denologix
INFORMATION MANAGEMENT

data quality FOR BIG DATA

Principles Remain, But Tactics Change

Data quality is a key priority in an information management program - poor quality data negatively affects business outcomes. However, IT leaders must rethink data governance and quality in the context of "big data."

Analyst(s): Roxane Edjlali | Ted Friedman

OVERVIEW

Extreme information challenges affect data governance and quality programs. Some existing policies and practices must be revisited to adapt to "big data;" many others (such as the need for business involvement) continue to be valid. Enterprise information architects, information managers and data management and integration leaders must modify data quality and governance practices when dealing with the challenges presented by big data.

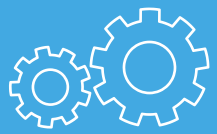
Key Findings

Data quality in the context of big data is driven by the use case. Use cases will not all have the same level of quality expectation: clickstream analysis and intrusion detection do not require the same level of precision.

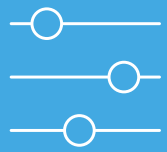
Big data project implementations require a data scientist role. Business users owning data quality initiatives will need the support of data scientists to define the data quality for big data.

As with other data quality improvement initiatives, data quality in the context of big data needs to be owned by the business.

Recommendations



Organizations that are considering using data they do not own will need to establish some degree of confidence in the data before it is leveraged.



As with regular data quality improvement projects, organizations will need to balance the fit of the data for each use case, the ability to reuse and the consistency aspects.

Assess the sources considered and evaluate the scope of governance that is applicable— particularly for external sources.

Include data scientists as part of the data quality team, working closely with data stewards.

What You Need to Know

Big data starts with the quantification of very large data volumes, but it goes beyond volume and also includes velocity, variety and complexity (see Figure 1).

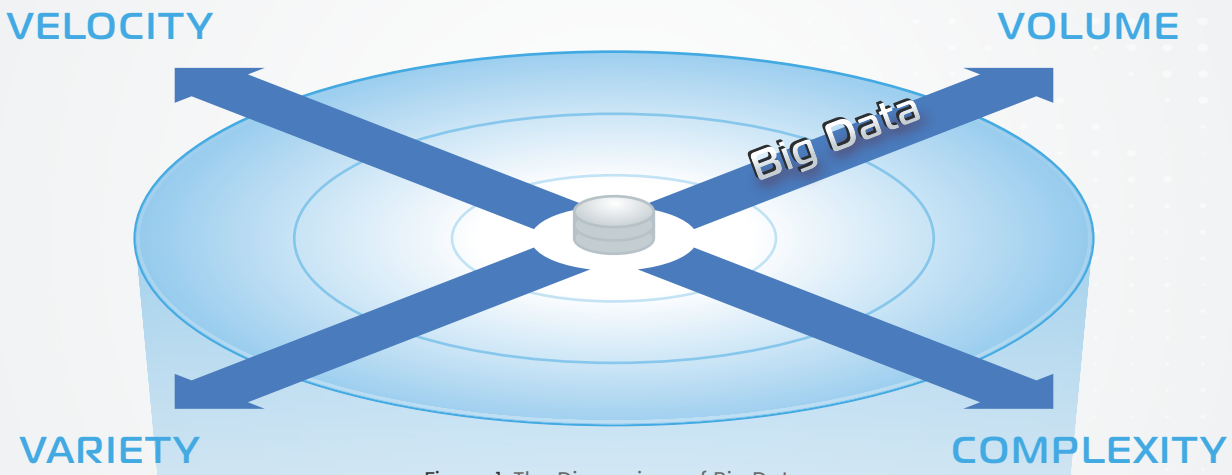


Figure 1. The Dimensions of Big Data
Source: Gartner

Each of the dimensions of big data/extreme information management affects the traditional perception and management of what quality means. In this report we have chosen to focus on the implications of volume on what data quality means. Variety, velocity and complexity will be addressed separately (see "'Big Data' Is Only the Beginning of Extreme Information Management").

ANALYSIS

We are entering a new era where the data sources that are being considered as part of the enterprise spectrum of information assets involve new sources of information, such as social media, clickstream data and external data sources such as Dun & Bradstreet. This means rethinking what data quality means and how to adapt it to these new sources of information.

Traditional data quality criteria assumed a fine-grained approach to data quality that required punctual manual intervention to manage exceptions, up to the record level if needed.

Clients would quote the following dimensions as being key to data quality:

- Completeness: identifying data elements that are required.
- Timeliness: providing data of an acceptable level of freshness.
- Accuracy: verifying that the data respects data accuracy rules.
- Adherence to a common language: data elements fulfill the requirements expressed in plain business language.
- Consistency: verifying that the data from multiple systems respects the data consistency rules.
- Technical conformance: meeting the data specification and information architecture guidelines.

Organizations looking at including big data sources as part of their information infrastructure need to define data quality in the context of big data. Traditional data quality approaches fail to adapt to the following constraints:

- Compared with data that has been manually captured, very large volumes of information that is machine-captured do not require the same granularity for data validation rules. Data does not need to be validated to verify potential user errors in collecting it. Instead, data quality should be looking for exceptions that give an indication of the validity of the data collection process. In the case of metering data, data quality would rather focus on detecting missing data (indicating that the device is out of service) or outliers (indicating either an abnormal situation or an issue with the device).
- Data is of a much finer grain. Examples include clickstream or metering data. As a result, each individual data value may not need to be checked for validity. In the context of big data, every single data point does not need to be cleansed. Data is considered as a whole instead of a collection of separate records. Managing data quality at the record level would not be practically possible given the volume of data. New approaches to data quality are needed that are more aligned with the use case.
- Data is not owned by the enterprise. For example, social media data used for sentiment analysis is not owned by the enterprise, and it can prove extremely challenging to define data validation rules, given the lack of control over who, where and when the data is being produced. The information collected on social websites is collected without any consideration of the use case and with no control over it. As a result, defining the validity, perishability and fidelity of the information is made even more challenging.

ANALYSIS

Because of the very different nature of the data, traditional data quality criteria will need to be revisited to address big data issues. Completeness, timeliness, accuracy and consistency will need to be adapted to the context of big data. Gartner's anecdotal research with clients has led to the following strategies for data quality in the context of big data.

"Good enough" data quality

Organizations need to adapt data quality to the use case, considering the data as a whole instead of looking at it record by record. Completeness, accuracy, consistency and timeliness will be considered for the whole dataset in support of the use case.

For example, in clickstream analysis, the objective is to optimize user retention and understand where users drop off. Far less important is carefully verifying the quality of the user data. However, it may be necessary to remove part of the noise. Removing noise in this particular example could be removing Web robots' interactions with a website and retaining only genuine users' interactions. Web robots will have a very different usage pattern from real users and, as such, will be detectable, allowing the data with the Web robots' signatures to be removed. At the other end of the spectrum, fraud detection will require finer data quality analysis. It will not be enough to split users into two categories: the real users and Web robots or other applications that access the website. The analysis will need to identify users sharing the same account, Web robots trying to break into user accounts, or unusual transactions. In this use case, data validation is of much greater importance and will need to deal with much greater granularity.

In the examples above, data accuracy can mean two different things depending on the use case.

Another specific concern with big data is perishability. Datasets can be highly perishable, with completely new datasets being considered on a daily basis, or more frequently. Tweet analysis, for example, will only cover tweets on a subject from a few hours after an event; Web log data may only be meaningful for a day; geolocation data for mobile users will only be valid for a very short time. Again, organizations will need to articulate how much data quality work can be done on perishable data. Organizations may have to decide if the data can be used as is.

In conclusion, organizations considering data quality in the context of big data should not attempt to overdo it, but should identify data of "good enough" quality, as required by the use case.

ANALYSIS

Data quality projects need to become much better at clearly assessing the necessary level of data quality

Traditional data quality approaches are overwhelmed with the volume of data involved in big data projects. For example, performing data profiling on the full dataset will be a very lengthy and resource-intensive process. Organizations need to become much more effective at limiting the scope of the data involved in data quality. A number of techniques can be used to reduce the scope of data quality — for example, identifying only the attributes to focus on, or identifying correlations between objects and retaining only one. Sometimes even limiting the attributes may not be restrictive enough. Managing data quality at a coarser level of granularity may be a better option. Identifying a coarser grain of data can be done by, for example, maintaining the fields and values that are the most frequent along the bell curve, or by discovering relations across two separate objects, looking at the correlation between the objects and retaining only one (one party per household, for example). As a result, the data to be considered can be greatly reduced (one client indicated a ratio of 1:1 million).

In the examples above, the data validation is driven by the use case. The use case acts as a series of funnels that parses out the data validity specifically for the use case. Taking a use case by use case approach to data validation prevents reuse and consistency issues across use cases. Going back to the clickstream example — how a "real" user is defined in the context of clickstream analysis may not correspond to a "real user" in the context of fraud detection, even though both use cases use the same data. The criteria selected to qualify the fitness of the data for the use case will lead to different data semantics. As a result, the ability to adapt the resulting dataset for wider use comes as a second consideration and may not be possible. When the aggregated view of data is stored, use case by use case data quality may lead to redundant work and inconsistent semantics, and even inconsistent or redundant data.

As with regular data quality improvement projects, organizations will need to balance the fit of the data for each use case, the ability to reuse the data and the consistency issues.

Organizations need to get better at identifying and compensating for data quality issues in data they do not own

Organizations have consumed and processed data from outside their enterprises for quite a long time. Examples include the receipt and translation of electronic data interchange transactions from trading partners as part of the supply chain operations of the business, or aggregation of point-of-sale data from franchisees in order to analyze consumer preferences and purchasing patterns. With the data flow for key business processes beginning (and possibly also ending) outside the organization's control, it is critical to establish data quality controls that measure, validate and ensure conformance with expectations about syntax, semantics and the fitness for purpose of the data. Failure to do so creates the risk of damage to internal operations (or the operations of a downstream partner) due to data quality flaws. Fortunately, organizations are able to create the necessary controls because the data in

ANALYSIS

in question is well understood and the expectations for quality are generally well known.

The big data phenomenon changes the game dramatically. Many of the emerging data sources that offer huge promise, primarily for analytical purposes, also bring extreme challenges, exactly because the structure and meaning are often not well understood, and expectations about their quality have not been established. The "fidelity" of the data for use in a new and different context may be completely unclear. For example, there is rapidly growing interest in leveraging data from social media (social networking websites and so on) to amplify the insight into the way consumers feel about products and services. However, because of the open nature of the environment, the creation of this data is largely ungoverned, so accuracy is highly questionable. The power of such massive amounts of information can be substantially degraded, if not completely destroyed, by quality issues, yet organizations consuming this data may not be aware of the degree of degradation. Consumers of such big data sources will need to develop techniques for verifying such data — perhaps by linking it to existing data known to be accurate (such as prior purchase history relative to similar products and services). Some degree of confidence in the data must be established before it is leveraged for the use case in question.

ANALYSIS

As for any other data quality initiative, business involvement and aligning with business strategy remain essential criteria for success

The big data phenomenon makes it even more important for organizations to align their governance efforts in support of the business strategy. In many respects, the best practices for successfully implementing data quality remain valid in the context of big data. While the need for sponsorship and business involvement is unchanged, new roles such as data scientists are required. Data scientists combine expertise in mathematics-based semantics in computer science with knowledge of the physics of digital systems. They will be best placed to work with the domain experts (the data stewards) to discover relationships within the data. For example, assessing the value of the various attributes by analyzing access frequency, detecting outliers or discovering correlations between attributes would be the initial stages in understanding data distribution.

As in traditional data quality initiatives, data stewards will continue to play the role of data champions, monitoring sudden changes in the data and, possibly in partnership with the data scientists, performing root cause analysis. For example, if data stewards see a sudden drop in user interaction on a website when no business justification seems to support it, this may suggest that the issue is with the instrumentation of the Web logs.

Understanding data quality for big data is a journey

Understanding data quality for large volumes of data can appear to be challenging at first. Organizations implementing data quality in big data initiatives have indicated that, as a first stage, monitoring data, identifying outliers using simple statistical methods, can help set the baseline for what to expect. Using this baseline as a starting point, organizations can further refine their analysis over time, including additional criteria (such as seasonality) as part of their models. Here again, just as for traditional data quality initiatives, data quality needs to be considered as a program rather than a project, and needs to have both the business and IT working together to make it progress.

denologix

INFORMATION MANAGEMENT

data911@denologix.com 1 800 393 1203 denologix.com