DATA INTEGRITY VS DATA QUALITY: AN INTRODUCTION

Big Data has been widely labeled as the new oil and the new black gold – parallels that describe the value of big data to our economy and business. However, the analogy only fits in limited situations.

Big Data becomes a truly valuable commodity only when the data is of high quality determined based on a range of qualitative and quantitative variables. These variables or dimensions may encompass data accuracy, completeness, consistency, timeliness, validity and uniqueness. Similarly, the data also needs to maintain its integrity to facilitate reliable decisions.

However, Data Integrity is often used as a proxy term for Data Quality. For data-driven business organizations, the parameters and metrics that define the quality and integrity of data present vastly different implications. A brief primer is therefore required to explore the differences between Data Quality and Data Integrity:

Data Quality vs Data Integrity

Data Quality refers to the characteristics that determine the reliability of information to serve an intended purpose including planning, decision making and operations. It is the state of complete features and attributes that define the usability of information to address specific needs in context of real-world circumstances and implications. This Wikipedia resource highlights a range of definitions that provide ample perspective on understanding the term, Data Quality.

Data Integrity refers to the characteristics that determine the reliability of the information in terms of its physical and logical validity. Data Integrity is based on parameters such as accuracy, validity and consistency of the data across its lifecycle. It is the absence of unintended change to the information between two successive updates or modification to data records. Data Integrity can be considered as a polar opposite to data corruption that renders the information as ineffective in fulfilling desired data requirements. This Wikipedia resource explores different types Integrity and constraints that fall within the scope of the term, Data Integrity.

Exploring Data Quality vs Data Integrity

Essentially, Data Integrity is a subset of Data Quality, which relates to characteristics beyond the validity of data as described below:

1. Completeness

An indication of the comprehensiveness of available data, as a proportion of the entire data set possible to address specific information requirements. This proportionality is measured as a percentage and is defined based on specific variables and business rules. For instance, consider a list health records of patients visiting the medical facility between specific dates and sorted by first and last names. The data resource will be considered as 100 percent complete even if it doesn't include the address or phone numbers of the patients, but includes all necessary health records, the first and last names within specific dates. The percentage of completeness reduces in absence of any critical data item.

2. Uniqueness

A discrete measure of duplication of identified data items within a data set or in comparison with its counterpart in another data set that complies with the same information specifications or business rules. For instance, consider the same list of health records as mentioned earlier that should cover 100 patients as per the real-world assessment. If the list contains more than 100 items, then one or

more patient must have had their data duplicated and listed as a separate entity. Depending upon the circumstances and business requirements for the data analysis, this duplication could lead to skewed results and inaccuracies. Mathematically, uniqueness may be defined as 100 percent if the number of data items in the real-world context is unique and equal to the number of data items identified in the available data set.

3. Timeliness

The degree to which the data is up-to-date and available within acceptable time frame, timeline and duration. The value of data-driven decisions not only depends on the correctness of the information but also on quick and timely answers. The time of occurrence of the associated real-world events is considered as a reference and the measure is assessed on a continuous basis. The value and accuracy of data may decay over time. For instance, data about the number of traffic incidents from several years ago may not be completely relevant to make decisions on road infrastructure requirements for the immediate future.

4. Validity

A measure of conformity to the defined business requirements and syntax of its definition. The scope of syntax may include the allowable type, range, format and other attributes of preference. It is measured as a percentage proportion of valid data items compared to the available data sets. In context of Data Integrity, the validity of data encompasses the relationships between data items that can be traced and connected to other data sources for validation purposes. Failure to establish links of valid data items to the appropriate real-world context may deem the information as inadequate in terms of its integrity. Data validity is one of the critical dimensions of Data Quality and is measured alongside the related parameters that define data completeness, accuracy and consistency – all of which also impact Data Integrity.

5. Accuracy

The degree to which the data item correctly describes the object in context of appropriate real-world context and attributes. The real-world context may be identified as a single version of established truth and used as a reference to identify the deviation of data items from this reference. Specifications of the real-world references may be based on business requirements and all data items that accurately reflect the characteristics of real-world objects within allowed specifications may be regarded as an accurate piece of information. Data accuracy directly impacts the correctness of decisions and should be considered as a key component for data analysis practices.

6. Consistency

This measure represents the absence of differences between the data items representing the same objects based on specific information requirements. The data may be compared for consistency within the same database or against other data sets of similar specifications. The discrete measurement can be used as an assessment of data quality and may be measured as a percentage of data that reflect the same information as intended for the entire data set. In contrast, inconsistent data may include the presence of attributes that are not expected for the intended information. For instance, a data set containing information on app users is considered as inconsistent if the count of active users is greater than the number of registered users.

The comparison of Data Quality vs Data Integrity largely centers around the dimension of validity associated with the data. In context of Data integrity, the attributes of data completeness accuracy and consistency are also closely related, followed by the completeness of information. The timeliness and uniqueness of data are more useful to understand the overall quality of data instead of the integrity of information. In addition to these six key dimensions of Data Quality, every organization may use their own metrics and attributes to understand the true value that the available information holds for them.