Importance of Data Quality for Analytics

Rajesh Jugulum

Introduction

As we know in this highly regulated environment, there is an ever-increasing need in creating and providing safeguards and tools to increase the transparency and accuracy of information. Such tools and mechanisms also need to satisfy business and regulatory requirements. This situation has significantly increased the role of data and analytics in business in general. Data and analytics capabilities should be viewed in the same way as other resources such as people, facilities, raw materials etc. Therefore, data and analytics capability management aspects have become critical functions in managing overall business and achieving business excellence. In this chapter, we will discuss the importance of data quality to perform high-quality analytics and take appropriate decisions basing on the analytics.

Data and Analytics as Key Resources

Harrington (2006) highlights the importance of managing processes, projects, change, knowledge, and resources for organizational excellence. In addition to these, in this data-driven world, importance of data and analytics capabilities cannot

Rajesh Jugulum, 2002 Feigenbaum Medalist.

R. Jugulum (✉)
Cigna/Northeastern University, Boston, USA
e-mail: rajesh_jugulum@yahoo.com

P. Sampaio and P. Saraiva (eds.), *Quality in the 21st Century*,
DOI 10.1007/978-3-319-21332-3_2

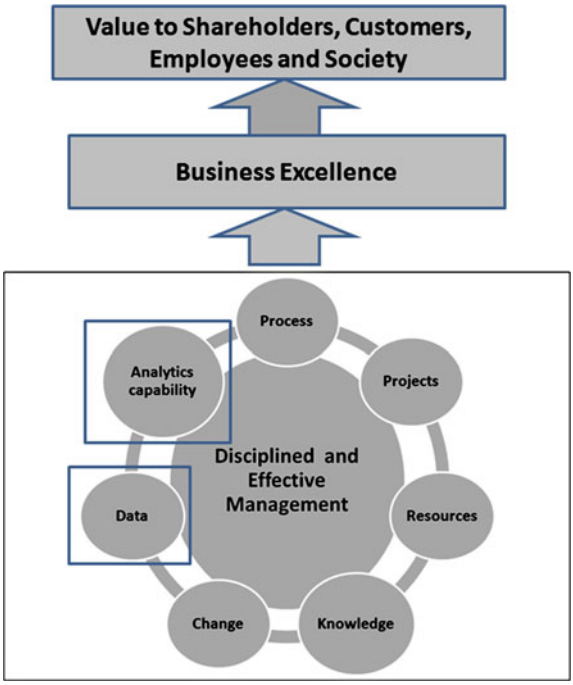

Fig. 1 Seven levers of a disciplined and effective organization

be overlooked. Because of the importance of data and analytics to derive insightful business outcomes, data and analytics capability management aspects have become critical functions in managing overall business and achieving business excellence.

Figure 1 shows seven levers of a disciplined and effective organization. Besides having good processes, projects, resources, great knowledge, and ability to change, we need to have capability to ensure high-quality data and ability to perform high-quality analytics to survive in the global competition and they have to be viewed like any other resources. Therefore, it is important to ensure that we have high-quality data across the organizations to derive meaningful business outcomes.

Importance of quality data was emphasized by famous statisticians much before the data quality field has experienced massive growth both in industry and academics. R.A. Fisher, a famous British statistician, said that the first task of a statistician is to conduct the cross-examination of the data for meaningful analysis of data and interpretation of results. C.R. Rao, a world-renowned Indian statistician provided a checklist (Rao 1997) for cross-examination of the data, where emphasis was primarily given to the data quality and measurement systems that we use for data collection.

Measuring Data Quality

To measure the level of data quality (DQ), we need to select the DQ dimensions of relevance to the specific business process. A DQ dimension, as defined by Wang and Strong (1996), is a set of DQ attributes that represent a single aspect of DQ. In Table 1, we list four core DQ dimensions that are typically used. After defining the DQ dimensional definitions, they are translated into set of business rules to measure the levels of DQ and reliability. DQ business rules help us quantify how good or how bad the data is. By applying the DQ rules to critical data elements (CDEs), we classify them as good or bad in the context of a chosen dimension. A critical data element can be defined as a data attribute that is "critical to success" or required to get the job done. Examples of CDEs are social security numbers, customer ids, data of birth, seniority of claim etc.

Let us assume that the CDE "Seniority of claim" always takes a two-digit value, the valid values of this CDE are restricted and they can be mapped to: Seniority of Claim Code-Senior Secured, Senior Unsecured, Subordinated Secured and Subordinated Unsecured.

We map this business rule to DQ rules for different DQ dimensions:

Data Quality Rule: The CDE "Seniority of claim" takes values from the set {10, 20, 30, 40}. As a result, this CDE is valid and its validity is 100 % as it takes predefined values. This rule maps to the validity dimension of DQ and similarly with other rules we can determine scores for other dimensions. After defining business rules, we perform profiling to understand the behavior patterns of CDEs, and after this step we can calculate DQ scores for CDEs.

Table 1 Four-core data quality dimensions

Dimension	Definition
Completeness	Completeness is defined as a measure of the presence of core source data elements that, exclusive of derived fields, must be present in order to complete a given business process
Conformity	Conformity is defined as a measure of a data element's adherence to required formats (data types, field lengths, value masks, field composition, etc.) as specified in either metadata documentation or external or internal data standards
Validity	Validity is defined as the extent to which data corresponds to reference tables, lists of values from golden sources documented in metadata, value ranges, etc.
Accuracy	Accuracy is defined as a measure of whether the value of a given data element is correct and reflects the real world as viewed by a valid real-world source (e.g., SME, customer, hard-copy record, etc.)

Measurement of Data Quality Scores

Once we have selected the DQ dimensions and measured them using associated DQ rules, the measurement results are called DQ scores. DQ scores are the direct indicators of the performance of the data. A DQ score may reflect the quality of the data at a certain level. In particular, it can be a score for a given DQ dimension of a CDE, an aggregated score of multiple DQ dimensions of a CDE, or even an aggregated score of multiple CDEs (across all related DQ dimensions) at either the taxonomy or business unit level or the enterprise level. A DQ score is a percentage between 0 and 100. It can be generally interpreted as the percent of nondefect data entries out of all data entries.

DQ scores at multiple levels need to be computed in a logical manner. In other words, we cannot get a DQ score at the CDE level without first getting DQ scores at the DQ dimension level. Similarly, we cannot derive a DQ score at the taxonomy or business unit level without first getting DQ scores at the CDE level. This is why we need to determine DQ dimensions and the related DQ rules first. They are used to profile the data and to calculate the DQ scores for different DQ dimensions. Once the dimension level scores are available, DQ scores at the CDE, taxonomy or business unit, and enterprise levels can be derived accordingly. Figure 2 describes the roll-up process that can be used to obtain DQ scores at various levels.

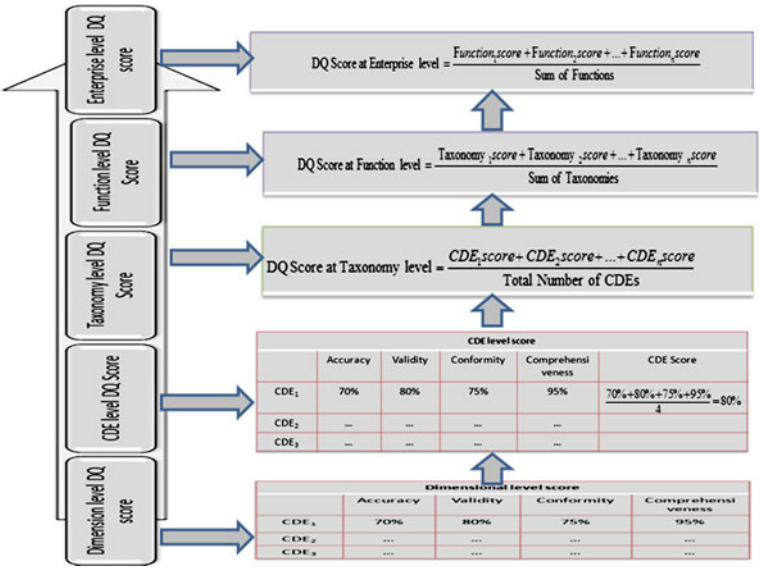

Fig. 2 Data quality scores at various levels

Different Types of Analytics

In this competitive world, the importance of having new insights with data and proper analytics to understand the customer expectations in a much better way has been emphasized by many organizations. There are different types of analytics and they can be chosen depending on the purpose. Gartner (2012) proposed an analytics ascendency model that shows how the value of analytics increases as we expand the capabilities. Table 2 shows different types of analytics including the types that Gartner proposed. Note that the tools associated with these analytics may overlap depending on the case we are addressing. A brief description of all these types of analytics is provided below:

Preparatory analytics: This type of analytics can also be called as cross-examination of data and is useful to evaluate existing DQ levels for variables/CDEs. Techniques like DQ business rules, DQ rule evaluation, and statistical process control are useful to assess the DQ levels.

Descriptive analytics: If we want to know what happened to particular process, operation, facility or CDE, we should perform this type of analytics by using tools like data mining, basic profiling, and descriptive statistics. This type of analytics helps us to understand the performance at a given time point by providing a snapshot with means and standard deviations.

Diagnostic analytics: Diagnostic analytics are performed to know when, where, why, and how a particular problem has occurred. Techniques like correlation analysis, hypothesis test, analysis of variance (ANOVA), and control charts are typically used in this type analytics.

Table 2 Different types of analytics

Purpose	Type of analytics	Tools/techniques
How good is the data?	Preparatory analytics	Data quality rules, data quality scores, statistical process control etc.
What happened?	Descriptive analytics	Basic profiling, data mining, descriptive statistics etc.
Why and when it happened?	Diagnostic analytics	Control charts, analysis of variance, hypothesis tests etc.
How did it happen? (root cause analysis)	Cause-related analytics	Cause and effect analysis, failure mode effect analysis etc.
What will happen?	Predictive analytics	Artificial neural networks, regression analysis etc.
How can we improve?	Prescriptive analytics	Design of experiments, simulations, scenario planning etc.
How confident can we be?	Reliability-based analytics	Failure analysis, confidence intervals, signal-to-noise ratios etc.

Cause-related analytics: Cause-related analytics are usually performed to identify the causes of the problems or failures. Tools like cause and effect diagram, cause and effect matrix, and failure mode effect analysis (FMEA) are used to perform cause-related analytics.

Predictive analytics: Predictive analytics, as the name suggests, are useful in predicting the behavior of a process, system or CDE. Because of this reason, this class of analytics can also be called as "what-if" type of analytics. Techniques like Artificial neural networks and regression analysis are useful to perform predictive analytics. Simulation analysis plays an important role here as it helps simulate various scenarios and perform what-if analysis.

Prescriptive analytics: Prescriptive analytics are useful in answering questions like how we can improve the performance. Tools like designed experiments, simulation analysis, and scenario planning are extremely useful in this class of analytics.

Reliability-based analytics: Reliability-based analytics are typically used in estimating reliability of a product or process or systems or set of models so that we can be more confident about results. With reliability-based analytics, we can assign a confidence level and failure rate for the performance. Failure analysis, confidence intervals and, signal-to-noise ratios etc. are usually used in reliability-based analytics.

It is important to note that the analytics types in Table 2, can be used with numerical, text, voice, web-based, or social media-related data with appropriate transformations/modifications. They can also be used in the context of big data.

Requirements for Executing Analytics

As described above, depending on the purpose we should be selecting appropriate set of analytics. In order to perform the analytics across an organization, it should have "analytics vision" to start with as shown in Fig. 3. Once we have a clear

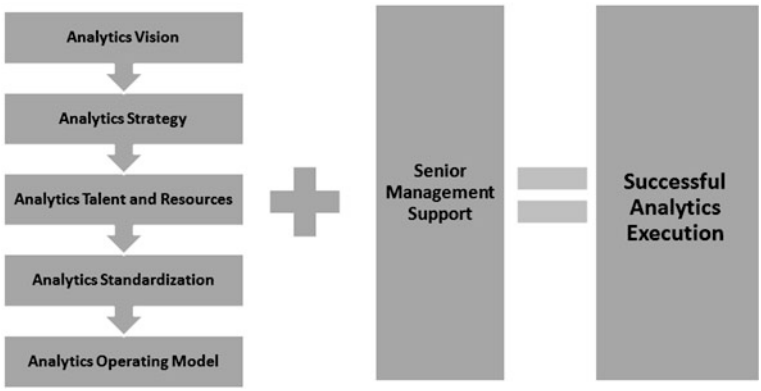

Fig. 3 Successful analytics execution

vision, we can design a suitable strategy for executing analytics. After this step, we should be looking at the talent and resources that we have for analytics. If there are gaps, we should start acquiring great talent and appropriate resources as part of investment strategy on analytics.

When we are planning to expand the use of analytics across the organization it is important to have a standardized approach along with an operating model for successful execution. Standardized approach helps us to use appropriate type of analytics in a given scenario although the tools may vary from application to application. The operating model comes in handy when we want to know how to deploy different types of analytics, associated methodologies, and interpretation of results etc.

In Fig. 3, it is needless to say that the most important requirement for analytics execution is senior management support. All other requirements cannot be fulfilled without the commitment from the senior management. After satisfying analytics requirements, the next stage is to define a process for execution. Next section outlines such a process.

Process of Executing Analytics

The first step in this process is to define the problem (as shown in Fig. 4) and understand purpose. Then we need to collect relevant data and ensure high DQ. After ensuring high DQ, we have to decide what type analytics we need to use based on Table 2. Based on the data and the constraints we have, we can build suitable frameworks/models. In the next step, we need to validate the frameworks/models by including data that was not part of model building activity. After validation step, we should have a plan to deploy these insights in decision-making

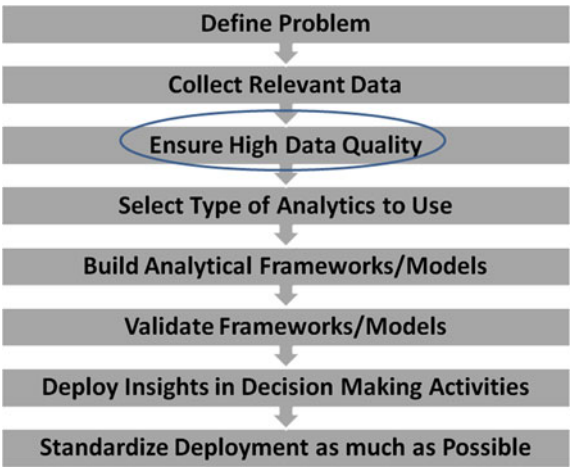

Fig. 4 Process of executing analytics

activities. Note that we need to standardize the method of deploying the insights as much as possible.

In Fig. 4, "ensure high quality data" step is highlighted because high-quality data is absolutely required to run sound analytics and get meaningful business outcomes. Often analytics fail because of poor-quality data and the loss associated with poor-quality data can be quite significant. A combination of high-quality data and reliable analytics will result in increased levels of customer, regulatory, and shareholder confidence by minimizing societal loss with maximum profits.

Conclusions

The conclusions of this chapter can be summarized as follows:

- Data and analytics capability management aspects have become critical functions in managing overall business and achieve business excellence. They should be viewed in the same way as other resources such as people, facilities, raw materials etc.
- For running high-quality data analytics, it is extremely important to have high-quality data. Therefore, preparatory analytics and cross-examination of data play a significant role.
- Different types of analytics exist and we should choose suitable type depending on the purpose and business requirements.
- Good data coupled with sound analytical techniques are key for organizational success because they provide very important insights and that will help in making sound decisions.

Acknowledgments The author would like to express his gratitude to Chuan Shi for his help and support.

References

Gartner (2012). *Big Data Strategy Components: IT Essentials.* Published on 15th October 2012.

Harrington, James (2006). The Five Pillars of Organizational Excellence, Referenced from Quality Digest (URL: http://www.qualitydigest.com/aug06/articles/05_article.shtml)

Jugulum, Rajesh (2014). *Competing with High Quality Data: Concepts, Tools and Techniques for Building a Successful Approach to Data Quality*, Wiley publication

Rao, C. R. (1997). *Statistics And Truth: Putting Chance To Work*, Wspc publications.

Wang, R. Y. and Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*. 12, 4, 5–33.

Author Biography

Author's Short Bio and Perspectives
Name: Rajesh Jugulum
Residence: Franklin, MA, USA
Education: Ph. D
Current Job: Director of global data strategies at Cigna
Adjunct Faculty, Northeastern University
Previous Jobs: held executive positions in quality related areas at Citi Group and Bank of America. Was also researcher at MIT.
Introduction to Quality: I was introduced to quality field during my Masters Degree at Indian Statistical Institute.
Favorite Definition of Quality: Quality must be measured in relation to loss imparted to society (Dr. Taguchi's definition).
Major Contributions to the Field: In multivariate analysis using principles of quality engineering and in the area of data/information quality.
One Trend Defining the Future of Quality: Data and information quality
Impact of Feigenbaum Medal: It gave me lot of recognition.
Favorite Book on Quality: Introduction to Quality Engineering: Designing Quality Into Products and Processes by Dr. Genichi Taguchi
Three Publications:

1. Genichi Taguchi and Rajesh Jugulum (2002) The Mahalanobis-Taguchi Strategy: A Pattern Technology, John Wiley & sons.
2. R. Jugulum, D. D. Frey. Toward a Taxonomy of Concept Designs for Improved Robustness. Journal of Engineering Design, Vol 18(2), 139–156, 2007.
3. Jugulum, Rajesh (2014). Competing with High Quality Data: Concepts, Tools and Techniques for Building a Successful Approach to Data Quality, Wiley publication.

Plans for the Future: Continue to contribute my part to quality field.
Quality Quote: Data quality and information quality are as important as quality of any product or service and they will be key aspects of big data world.

http://www.springer.com/978-3-319-21331-6

Quality in the 21st Century
Perspectives from ASQ Feigenbaum Medal Winners
Sampaio, P.; Saraiva, P. (Eds.)
2016, XIX, 118 p. 28 illus., 23 illus. in color., Hardcover
ISBN: 978-3-319-21331-6