DATA INTEGRITY VS DATA QUALITY: AN INTRODUCTION

Big Data has been labeled the new oil—parallels that describe the value of big data to our economy and business. Like oil, data has degrees of value based on its quality, integrity, and refining process. Different aspects of its quality encompass data accuracy, completeness, consistency, timeliness, validity, and uniqueness.

The terms quality and integrity can get mixed, but for data-driven businesses, the parameters and metrics that define the quality and integrity of data have vastly different implications. And that's why we put together this brief primer—so you can fully understand the differences between Data Quality and Data Integrity.

What is data quality?

Data Quality characterizes how reliable the information is to serve some intended purpose. This purpose might be:

- Planning
- Decision-making
- Operations

When the data is complete, full of all features and attributes, it is usable information to address specific real-world circumstances.

(Learn all about data quality management.)

What is data integrity?

Data Integrity characterizes how reliable the information is in terms of its physical and logical validity. Data Integrity is based on parameters such as:

- Accuracy
- Validity
- Consistency of the data across its lifecycle

Integrity is the absence of unintended change to the information between two successive updates or modification to data records. Let's consider Data Integrity as the polar opposite to data corruption, which renders the information ineffective for fulfilling desired data requirements.

Data quality vs data integrity

Data that has integrity can be quality data, but not all quality data has integrity. Below, we describe some characteristics of quality and integrity:

1. Completeness

Completeness is an indication of the comprehensiveness of the available data. It can be measured as a percentage to state how much of the entire possible data set is available. The proportion required to classify the data as “complete” determines on the business and particular reason for it.

For instance, consider a list of health records of patients visiting the medical facility between specific dates and sorted by first and last names. The data resource will be considered 100% complete as long as it includes all necessary health records, and the first and last names, within specific dates. In fact, even if it doesn’t include the address or phone numbers of the patients, we can consider it complete because the ask did not include this information.

Conversely, the percentage of completeness reduces as any critical data item(s) are absent.

2. Uniqueness

Uniqueness is a measurement of duplication. Does the data, or similar data, exist multiple times within the same dataset? It is a discrete measure on particular items within a data set.

Example: Consider the same list of health records as mentioned earlier. There are 100 patients in a hospital. If the list contains more than 100 patients, then one or more patients must have had their data duplicated and listed as a separate entity. Whichever patient's list record is duplicated is considered not unique. 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% 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

Timeliness is the degree to which data is up-to-date and available within an 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

Data validity is a test of whether the data is in the proper format. Does the data input match the required input format? Examples include:

- Is a birth date written as Month, Day, Year or as Day, Month, Year?
- Are times based on local time zones, user device time, or the global UTC time?

The scope of syntax may include the allowable type, range, format, and other attributes of preference.

In particular, validity is measured as a percentage proportion of valid data items compared to the available data sets (i.e.; "The 90% of the data is valid.") In the context of Data Integrity, the validity of data also includes 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

Accuracy is 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

Consistency measures the similarities between 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 reflects 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 inconsistent if the count of active users is greater than the number of registered users.

So, timeliness and uniqueness of data are useful to understand the overall quality of data instead of the integrity of information. Data completeness, accuracy, and consistency are good measurements of data integrity.

And, what each data item will actually be is unique to each organization. That responsibility is left up to you. Now let's turn to look at data integrity in the real world.

Data integrity in practice

Data quality and integrity are important in the machine learning and analytics worlds. When data is the resource from which all decisions are based, then quality data allows for quality decisions. But what happens when your data is invalid, inaccurate, or inconsistent?

Let's see!

Decisions on invalid data...

Maybe for legal reasons. Differing formats. Anomaly detection.

Invalid data changes the actual input data, and, if left as is, the decisions that are made are completely wrong. It'd be like determining to feed a person a hearty breakfast because they always eat dinner at 7pm, but their time data is invalid and set to a different time zone, and, really, they should be eating their dinner.

When data points are invalid, it makes decisions around detected anomalous data points weaker, and decisions about the actions to take weaker as well. Invalid data can sometimes be redeemed by interpreting it (such as converting errors to what you think they should be), but that comes at the cost of time, and labor, and the fact that the truth gets lost in translation.

Decisions on inaccurate data...

Accurate data is easy. Generally, when gathering data, people ask questions relevant to their domain—they understand what is useful to their business and what is not.

Where inaccuracy most commonly occurs is when data comes from an outside source and is retrofitted to suit one's needs. In machine learning, this is called Domain Switching, and, with no surprise, a machine learning model trained to predict someone's opinion of a movie from a dataset of IMDB movie reviews will perform poorly when set to predict someone's opinion of their Friday night date.

When the purpose of the model has switched, the dataset has become inaccurate, and the model's performance suffers.

Decisions on inconsistent data...

Like in the real world, if the data is inconsistent, then the outcomes are unpredictable. All models and decisions, if they are to be modeled, require patterns in behavior. Consistency is a prerequisite to pattern detection, and if the data is inconsistent no patterns can be detected.

In the early days of AI imaging, converting handwriting was a hard task because everyone's penmanship was inconsistent. The Palm Pilot, one of the first handheld touchscreen devices, developed its own written alphabet to help its users and its device communicate with one another.

To combat data inconsistency, the solution can vary. You might:

- Increase the size of the data set.
- Redefine the data required so that what is collected can be consistent.

Related reading

- BMC Machine Learning & Big Data Blog
- Data Architecture Explained: Components, Standards & Changing Architectures
- Artificial Intelligence (AI) vs Machine Learning (ML): What's The Difference?
- Data Streaming Explained: Pros, Cons & How It Works
- What Is a Canonical Data Model? CDMs Explained
- DataOps Explained: Understand how DataOps leverages analytics to drive actionable business insights