collibra.com

The 6 dimensions of data quality

Ankur GuptaDirector Product Marketing

Measuring data quality is critical to understand if you want to use enterprise data confidently in operational and analytical applications. Only good quality data can power accurate analysis, which in turn can drive trusted business decisions.

According to one Gartner estimate, poor data quality can result in additional spend of $15M in average annual costs. Although it is not just about financial loss. Poor quality of data affects your organization at multiple levels:

- Higher processing cost: The rule of ten states that it costs ten times as much to complete a unit of work when the data is flawed than when the data is perfect
- Unreliable analysis: With lower confidence in reporting and analysis, bottomline management is never easy
- Poor governance and compliance risk: Compliances are no longer optional, and business survival gets challenging without them
- Loss of brand value: When organizations constantly make erroneous operations and decisions, the brand value decreases quickly

Bad quality data impacts an organization's business strategy of fueling growth and driving innovation. The immediate concern is how an organization can measure data quality and find ways to improve it.

How is data quality measured?

Data quality may be easy to recognize but it is difficult to determine precisely. You can consider multiple attributes of data to get the correct context and measurement approach to data quality. For example, patient data in healthcare must be complete, accurate, and available when required. For a marketing campaign, customer data needs to be unique, accurate, and consistent across all the engagement channels. Data quality dimensions capture the attributes that are specific to your context.

What is a data quality dimension?

Data quality dimensions are measurement attributes of data, which you can individually assess, interpret, and improve. The aggregated scores of multiple dimensions represent data quality in your specific context and indicate the fitness of data for use.

On average, 47% of recently created data records have at least one critical (e.g., work-impacting) error. High-quality data is the exception, with only 3% of the DQ scores rated acceptable (with >97% acceptability score). So, only 3% of companies' data meets basic quality standards.

Scores of data quality dimensions are typically expressed in percentages, which set the reference for the intended use. For example, when you use 87% accurate patient data to process billing, 13% of the data cannot guarantee you correct billing. In another example, a 52% complete customer data set implies lower confidence in the planned campaign reaching the right target segment. You can define the acceptable levels of scores for building more trust in data.

Data quality dimensions serve as a guide for selecting the most suitable dataset. When presented with two datasets of 79% accuracy and 92% accuracy, analysts can choose the dataset with higher accuracy to ensure that their analysis has a more trusted foundation.

What are the 6 dimensions of data quality?

You can measure data quality on multiple dimensions with equal or varying weights, and typically the following six key dimensions are used.

1. Completeness

This dimension can cover a variety of attributes depending on the entity. For customer data, it shows the minimum information essential for a productive engagement. For example, if the customer address includes an optional landmark attribute, data can be considered complete even when the landmark information is missing.

For products or services, completeness can suggest vital attributes that help customers compare and choose. If a product description does not include any delivery estimate, it is not complete. Financial products often include historical performance details for customers to assess alignment with their requirements. Completeness measures if the data is sufficient to deliver meaningful inferences and decisions.

2. Accuracy

Data accuracy is the level to which data represents the real-world scenario and confirms with a verifiable source. Accuracy of data ensures that the associated real-world entities can participate as planned. An accurate phone number of an employee guarantees that the employee is always reachable. Inaccurate birth details, on the other hand, can deprive the employee of certain benefits.

Measuring data accuracy requires verification with authentic references such as birth records or with the actual entity. In some cases, testing can assure the accuracy of data. For example, you can verify customer bank details against a certificate from the bank, or by processing a transaction. Accuracy of data is highly impacted on how data is preserved through its entire journey, and successful data governance can promote this dimension of data quality.

High data accuracy can power factually correct reporting and trusted business outcomes. Accuracy is very critical for highly regulated industries such as healthcare and finance.

3. Consistency

This dimension represents if the same information stored and used at multiple instances matches. It is expressed as the percent of matched values across various records. Data consistency ensures that analytics correctly capture and leverage the value of data.

Consistency is difficult to assess and requires planned testing across multiple data sets. If one enterprise system uses a customer phone number with international code separately, and another system uses prefixed international code, these formatting inconsistencies can be resolved quickly. However, if the underlying information itself is inconsistent, resolving may require verification with another source. For example, if a patient record puts the date of birth as May 1st, and another record shows it as June 1st, you may first need to assess the accuracy of data from both sources. Data consistency is often associated with data accuracy, and any data set scoring high on both will be a high-quality data set.

4. Validity

This dimension signifies that the value attributes are available for aligning with the specific domain or requirement. For example, ZIP codes are valid if they contain the correct characters for the region. In a calendar, months are valid if they match the standard global names. Using business rules is a systematic approach to assess the validity of data.

Any invalid data will affect the completeness of data. You can define rules to ignore or resolve the invalid data for ensuring completeness.

5. Uniqueness

This dimension indicates if it is a single recorded instance in the data set used. Uniqueness is the most critical dimension for ensuring no duplication or overlaps. Data uniqueness is measured against all records within a data set or across data sets. A high uniqueness score assures minimized duplicates or overlaps, building trust in data and analysis.

Identifying overlaps can help in maintaining uniqueness, while data cleansing and deduplication can remediate the duplicated records. Unique customer profiles go a long way in offensive and defensive strategies for customer engagement. Data uniqueness also improves data governance and speeds up compliance.

6. Integrity

Data journey and transformation across systems can affect its attribute relationships. Integrity indicates that the attributes are maintained correctly, even as data gets stored and used in diverse systems. Data integrity ensures that all enterprise data can be traced and connected.

Data integrity affects relationships. For example, a customer profile includes the customer name and one or more customer addresses. In case one customer address loses its integrity at some stage in the data journey, the related customer profile can become incomplete and invalid.

While you regularly come across these six data quality dimensions, many more dimensions are available to represent distinctive attributes of data. Based on the context, you can also consider data conformity to standards (do data values comply with the specified formats?) for determining data qualityData quality is multi-dimensional and closely linked with data intelligence, representing how your organization understands and uses data.

Measuring data quality dimensions helps you identify the opportunities to improve data quality. With adaptive rules and a continuous ML-based approach, predictive data quality brings you trusted data to drive real-time, consistent, innovative business decisions.

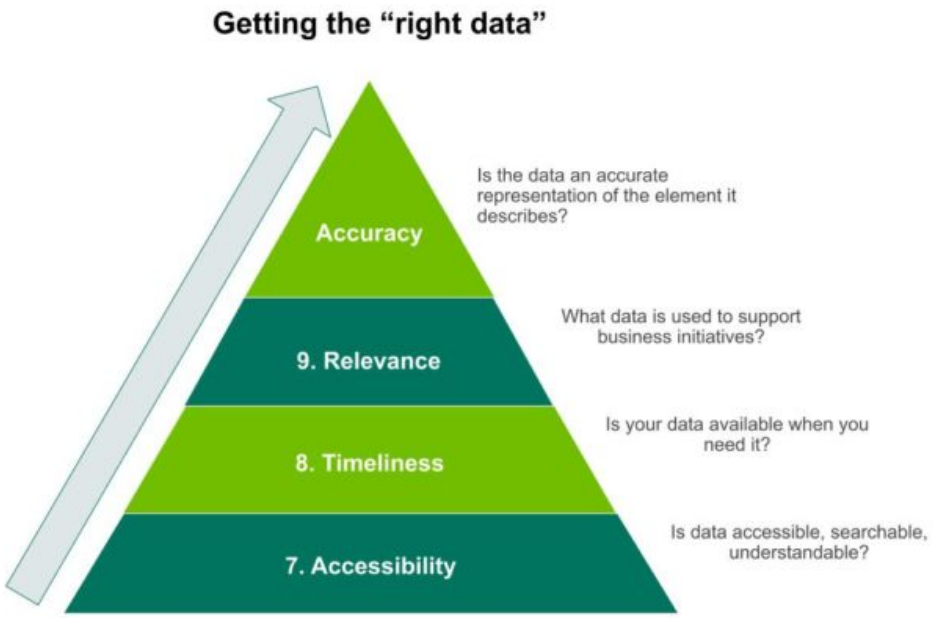
Beyond accuracy: What data quality means to data consumers

Data quality from the perspective of data producers and managers focuses mostly on accuracy. Matching data as closely as possible to the real-world entity is their goal. Their data cleaning, fixing, and management efforts are directed towards improving data accuracy.

From data consumers' perspective, we should add three more dimensions to data quality. When data consumers shop for quality data, their challenges are more oriented to the supply-chain of data. Their first requirement is data **accessibility**. They want to know where data resides and how to retrieve it.

Their next concern is **timeliness**. The value of data is in using it. Accessible data has no value if it is not available for timely use. Timeliness defines if the data is available when required. Trusted data available in real-time or near real-time can reduce errors and streamline operational processes. Timely data availability can drive successful business innovation and maintain a competitive edge.

Data consumers want to access data when they want, and they want the most recent data to power their projects.



Source: Gartner (Aug 2020) – Data Quality Fundamentals for Data & Analytics Technical Professionals

Once data consumers are satisfied with data accessibility and timeliness, their focus shifts to **relevance**. They want to shop for data that correctly aligns with their requirements. They do not want to waste their efforts on data that is not immediately relevant to their planned projects. Only then comes data accuracy, which ensures that the selected data will correctly deliver the results.

Going beyond accuracy, data producers and consumers jointly need to evolve a strategy that rethinks data quality. Data consumers must define what's most important and creators must focus on delivering that most important data. They need to assess the factors impacting effective data shopping, and ask the following questions:

- Is the data well understood?
- Is it driven by data intelligence?
- Does the data have sufficient metadata to understand how they can use data to power their specific analysis?
- Can they access data lineage as the data moves between sources and goes through aggregations, manipulations and transformations?

Only then the data quality can be addressed successfully and improved continuously.

Ankur is a passionate data-driven marketer and a storyteller who loves helping businesses achieve growth and excellence. He holds an MBA from Cornell and engineering from Indian Institute of Technology Delhi.