JMSBI, Volume 1, Issue 1,
October. 2016, pages 12-20
http://www.ibii-us.org/Journals/JMSBI/
ISBN 2472-9264 (Online), 2472-9256 (Print)
DOI https://doi.org/10.5281/zenodo.376753

Big Data Has Unique Needs for Information Governance and Data Quality

Charles A. Mathes

Leadership Studies Department, Our Lady of the Lake University, San Antonio, TX, USA
cmathes@lake.ollusa.edu

Received: May 16, 2016; revised: September 19, 2016; published: October 07, 2016

ABSTRACT

Enterprises that are venturing into the technical environment of big data and are attempting to create a data lake environment need to take precautions. The principles of information governance and data quality need to be applied to the new world of big data to avoid the trap of the data lake turning into a data swamp. Applying the seven V's of big data as foundational principles for information governance and data quality will help ensure the long-term success of the expertise big data environment.

Keywords: Big data; Data lake, Seven V's; Information governance; Data quality

1. INTRODUCTION

Data is growing at an astonishing rate. As individuals, small organizations, and all the way to global enterprises generate, use, and store data. The types and sources of data continue to expand and convert to digital media. From wearables to appliances to industrial equipment are becoming sensor enabled, all of these devices are fueling the growth of data. Increasingly business transactions, social interactions, and entertainment are becoming digitally driven. The mountain of data will either be turned into usable information or lose its value and turn into dark data. As enterprises embark on their big data adventure, there is a desire to capture and retain the value inherited with insights. Data lakes are a method for economically storing massive amounts of data but in order for the data to be useful as information, thereby an asset, the integrity of the data must be maintained and a means to retrieve the data out of the data lake needs to be planed before investing time and effort to populate the data lake.

2. BIG DATA ENVIROMENT

There has been established industry leading practices and standards for the care and feeding of enterprise data for an extended period of time. Data Management Association (DAMA) created a framework guide in 2009 called Data Management Body of Knowledge (DMBOK). This frame work has data governance as the center of activity and the central knowledge domain that connects all the other domains.

Fig. 1 DAMA DMBOK Guide Knowledge Area Wheel

- Data Architecture Management
- Data Development
- Data Operations Management
- Data Security Management
- Data Integration and Interoperability
- Document and Content Management
- Reference and Master Data
- Data Warehousing and Business Intelligence
- Metadata Management
- Data Quality Management

From the perspective of an enterprise with a more traditional business model and use of structured data, frameworks such as DAMA DMBOK gave appropriate guidance from simple to highly complex environments. But now, the volume, types, and speed of data are challenging how data professionals address the needs to maintain the integrity of data in the new big data era. Yet the underlining principles remain consistent.

Fig. 2 Image of stylized Data Lake

In the new big data era, there are a variety of technologies and vendors that foster the creating, using, and storing of data in a wide variety of formats. The structure around these repositories are referred to as Modern Data Architecture (MDA) and the large data stores are known as Data Lakes. Their primary function is to store large amounts of data in a financially economic way. The incoming flows of data into the data lake can have a variety of formats. Structured data is information that can be store in rows and columns. This is the traditional data format that existing infrastructure and tools have been developed over time. Unstructured data is typically under-organized and does not fit well in the more traditional tools. Unstructured data

can exist in multiple general formats such as e-mails, PDF documents, sensor feeds, images, audio, and video as well as other general formats.

One of the issues that have plagued Information Technology (IT) and data management professions as well as the end consumer of data alike is that there is a Pareto relationship of around 80% of the time and effort is spent on prepare the data to be usable and around 20% of the time and effort is spent actually using the data. Therefore, there are opportunities to apply principles to improve the effectiveness and efficiencies of data storage and retrieval.

3 SEVEN FOUNDATIONAL PRINCIPLES OF BIG DATA

The three "V" words commonly used to describe Big Data – volume, velocity, and variety – define the proportional dimensions and challenges specific to big data but fail to fully describe the whole concept of Big Data. The other "V" s are aspirational qualities of all data and provide a more complete picture to describe the attributed of big data and what is necessary to maintain the integrity of the data as well as be able to leverage data as a value generating asset are summarized in the seven foundation V's of big data.

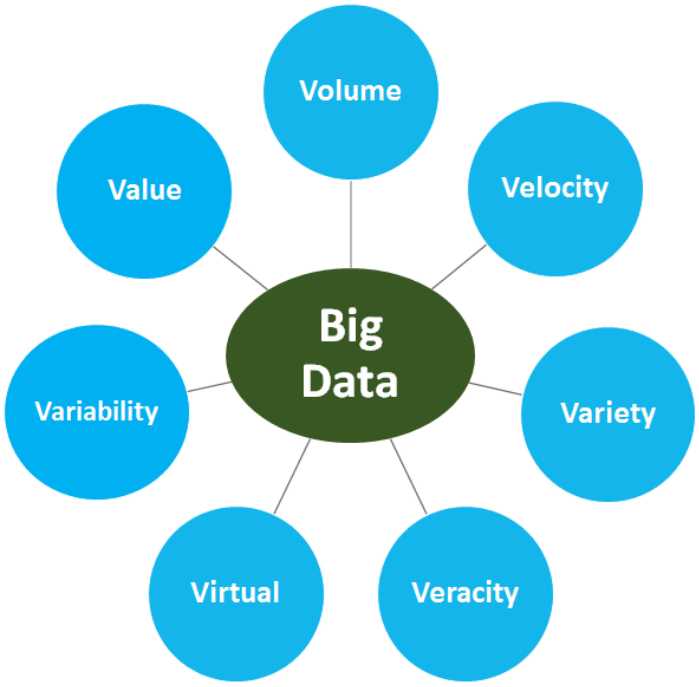


Fig. 3 Seven V's of Big Data

- Volume
- Velocity
- Variety
- Veracity
- Virtual
- Variability
- Value

Companies are increasingly turning to Big Data as a means of better using structured and unstructured data generated by operations not only to enhance safety, efficiency and productivity, but to predict events before they happen.

Having a data quality issue is much more than just an inconvenience, missing or misleading data can be very expensive and can even cost lives. "Poor data can cost businesses 20%–35% of their operating revenue", Chad Luckie May 2012.

Many of these principles complement each other and it is common for data to experience two or more of these principles at the same time and as data goes through its life cycle, the principles describing the data may change as well as the governing needs of the data.

3.1 Volume

Volume is the scale of data. While big data is not all about the size of data, the growth in the volume in data is impressive. According to International Data Corporation (IDC) estimates, by 2020, business transactions on the internet business-to-business and business-to-consumer

will reach 450 billion per day. And according to IBM on their big data blog, over 90% of all the data in the world has been created in the past two years. The size of data use to be measured in megabytes, now data is being measured in terms of exabytes (1,000,000,000,000,000,000 bytes) and zettabytes (1,000,000,000,000,000,000,000 bytes). The Industrial Internet of Things (IIoT) is a source of sensor data. Dylan Twenty reports that jet engines from GE generate 500 gigabytes of data during every flight. This data is store for analysis of the health of the engines. Sensor and device data feeds are the major contributor of digital data growth yet social media, Voice over Internet Protocol (VoIP), and enterprise data are all contributing to the massive growth of digital data. If data is not properly sorted and tagged, then the usefulness and value of the data is dramatically reduced.

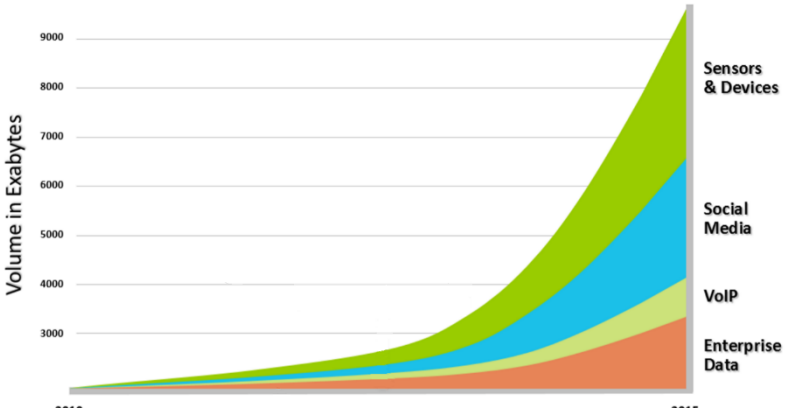


Fig. 4: Volume - digital data growth chart - Di Como

3.2 Velocity

Velocity is the rate of change that the data experiences. The velocity of data is described as data at rest, data in use, and data in motion. Data at rest is typically associated with master data, archived data and other data sources that are static. Data at rest is data that is not changing. Data in use is typically associated with transactional data.

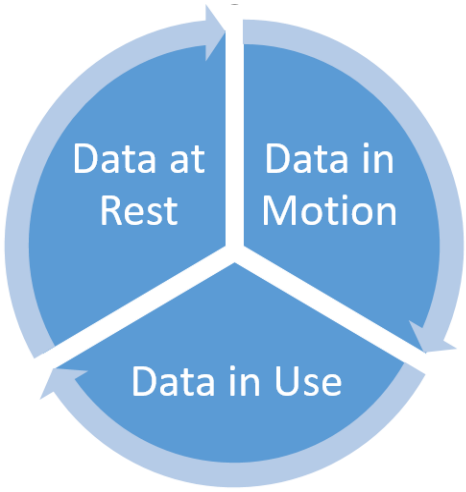


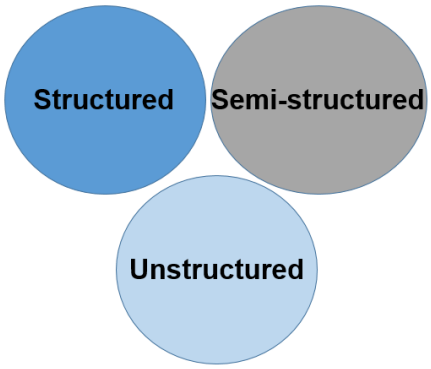
Fig. 5 Velocity

An example of transactional data would be the interconnected processes of a sales order. The financial transaction that occurs with the financial institution, inventory updates and the corresponding inventory checking in the warehouse and material requirements planning (MRP), and the delivery process. Data in motion is the movement of data from one application to another application. An example of data in motion could be back-up and archiving, retrieving data from one application to another in order to complete a transaction, or sensor data flowing to the primary repository for processing. Depending on the use case, some data in motion can be processed near the source, thus diminishing the need to move the majority of the data. This is referred to as data on the edge. Each type of velocity has its own governance requirements.

3.3 Variety

Variety is various forms that data can take. It is common to think of data variety as structured data, unstructured data, and semi-structured data. Structured data is the more traditional

Fig. 6 Variety



enterprise data that fits into rows and columns. Structured data is easily stored in databases and there is a wide array of tools available to retrieve date from the databases. Unstructured data is data that does not fit neatly into rows and columns. Unstructured data can exist in multiple general formats such as e-mails, PDF documents, sensor feeds, images, audio, and video as well as other general formats. Unstructured data is more difficult to classify and the current commercial tools to retrieve unstructured data are still early in their maturity lifecycle. Semi-structured data is unstructured data that has identifying tags called metadata to help identify the data for later retrieval. "Metadata summarize basic information about data which can make finding and working with particular instances of data easier." Keith Holdaway 2014.

3.4 Veracity

Veracity is correctness or accuracy of the data along with the context of the data that leads to trust. Three aspects of veracity are data lineage, traceability, and integrity. Data lineage is knowing the source of the data. If the data comes from a trusted source such as the enterprise accounting system that has controls built into it, then the data itself is more trusted. Traceability is the ability to accurately trace where the data came from. If the report is generated out of the corporate data lake but the individual data elements have their source from a trusted system and can be shown that the data elements have not been modified, then the trust of the source system in inherited. On the other hand, if the source of data is coming from an individual's spreadsheet that were updated on their workstation, then a compelling argument can be made that the data should not be trusted.

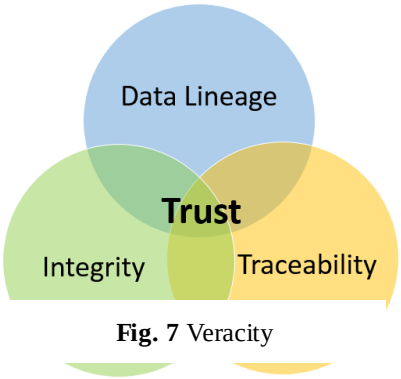


Fig. 7 Veracity

3.5 Virtual

Virtualization is extending the applications and their respective data sources to an abstraction layer so the data from disparate systems appears as a unified table. There are five patterns of data virtualization use. Data federation, data warehouse extension, enterprise data sharing, real-time enterprise data, and cloud data integration. Data federation is used when there are multiple and comparable source applications. Data federation is useful for creating federated views, data services, data mash-ups, caches, virtual data marts, and virtual operational stores. Data warehouse extension are useful for Master Data Management (MDM) hub extensions, data warehouse federation, hub and virtual spoke, complementing Extract Transform Load (ETL) interfaces, data warehouse prototyping, and data warehouse migrations. Enterprise data sharing is useful for shared data services, creating a data abstraction layer, standard-compliance data services, and data virtualization competency center. Real-time enterprise data is for the fast paced enterprise that needs real-time access to their data. Cloud data integration for access and delivery of data to the cloud.

Before virtualizations, traditional techniques were to move the date from disparate systems to the application for processing. There are now tools and techniques maturing in the marketplace to allow a compostable landscape, where logic is brought to the data.

3.6 Variability

Variability refers to data whose meaning is constantly changing. Words do not have static definitions, and their meaning can vary wildly in context. Individual words without context can be very misleading. Same can be true for a numeric stream. Data flows can be highly inconsistent with periodic peaks and valleys. For example, is the variance due to seasonal trends or is there a true anomaly indicating a problem? When the data set is divergent from the average or mean or in other words, the data is outside the range of normal taken in context of the previous V's. The four measures of variability that are commonly used are range, mean, variance and standard deviation. These variability measurements can occur in any of the previous V's and can be depended on their stage in the data lifecycle. Establishing variability standards allows for using tools to monitor the data and manage the exceptions when they occur. Historical context can be used to generate regression analysis for insights on correlations or Principle Components Analysis (PCA) to help avoid the trap of making decisions on spurious relationships, and Monte Carlo simulations can lead to great insights and predictions.

3.7 Value

Value is the accumulation of applying both tactical and strategic governance to big data in the previous 6 "V's" Enterprises commonly treat data as a cost but should treat data as one of the most valuable asset of the enterprise. Trusted information can be used for descriptive, predictive, and prescriptive analytics. Descriptive analytic is reporting on data generated events that occurred in the past. Dashboards are a common use of descriptive analytic. Predictive analytic uses data generating events that had occurred and makes a prediction on what is going to occur next. Analytics such as regression analysis are used for prediction. It is important to note that a co-variance in variables that show relationships does not imply causation. Prescriptive analytics is the next step from predictive analytic. Prescriptive analytics uses data generating events that had occurred and makes a prediction on what is going to occur next and provide an advice on how to react to the prediction. Correct, complete, and trusted data will enhance people, process, and technology.

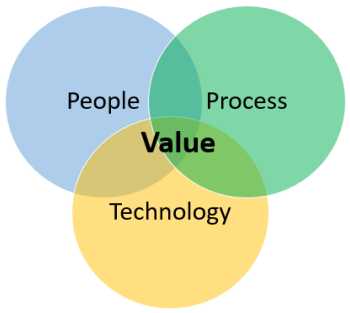


Fig.8 Value

4 INPUT PROCESSING OUTPUT

Every process regardless if it is automated or manual, an over-arching macro process or low-level details micro process, all contain the basic components of Input, Process, Output. And if it is governed, then it has controls and mechanism. In the context of computer systems,

inputs are the data feeds that go into the computer program. The inputs can be as simple as a person typing on the keyboard or as complex as sensor data off of a jet engine. The higher the quality of the inputs the better. Processing is the computer program that takes the inputs and applies logic to the date and makes some kind of decision. Outputs are the results of the logic and the decisions that were made. If there is information governance in place, then the controls are the rules and standards. Mechanisms are the way to apply and enforce the controls. If the process is more sophisticated, then there can be a feedback loop for continuous learning.

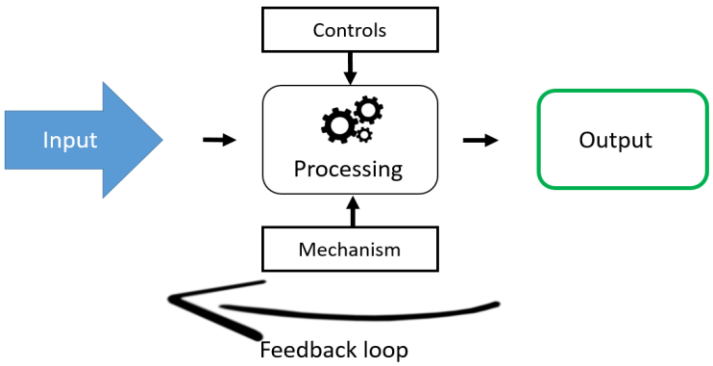


Fig.9 Input Process Output

5 METAPHOR FOR DATA QUALITY

The following metaphor is to illustrate that making decision with partial information without considering the bigger picture can lead to inaccurate conclusions.

5.1 Anscombe's Quartet

The following table contains 11 observations from four different datasets. Each dataset is statistically very similar. When statistical analysis is applied, the common measures of R^2, Means, and P values are identical for the four data sets up to three significant digits, strongly indicating that these data sets represent the same values.

	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Table 1. Anscombe's Quartet

But with further investigation and visualizing the data, it become obvious that these four data sets are distinctive and unique after all.

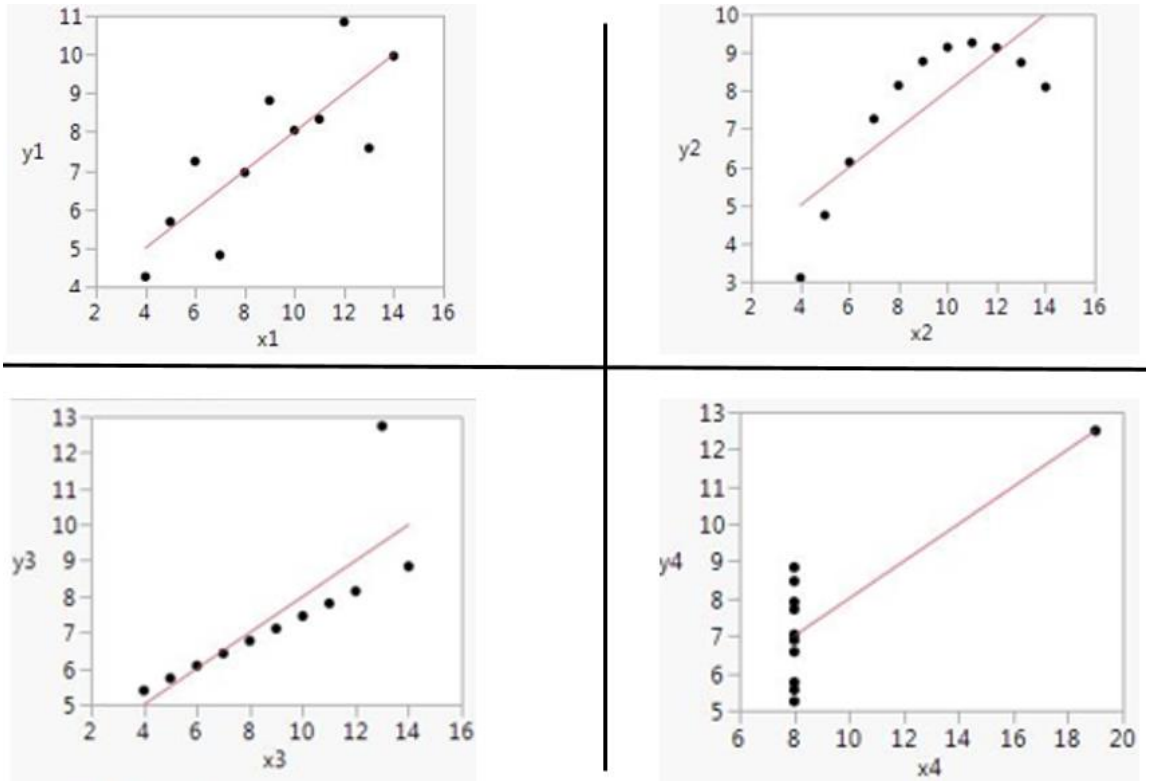


Fig. 10 Results of Anscombe's Quartet (Scatter Plot from SAS)

Illustrating the point that ungoverned and incomplete data can lead to inaccurate conclusions.

6 CONCLUSION

Enterprises that leverage and use the data they currently have access to as valuable information used for descriptive, predictive, and prescriptive analytics and treating data as a value generating asset that complements people, process, and technology will have a competitive advantage. The big data era is upon us and the established trend is that the volume of data will continue to grow at an astonishing rate. The fundamental concepts and processes for managing and governing data with the goal of maintain data integrity and usefulness have not changed but the application and details of information governance are going through a significant metamorphosis. Big data introduces challenges that go beyond the proportional properties of volume, velocity, and variety. Information governance also requires management of veracity, virtual, variability, and value. At any point in time during the lifecycle of data, two or more of these seven properties can unite and create unique sets of circumstances and requirements. Information governance and data quality professionals are having to be flexible and adaptive to this new landscape of big data.

It is the suggestion of this author that next steps in the study of information governance in the big data era would be to create an instrument for capturing information governance maturity and effectiveness and conduct an experiment with a convenience sample of mid-sized to large enterprises. The experiment would by necessity have to be quasi experimental due to missing control group.

E(R) O1 X O2 where E(R) is Experiment Result, O1 is first observation, X is the manipulation, O2 is second observation.

Observation 1, conduct survey on information governance maturity and perceived effectiveness of information governance on big data to create a base line observation.
Manipulation, implement big data information governance in the enterprises.

Observation 2, conduct survey on information governance maturity and perceived effectiveness of information governance on big data after implementing information governance to measure if maturing and perceived effectiveness changed.

In an effort to avoid Type II Errors, the hypothesis should include implementing information governance would not have a significant effect (P > .05) on maturity and perceived effectiveness of information governance. The null hypothesis should include that implementing information governance would have a significant effect (P > .05) on maturity and perceived effectiveness of information governance.

REFERENCES

[1] Mosley M., Et al…, The data management body of knowledge, DAMA-DMBOK Guide (Technics), USA, 2009.

[2] Luckie C., http://www.fathomdelivers.com/blog/analytics-and-big-data/big-data-facts-and-statistics-that-will-shock-you/, May 8, 2012

[3] IDC, http://wikibon.org/blog/unstructured-data

[4] IBM, http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[5] Tweney D., http://venturebeat.com/2015/06/18/here-comes-the-industrial-internet-and-enormous-amounts-of-data/

[6] Di Como P., Review of performance evaluation benchmarks of apache Hadoop, https://www.politesi.polimi.it/bitstream/10589/93418/1/PUSTINA_749598_ReviewOfPerformanceEvaluationBenchmarksHadoop.pdf, 2014

[7] Holdaway K., Harness oil and gas big data with analytics (Wiley), USA, p.310, 2014

[8] Kitchin R., The Data Revolution: Big Data, Open Data, Data Infrastructures and their consequences (Sage), USA, 2014