depend on various technologies and techniques. We need to follow a data quality strategy to ensure the data is of high quality. There are different phases that provide the ability to manage, monitor, and improve data quality, as given below:

- **Parsing and standardization**: A process to extract pieces from data to validate if it follows a specific pattern. If it doesn't fit the pattern, the data is formatted to provide consistent values.
- **Generalized cleansing**: A process to remove errors and inconsistencies in data.
- **Matching**: A process to compare, identify, or merge related entities across two or more sets of data.
- **Profiling**: A process to analyze the content of a dataset for validating the accuracy, consistency, and uniqueness of data.
- **Monitoring**: A process to continuously access and evaluate the data to ensure it is fit for the purpose.
- **Enrichment**: A process to enhance data quality by using data from various internal and external sources.

GENERALIZED CLEANSING

Generalized cleansing is a process to modify the data to meet data quality requirements for an organization based on defined business rules. It can be as easy as changing the titlecase of a letter to search and replacing any part of a string. Below are a few generalized cleansing operations that are carried out:

- Adding or removing a punctuation mark.
- Expanding or contracting abbreviations (e.g. NC to North Carolina).
- Case folding (e.g. changing the titlecase from capital to lowercase or vice versa).
- Replacing part of a string.
- Creating expressions by combining different values.
- Using regular expressions to extract terms and reduce words to their root forms.

PARSING

Data usually conforms to specific patterns. For example, pin codes, telephone numbers, and email addresses have certain data patterns. Parsing is a way to analyze these strings of characters and symbols using certain rules to validate whether the data meets the pattern. For example, there is a requirement to have a telephone number in a format of (TTT) TTT-TTTT. This format can be a rule set up in a software engine, generally known as a parser. When a telephone number (987) 786-4532 is supplied to the parser, it runs this number through the rule to validate if it meets the requirements. A parser can also use the authoritative reference data sources to verify the accuracy and reliability of the telephone number. Authoritative reference data is a source of information that is considered very trustworthy and that is compiled from internal master data sources or from official publications.

This is basically how data parsing works. This is an example of a simple parser. In the real world, a parser is usually complicated with several rules embedded in it.

We just touched upon validation and verification in our previous example. Before getting into the details of other data quality phases and techniques, let's comprehend the meaning of the terms *data verification* and *data validation*.

Verification + **V**alidation = **V**alue addition to data quality

People working on data quality initiatives hear these two words quite often. Most of the time, we get confused about the difference between data verification and validation and use them interchangeably. However, there is some significant difference between the two that sets them apart.

VERIFICATION AND VALIDATION

Data validation is a process to compare the dataset to a set of business rules to determine whether it conforms to the business and IT system's data requirements and ensure that it is logical and reasonable to use. There can be rules such as:

- Checking the data format.
- Checking that the data is of an appropriate type, e.g. salary should be a whole number.
- Validating whether the data is present for a not null column.

On the other hand, data verification is a process to ensure that the incoming data exactly matches the source from which it originated. This ensures that the data is accurate and error-free. Verification is mostly done through the following methods:

- Data is checked by comparing it with the original source. For example, invoice details in a system are verified with the original invoice document.
- Double entry checks where data is entered twice and it is compared for discrepancies. For example, password verification is done by asking the users to enter the password twice.
- Data can also be verified by phone calls and email verification, e.g. calling a new prospect to verify details before moving forward with the marketing process. Similarly, it can also mean verifying the genuineness of a new customer by sending an email to their email address.

STANDARDIZATION

Parsing aids in dividing data into parts and validating whether the data meets the standards — which can be industry standards, standards set by governments, or user-defined standards. When a text does not meet a standard, it can be standardized by doing certain transformations. Continuing with our previous example, the representation (TTT) TTT-TTTT can be considered a standard form for the telephone number. If a telephone number (987)7658974 is passed through the parser, it will recognize it as an invalid phone number. In such cases, the missing hyphen can be inserted in between to transform it to (987) 765-8974, which is the standard form. This process of transformation is known as standardization.

PARSE > CLEAN > TRANSFORM CYCLE

Parsing, cleaning, and data transformation are vital and iterative processes for maintaining high-quality data in this digital age, where the major chunk of data is unstructured. If an organization is interested in some natural language processing activity, we cannot straight away store the raw text data for getting some insights out of it or fitting it to a machine learning model. First, we need to parse it so that the text is split into words as well as clean it to make it suitable for analysis.

MATCHING

Matching is a process to compare, identify, or merge related entities across two or more sets of data. For example, in an organization, customer account information is stored across various IT applications such as CRM, order management, and account receivables. Data-matching techniques can be employed to remove duplicate content and identify key links between these systems of records to provide a single source of truth. Finding out the "relevant associations" between different records is the key for matching techniques.

Some of the terminologies associated with matching are:

- **Linking**: This is the most straightforward data matching task. It involves linking records to the fixed structured reference set.
- **Deduplication**: All records from a dataset are matched to records from another dataset or even to each other to merge or eliminate duplicate records, e.g. merging multiple

records of a customer who has registered multiple times on a website. The primary challenge in deduplication is to decide which fields in the data need to be considered from the duplicated records and retained in the master record. This can be achieved by framing specific rules like picking the data that was most recently changed, picking data from a trusted source, or choosing fields with more details to make the data meaningful. Let's say that a customer has different names, M. Kopchak and Michael Kopchak, in two records; we can retain the name as Michael Kopchak in the master record as it is more meaningful and detailed than the previous one. This ability to choose the surviving fields of similar records that will be retained in the master record is known as survivorship.

- **Auto-tagging**: In this technique, the documents or records are matching to a fixed set of tags. An example includes segregating products based on different product type.

Sometimes, there are unique identifiers absent in a dataset that hinder the process of data matching. In such cases, probabilities to be similar content is weighed, that can be applied to match the records. Broadly, there are two main matching techniques:

1. **Deterministic Matching**
   With this technique, we try to find an exact match between records. This technique is generally straightforward and is an ideal technique when the record contains some unique identifiers like Social Security numbers, customer IDs, and employee IDs. However, sometimes, collecting the unique identifiers is difficult and impossible. For example, a customer probably won't provide information about his Social Security number while buying groceries. In such cases, several pieces of his information such as address, phone number, age, and email address can be matched separately to generate matching scores that can be rolled together to get an overall matching score. Deterministic matching is reasonably easy to define and implement.

2. **Fuzzy Matching**
   It is not always possible to have exact matches or employ deterministic matching techniques. Previously, I worked on a project where I had to extract product information from websites such as Amazon and match it with the products in our internal procurement database to do various business improvement analysis. We retrieved certain products, e.g. an iPhone with 64 GB. By manual inspection, we match it to the Apple iPhone that was in our database. However, this process cannot be done for each and every record when there are thousands of products to be matched. For this type of case, we have to rely on fuzzy matching techniques. In this technique, the records are matched based on the degree of similarity between two or more datasets. Most of the time, in fuzzy matching techniques, probability and statistical techniques are used to generate matching scores. Regular expressions are also used widely to extract parts of the potential matching attributes.

Listed below are some fuzzy matching algorithms that are part of various tools or that are available as multiple open-source libraries, such as the stringdist package in R.

LEVENSHTEIN DISTANCE ALGORITHM

This algorithm is used to measure the similarity between two strings. For example, if there are two strings — s as "data" and t as "data" — the Levenshtein distance LD(s,t) is zero. But for s = "data" and t = "date," the Levenshtein distance LD(s,t) is 1, as "a" in string s has to be replaced with "e" in string t to make them similar. So, the Levenshtein distance can be defined as the minimum number of insertions, substitutions, or deletions required to transform a string s to string t.

JACCARD DISTANCE

This algorithm measures the similarity between the values of two or more attributes in a dataset by comparing which values of the attributes are similar and which are distinct. If A and B are two data attributes, the Jaccard distance can be computed by using the below formula:

J(A,B) = |A n B|/|A U B|

For example, if an attribute A has values {J,K,L,T} and another attribute B has values {U,V,J,K,L,N}, then:

J(A,B) = {J,K,L}/{U,V,N,J,L,K,T} = 3/7 = 0.42

So, attributes A and B in this example have 42% similarity based on the Jaccard distance computation. The higher the similarity percentage, the more similar the attributes.

JARO-WINKLER DISTANCE

It is also an algorithm to compute the measure of similarity between different strings. It is computed using the below formula:

$$d_i = \begin{cases} 0\ , if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|a_1|} + \frac{m}{|a_2|} + \frac{m-t}{|m|}\right) \end{cases}$$

Here, m is the number of matching values and t is the number of transpositions. Transposition is defined as the number of matching values in different sequence orders divided by two.

For example, where string = data and = date, the Jaro-Winkler distance is computed as below:

$$d_i = \frac{1}{3}\left(\frac{3}{4}+\frac{3}{4}+\frac{3-0}{3}\right)\underline{= 0.833}$$

So, in this example, the strings and has a similarity of 83.3%.

PROFILING

Data profiling is the process of analyzing the content of a dataset for validating the accuracy, consistency, and uniqueness of the data. This is carried out using several statistical analysis techniques to provide several informative summaries about the data that aids in providing insights into the data quality. Below are some of the analyses performed:

- **Completeness analysis**: Used to check whether the data is correct. Analyzing each data attribute for missing or null values helps identify potential data issues. Questions like how often a column is populated versus how often it is blank or null can help eliminate data issues flowing into our target databases.
- **Distribution analysis**: Done to identify the data distribution of the values of an attribute. This helps in identifying the presence of outliers that can distort the overall distribution of data. Outliers are the values in an attribute that differ from the majority of the values in that particular column.
- **Uniqueness analysis**: Helps to determine the records that are uniquely identified by an attribute or a group of attributes. This assists in identifying duplicates and also in identifying whether linking between datasets is possible.
- **Statistical analysis**: Range analysis can be done on numerical and date types of data to identify appropriate value ranges. Calculating different summary statistics for a dataset (such as the minimum and the maximum value of an attribute, mean, mode, standard deviation, five quantile values, and top and bottom five values) can provide insight into the quality of a dataset.

All of these practices are part of various data quality profiling tools, and the functionalities are a part of the below techniques:

- **Column profiling**: This profiling technique helps us analyze the data distribution of columns including outlier analysis, as well as statistical and uniqueness analysis. Some tools also suggest rules that can be added to a column.
- **Mid-stream profiling**: As the name suggests, data profiling can be performed in the middle of the data stream without creating a mapping.
- **Join profiling**: This profiling technique is used to determine the data overlap between two or more datasets and analyze the referential integrity between those datasets, e.g. validating the primary and foreign keys between two or more join conditions.

MONITORING

Data monitoring is a process to continuously access and evaluate the data to ensure it is fit for the purpose. It helps to track unusual or abnormal data behavior and changes in data quality. Data monitoring is done to ensure that all existing and incoming data meets business rules. Through on-going data monitoring, it provides the ability to ensure that we are well-positioned to capitalize on the information requirements and that high-quality data standards are well maintained. If we cannot monitor the data, we cannot manage it.

The first step in monitoring is usually to collect data. Based on the data, various metrics are set up to be monitored. For example, "there cannot be more than 20% worth of null values in the employee's salary attribute" can be a metric set up to be monitored.

One of the other techniques is to create a baseline for normal performance and compare the results over time. The value of the metric is termed as a threshold.

When the data is collected, it is compared with the threshold value of the metric to ensure it meets the set-up criteria. If the data doesn't meet the threshold value criteria, it indicates that the data is weak and doesn't meet the data quality requirements.

DATA ENRICHMENT

Data enrichment is a process of enhancing data quality by using data from various internal and external sources. It augments the existing data by enriching it with supplemental datasets. We can have useful data, but enrichment benefits us by making it better.

There are various methods available both internally and externally that can be used to enrich the data.

The data can be enriched by integrating various internal data sources. For example, let's say that we have some customers in our organization who are our suppliers, too. In this case, the customer data is available in the organization's order management system as well as in procurement systems in the form of suppliers. There might be a subtle difference in these systems of records based on the design of the applications or the information recorded in the system. The customer information can be enriched by combining these datasets and creating a single source of truth.

Similarly, if we are in a business where location plays a significant role (such as in transportation, logistics, insurance, and door-to-door retail services), then the customer and user information can be enrichment using location intelligence data services.

Big data and machine learning can also be utilized for data enrichment. They are changing the previous way of business. They are providing businesses with insights and predictions for driving data-driven insights. Machine learning techniques can be employed for enriching the datasets by providing various techniques for improving the quality of a dataset. For example, machine learning models can be used to clean and impute the records that best suits the dataset. They can also be used to classify a massive number of documents and add tags that make it easy to manage the datasets.

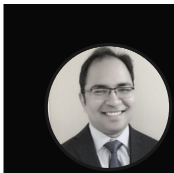
SUMMARY

There is a crucial need for organizations to have a data quality strategy in place to maintain and improve the quality of the data. It helps reduce data inaccuracies and eliminate errors that can hinder our ability to draw actionable insights needed to move the business forward. With the growth of technologies and the addition of new data sources, there is a need to ensure the continued usability of the data. This requires us to develop a sophisticated approach to maintain data quality, i.e. a more proactive and optimized way to handle data. The methods and process outlined here provide a model for implementing a data quality program.

However, we just touched the tip of the iceberg and provided a head start on driving data quality initiatives. If you have truckloads of data and are wondering how to manage high-quality data, below are some additional readings that can provide you with in-depth knowledge of this topic.

ADDITIONAL READINGS

1. Data Quality Assessment by Arkady Maydanchik
2. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (TM) by Danette McGilvray



Written by **Sibanjan Das**, Business Analytics and Data Science Consultant

Sibanjan Das is Business Analytics and Data Science consultant. He has over seven years of experience in IT industry working on ERP systems, implementing predictive analytics solutions in business systems and Internet of Things. Sibanjan holds a Master of IT degree with a major in Business Analytics from Singapore Management University. Connect with him at his twiiter handle @sibanjandas for following the latest happening on the Data Science, Big Data and AI.

DZone communities deliver over 6 million pages each month to more than 3.3 million software developers, architects and decision makers. DZone offers something for everyone, including news, tutorials, cheat sheets, research guides, feature articles, source code and more. "DZone is a developer's dream," says PC Magazine.

DZone, Inc.
150 Preston Executive Dr. Cary, NC 27513
888.678.0399 919.678.0300