Implications of Big Data Analytics

P. Joseph Charles[1], R.Suba[2]
[1, 2] Assistant professor, Dept of information technology
[1, 2] Department of information technology
St. Joseph's College, tirchy

Abstract- *Big Data is associated with a new generation of technologies and architectures, which can harness the value of extremely large volumes of very varied data through real time processing and analysis. It involves changes in data types, accumulation speed, and data volume.Size is the first, and at times, the only dimension that leaps out at the mention of big data. This paper attempts to offer a broader definition of big data that captures its other unique and defining characteristics. Academic journals in numerous disciplines, which will benefit from a relevant discussion of big data, have yet to cover the topic. This paper presents a consolidated description of big data by integrating definitions from practitioners and academics. The paper's primary focus is on the analytic methods used for big data. A particular distinguishing feature of this paper is its focus on analytics related to unstructured data, which constitute 95% of big data. This paper also reinforces the need to devise new tools for predictive analytics for structured big data. In this paper, we focus on concepts, methods and analytics used in big data.*

I. INTRODUCTION

This paper documents the basic concepts relating to big data. It attempts to consolidate the hitherto fragmented discourse on what constitutes big data, what metrics define the size and other characteristics of big data, and what tools and technologies exist to harness the potential of big data. In recent decades, the increasing importance of data to organizations has led to rapid changes in data collection and management. Traditional information management and data analysis methods ("analytics") are mainly intended to support internal decision processes. They operate with structured data types, existing mainly within the organization. Throughout the history of IT, each generation of organizational data processing and analysis methods acquired a new name. With the launch of Web 2.0, a large amount of valuable business data started being generated beyond the organization by consumers and, generally, by web users. This data can be structured or unstructured, and can come from multiple sources such as social networks, products viewed in virtual stores, information read by sensors, GPS signals from mobile devices, IP addresses, cookies, bar codes, etc. Some types of data, such as text and voice, have existed for a long time, but their volume is now exacerbated by the Internet and by other digital structures. This brings about a new era in existing technologies for data analysis. It is argued that the explosion in data volume is largely attributed to unstructured data, which partly comes from new sources through the passive behavior of the user (e.g. the case of online search terms or the user's location detected by mobile phone apps). The sudden rise of big data has left many unprepared. In the past, new technological developments first appeared in technical and academic publications. The knowledge and synthesis later seeped into other avenues of knowledge mobilization, including books. The fast evolution of big data technologies and the ready acceptance of the concept by public and private sectors left little time for the discourse to develop and mature in the academic domain. Authors and practitioners leapfrogged to books and other electronic media for immediate and wide circulation of their work on big data. Thus, one finds several books on big data, including Big Data for Dummies, but not enough fundamental discourse in academic publications.

Today, many organizations are collecting, storing, and analyzing massive amounts of data. This data is commonly referred to as "big data" because of its volume, the velocity with which it arrives, and the variety of forms it takes. Big data is creating a new generation of decision support data management. Businesses are recognizing the potential value of this data and are putting the technologies, people, and processes in place to capitalize on the opportunities. A key to deriving value from big data is the use of analytics. Collecting and storing big data creates little value; it is only data infrastructure at this point. It must be analyzed and the results used by decision makers and organizational processes in order to generate value.

Big data and analytics are intertwined, but analytics is not new. Many analytic techniques, such as regression analysis, simulation, and machine learning, have been available for many years. Even the value in analyzing unstructured data such as e-mail and documents has been well understood. What is new is the coming together of advances in computer technology and software, new Sources of data (e.g., social media), and business opportunity. This confluence has created the current interest and opportunities in big data analytics. It is even spawning a new area of practice and study called "data science" that encompasses the techniques, tools, technologies, and processes for making sense out of big data. Given that the discourse on big data is contextualized in predictive analytics frameworks, we discuss how analytics have captured the imaginations of business and government leaders and describe the state-of-practice of a rapidly evolving industry. We also highlight the perils of big data, such as spurious correlation, which have hitherto escaped serious inquiry. The discussion has remained focused on correlation, ignoring the more nuanced and involved discussion on causation. We conclude by highlighting the expected developments to realize in the near future in big data analytics.

II. CONCEPTS IN BIG DATA

In 2011, a report of the International Data Corporation has defined Big Data as "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis" [2]. This definition highlights the most critical Big Data problem: that of uncovering value from data sets of huge dimensions, given the wide variety of the data types and the rapid generation of the data.It is difficult to recall a topic that received so much hype as broadly and as quickly as big data. While barely known a few years ago, big data is one of the most discussed topics in business today across industry sectors. A major reason for creating data warehouses in the 1990s was to store large amounts of data. Back then, a terabyte was considered big data.

2.1 Big Data Characteristics

The characteristics of BD can be synthesized by 3Vs [5]

1. *Volume*: the increase in data volume in enterprise-type systems is caused by the amount of transactions and other traditional data types, as well as by new data types. Too much data becomes a storage problem, but also has a great impact on the complexity of data analysis;

2. *Velocity*: refers to both the speed with which data is produced and that with which it must be processed to meet demand. This involves data flows, the creation of structured records, as well as availability for access and delivery. The speed of data generation, processing and analysis is continuously increasing due to real-time generation processes, requests resulting from combining data flows with business processes, and decision-making processes. The velocity of the data processing must be high, while the processing capacity depends on the type of processing of the data flows;

3. *Variety*: Data come from different data sources. For the first, data can come from both internal and external data source. More importantly, data can come in various format such as transaction and log data from various applications, structured data as database table, semi-structured data such as XML data, unstructured data such as text, images, video streams, audio statement, and more. There is a shift from sole structured data to increasingly more unstructured data or the combination of the two.

However, the three important V's are shown in figure.1

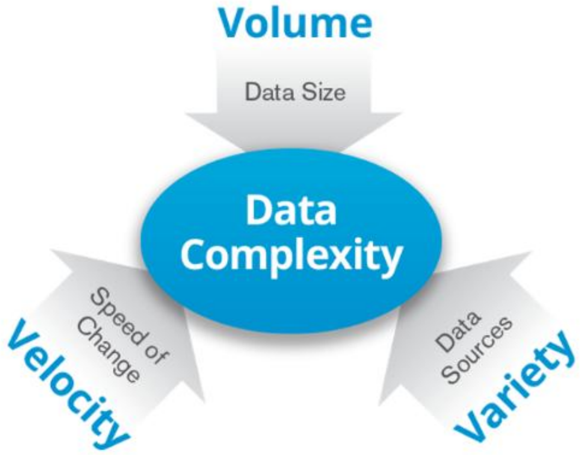

Figure 1: The Three V's of Big Data

III. BIG DATA ANALYTICS

Big Data Analytics (BDA) is a new approach in information management, which provides a set of capabilities for revealing additional value from BD. It is defined as "the process of examining large amounts of data, from a variety of data sources and in different formats, to deliver insights that can enable decisions in real or near real time" [3]. BDA can be used to identify patterns, correlations and anomalies [3], [6]. BDA is a different concept from those of Data Warehouse (DW) or Business Intelligence (BI) systems. Gartner defines a DW as "a storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. The warehouse then combines that data in an aggregate, summary form suitable for enterprise wide data analysis and reporting for predefined business needs" [4]. BI is defined as "a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making" [7].

The stored data does not generate business value, and this is true of traditional databases, data warehouses, and the new technologies such as Hadoop for storing big data. Once the data is appropriately stored, however, it can be analyzed, which can create tremendous value. A variety of analysis technologies, approaches, and products have emerged that are especially applicable to big data, such as in-memory analytics, in-database analytics, and appliances

It is helpful to recognize that the term analytics is not used consistently; it is used in at least three different yet related ways [8].

A starting point for understanding analytics is to explore its roots. Decision support systems (DSS) in the 1970s were the first systems to support decision-making [9]. DSS came to be used as a description for an application and an academic discipline. Over time, additional decision support applications such as executive information systems, online analytical processing (OLAP), and dashboards/scorecards became popular. Then in the 1990s, Howard Dresner, an analyst at Gartner, popularized the term business intelligence. A typical definition is that “BI is a broad category of applications, technologies, and processes for gathering, storing, accessing, and analyzing data to help business users make better decisions” [10]. With this definition, BI can be viewed as an umbrella term for all applications that support decision making, and this is how it is interpreted in industry and, increasingly, in academia. BI evolved from DSS, and one could argue that analytics evolved from BI (at least in terms of terminology).

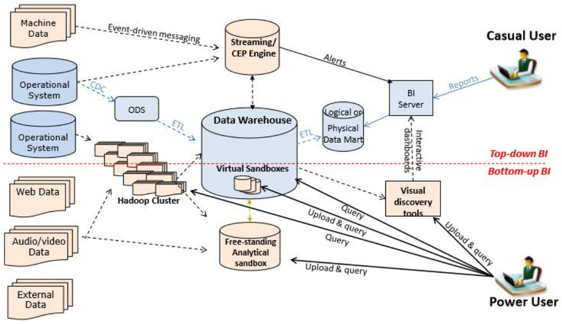

Figure 2: An integrated Analytics Architecture

Thus, analytics is an umbrella term for data analysis applications. BI can also be viewed as “getting data in” (to a data mart or warehouse) and “getting data out” (analyzing the data that is stored). A second interpretation of analytics is that it is the “getting data out” part of BI. The third interpretation is that analytics is the use of “rocket science” algorithms (e.g., machine learning, neural networks) to analyze data. These different takes on analytics do not normally cause much confusion, because the context usually makes the meaning clear. The progression from DSS to BI to analytics is shown in Figure 3.

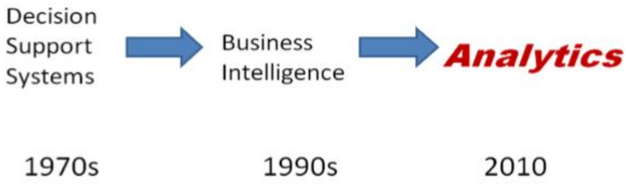

Figure 3: From DSS to BI to Analytics

3.1 Examples of big data Analytics

3.1.1 Introducing a New Coffee Product at Starbucks

Starbucks was introducing a new coffee product but was concerned that customers would find its taste too strong. The morning that the coffee was rolled out, Starbucks monitored blogs, Twitter, and niche coffee forum discussion groups to assess customers’ reactions. By mid-morning, Starbucks discovered that although people liked the taste of the coffee, they thought that it was too expensive. Starbucks lowered the price, and by the end of the day, all of the negative comments had disappeared.

Compare this fast response with a more traditional approach of waiting for the sales reports to come in and noticing that sales are disappointing. A next step might be to run a focus group to discover why. Perhaps in several weeks Starbucks would have discovered the reason and responded by lowering the price.

3.1.2 Drilling for Oil at Chevron

Each drilling miss in the Gulf of Mexico costs Chevron upwards of $100 million. To improve its chances of finding oil, Chevron analyzes 50 terabytes of seismic data. Even with this, the odds of finding oil have been around 1 in 5. In the summer of 2010, because of BP’s Gulf oil spill, the federal government suspended all deep water drilling permits. The geologists at Chevron took this time to seize the opportunity offered by advances in computing power and storage capacity to refine their already advanced computer models. With these enhancements, Chevron has improved the odds of drilling a successful well to nearly 1 in 3, resulting in tremendous cost savings.

IV. CONCLUSIONS

The paper first defined what is meant by big data to consolidate the divergent discourse on big data. Big Data is an opportunity to gain new insights into emerging types of data and content, in order to create more agile businesses and to answer to questions which were previously considered unanswerable. We presented various definitions of big data, highlighting the fact that size is only one dimension of big data. Other dimensions, such as velocity and variety are equally important. The variety of data sources, the quality of the data to be integrated and data visualization are some of the challenges for BD integration. The paper’s primary focus has been on analytics to gain valid and valuable insights from big data. We highlight the point that predictive analytics, which deals mostly with structured data, overshadows other forms of analytics applied to unstructured data, which constitutes 95% of big data.

Technological advances in storage and computations have enabled cost-effective capture of the informational value of big data in a timely manner. Consequently, one observes a proliferation in real-world adoption of analytics that were not economically feasible for large-scale applications prior to the big data era. For example, sentiment analysis (opinion mining) have been known since the early 2000s [1].

However, big data technologies enabled businesses to adopt sentiment analysis to glean useful insights from millions of opinions shared on social media. The processing of unstructured text fueled by the massive influx of social media data is generating business value by adopting conventional (pre-big data) sentiment analysis techniques, which may not be ideally suited to leverage big data. Since big data are noisy, highly interrelated, and unreliable, it will likely lead to the development of statistical techniques more readily apt for mining big data while remaining sensitive to the unique characteristics.

Going beyond samples, additional valuable insights could be obtained from the massive volumes of less ‘trustworthy’ data. Next, a strategic plan is developed for evaluating the different Big Data alternatives. Performance criteria can be used to select different suppliers. The clarification of success criteria allows the best estimation of value. The strategic plan must align Big Data technologies with existing infrastructures for BI and analytics.

REFERENCES

[1] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

[2] J. Gantz, D. Reinsel, "Extracting value from chaos", IDC iView, 2011, pp 1–12.

[3] Canada Inforoute, "Big Data Analytics in health", White Paper, Full Report, April 2013.

[4] Gartner, "IT glossary: big data" [webpage on the Internet]. Stamford, CT; 2012.Retrieved from: http://www.gartner.com/it-glossary/big-data.

[5] E. Mcnulty, "Understanding Big Data: The Seven V’s", Dataconomy, May 22, 2014, Retrieved from: http://dataconomy.com/seven-vs-bigdata/.

[6] A. Alexandru, D. Coardos, "BD in Tackling Energy Efficiency in Smart City", Scientific Bulletin of the Electrical Engineering Faculty, vol. 28, no. 4, pp. 14-20, 2014, Bibliotheca Publishing House, ISSN 1843-6188.

[7] Frost & Sullivan White Paper, "Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations", 2012, Retrieved from http://www.emc.com/collateral/ analyst-reports/frost-sullivan-reducing-information-technology-complex ities-ar.pdf.

[8] Watson, H.J. (2013a) “All about Analytics", International Journal of Business Intelligence Research, (4)2, pp.13-28.

[9] Power, D.J. (2007) “A Brief History of Decision Support Systems", DSSResources.com, http://DSSResources.COM/history/dsshistory.html, version 4.0 (current March 7, 2014).

[10] Watson, H. J. (2009a) "Tutorial: Business Intelligence – Past, Present, and Future", Communications of the Association for Information Systems,(25)39http://aisel.aisnet.org/cais/vol25/iss1/39 (current March 7, 2014).