International Journal of Computer Networks and Communications Security

VOL. 5, NO. 8, AUGUST 2017, 181–187
Available online at: www.ijcncs.org
E-ISSN 2308-9830 (Online) / ISSN 2410-0595 (Print)

Big Data: Concepts, Approaches and Challenges

Lawal Idris Bagiwa

Department of Computer Science, College of Science and Technology, Hassan Usman Katsina Polytechnic
P.M.B. 2052, Katsina, Nigeria

lbagiwa@yahoo.com

ABSTRACT

These days, many people in the information technology world and in corporate boardrooms are talking about "Big Data." Many believe that, for organizations that get it right, Big Data will be able to unleash new organizational capabilities and value. But what does the term "Big Data" actually entail, and how will the insights it yields differ from what managers might generate from traditional analytics? This paper presents an overview of Big Data and highlights different types of Big Data analytics as well as the concepts and approaches to their implementation. It serves as a starting point for individuals and organizations, who are interested in this technology. The Big data challenges discussed are Scale, Performance, Continuous Availability, Workload Diversity, Data Security, Manageability and Cost.

Keywords: *Big Data, Big Data Concepts, Big Data Technologies, 3Vs, Big Data Challenges*.

1 INTRODUCTION

Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" [10]. There is no question that organizations are swimming in an expanding sea of data that is either too voluminous or too unstructured to be managed and analyzed through traditional means. Any data that exceeds our current capability of processing can be regarded as "big" [5, 23]. Among its burgeoning sources are the clickstream data from the Web, social media content (tweets, blogs, Facebook wall postings, etc.) and video data from retail and other settings and from video entertainment [26]. But Big Data also encompasses everything from call center voice data to genomic and proteomic data from biological research and medicine. Every day, Google alone processes about 24 petabytes (or 24,000 terabytes) of data [4]. Yet many organizations either not benefitting at all from this pool of information or only very few do [25]. Gartner [6] estimates that 'Through 2016, 85% of Fortune 500 organizations will be unable to exploit Big Data for competitive advantage' [5]. Which means only 15% will be venturing beyond the 'known-knowns' of their business, using Big Data and analytics to uncover new insights and deliver a sharper competitive edge. Lack of awareness about the concept and benefits of Big Data has been identified as the major barrier to effective utilization of Big Data. Therefore, 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.

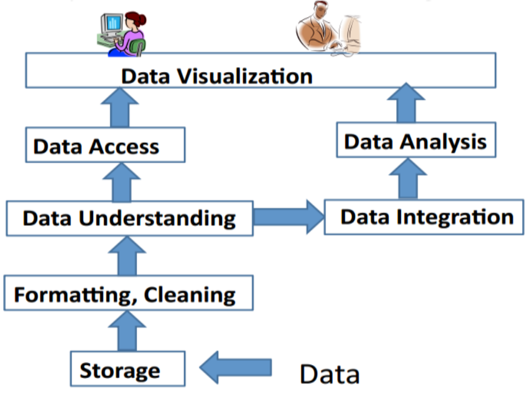

Fig. 1. Computational View of Big Data

Figure 1 presents the flow of data in Big Data technology from acquisition to deep analysis and visualization.

2 BIG DATA CONCEPTS

Big Data represents a new era in data exploration and utilization [26]. Big Data is a broad term for data sets so large or complex that traditional data processing applications are inadequate [5]. The term **Big Data** is being increasingly used almost everywhere on the planet – online and offline. And it is not related to computers only. It comes under a blanket term called Information Technology, which is now part of almost all other technologies and fields of studies and businesses [10, 11]. Some experts say that the Big Data Concepts are three Vs [7, 19]: Volume, Velocity and Variety as shown in figure 2.

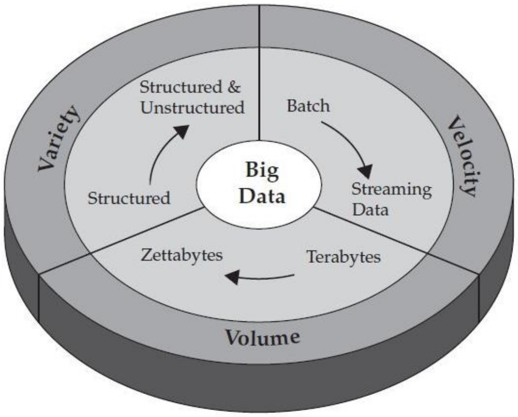

Fig. 2. Big Data by its volume, velocity, and variety or simply 3Vs

Volume refers to the vast amount of data generated every second. Just think of all the emails, Twitter messages, photos, video clips and sensor data that are produced and shared every second nowadays. People are not talking about terabytes, but zettabytes or brontobytes of data. On Facebook alone, 10 billion messages can be sent per day, click the like button 4.5 billion times and upload 350 million new pictures each and every day [13, 24].

Velocity refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in minutes, the speed at which credit card transactions are checked for fraudulent activities or the milliseconds it takes trading systems to analyze social media networks to pick up signals that trigger decisions to buy or sell shares. Big Data technology now allows us to analyze the data while it is being generated without ever putting it into databases [16].

Variety refers to the different types of data that can now be used. In the past, the focused was on structured data that neatly fits into tables or relational databases such as financial data (for example, sales by product or region). In fact, 80 percent of the world's data is now unstructured and therefore can't easily be put into tables or relational databases—think of photos, video sequences or social media updates. With Big Data technology, different types of data can be harnessed including messages, social media conversations, photos, sensor data, videos or voice recordings and bring them together [23]. Some others add few more Vs to the concept like Veracity (Reliability) and Value

Veracity refers to the messiness or trustworthiness of the data. With many forms of Big Data, quality and accuracy are less controllable, for example twitter posts with hashtags, abbreviations, typos and colloquial speech. Big Data and analytics technology now allows us to work with these types of data. The volumes often make up for the lack of quality or accuracy.

Value refers to our ability to turn our data into value. It is important that businesses make a case for any attempt to collect and leverage Big Data. It is easy to fall into the buzz trap and embark on Big Data initiatives without a clear understanding of the business value it will bring. Table 1 demonstrates Big Data origin and target use domains.

Table 1: Big Data origin and target use domains

Big Data Origin	Big Data Target Use
1. Science	(a) Scientific discovery
2. Telecom	(b) New technologies
3. Industry	(c)Manufacturing, process
4. Business	(d) Control, transport
5.Living Environment,	(e) Personal services,
6. Social media and networks	(f) Campaigns
7. Healthcare	(g) Living environment

2.1 *Big Data Analytics*

Big Data analytics is the process of examining large data sets containing a variety of data types -- i.e., Big Data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits [25]. Big Data analytics has been categorized into four major processes as follows:

Prescriptive analytics is really valuable, but largely not used. According to Gartner [13], 13 percent of organizations are using predictive but only 3 percent are using prescriptive analytics. Where Big Data analytics in general sheds light on a subject, prescriptive analytics gives a laser-like focus to answer specific questions. For example, in the health care industry, one can better manage the patient population by using prescriptive analytics to measure the number of patients who are clinically obese, then add filters for factors like diabetes and LDL cholesterol levels to determine where to focus treatment. The same prescriptive model can be applied to almost any industry target group or problem [20, 22].

Predictive analytics use Big Data to identify past patterns to predict the future. For example, some companies are using predictive analytics for sales lead scoring. Some companies have gone one step further use predictive analytics for the entire sales process, analyzing lead source, number of communications, types of communications, social media, documents, CRM data, etc. Properly tuned predictive analytics can be used to support sales, marketing, or for other types of complex forecasts [8].

Diagnostic analytics are used for discovery or to determine why something happened. For example, for a social media marketing campaign, one can use diagnostic analytics to assess the number of posts, mentions, followers, fans, page views, reviews, pins, etc. There can be thousands of online mentions that can be distilled into a single view to see what worked in your past campaigns and what didn’t [3, 27]

Descriptive analytics or data mining are at the bottom of the Big Data value chain, but they can be valuable for uncovering patterns that offer insight. A simple example of descriptive analytics would be assessing credit risk; using past financial performance to predict a customer’s likely financial performance. Descriptive analytics can be useful in the sales cycle, for example, to categorize customers by their likely product preferences and sales cycle.

Harnessing Big Data analytics can deliver big value to business, adding context to data that tells a more complete story. By reducing complex data sets to actionable intelligence one can make more accurate business decisions [21].

"Big Data Analytics" has recently been one of the hottest buzzwords. It is a combination of "Big Data" and "Deep Analysis." The former is a phenomenon of Web2.0 where a lot of transaction and user activity data have been collected, which can be mined for extracting useful information. The latter is about using advanced mathematical/statistical techniques to build models from the data. In reality, it also identified that two areas are quite different and disjointed, and people working in each area have pretty different backgrounds, Big Data and Deep Analysis [14, 15].

2.2 *Big Data Area*

People working in this area typically come from a Hadoop, PIG/Hive background. They usually have implemented some domain-specific logic to process a large amount of raw data. Often the logic is relatively straight-forward based on domain-specific business rules. Research shows that most of the people working with Big Data come from a computer science and distributed parallel processing systems background, but not from the statistical or mathematical discipline [7].

2.3 *Deep Analysis Area*

On the other hand, people working in this area usually come from statistical and mathematical background, where the first thing being taught is how to use sampling to understand a large population's characteristic. Notice the magic of "sampling" is that the accuracy of estimating the large population depends on the size of sample but not the actual size of the population. In their world, there is never a need to process all the data in the population in the first place. Therefore, Big Data Analytics is unnecessary under this philosophy [9].

2.4 *Typical Data Processing Pipeline*

Figure 3 presents the typical data processing pipeline, though different and complex models could be found from different literatures.

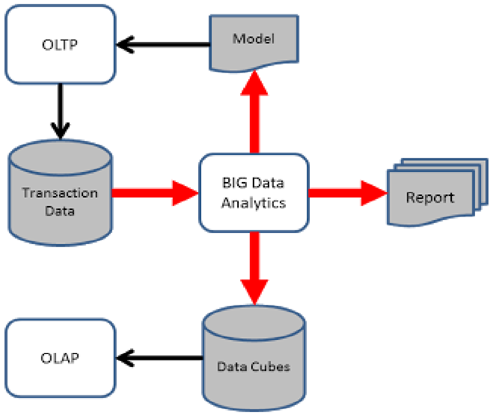

Fig. 3. Typical Data Processing Pipeline

In this model, data is created from the OLTP (On Line Transaction Processing) system, flowing into the Big Data Analytics system, which produced various outputs; including data mart/cubes for OLAP (On Line Analytical Processing), reports for the consumption of business executives, and predictive models that feedback decision support forOLTP.

2.5 Big Data + Deep Analysis

The Big Data analytics box is usually done in a batch fashion (e.g. once a day), usually Big Data processing and deep data analysis happen at different stages of this batch process.

In figure 4, the Big Data processing part (colored in orange) is usually done using Hadoop/PIG/Hive technology with classical ETL logic implementation. By leveraging the Map/Reduce model that Hadoop provides, we can linearly scale up the processing by adding more machines into the Hadoop cluster. Drawing cloud computing resources (e.g. Amazon EMR) is a very common approach to performing this kind of tasks. The deep analysis part (colored in green) is usually done in R, SPSS, or SAS using a much smaller amount of carefully sampled data that fits into a single machine's capacity (usually less than couple hundred thousand data records). The deep analysis part usually involves data visualization, data preparation, model learning (e.g. linear regression and regularization, K-nearest-neighbor/Support vector machine/Bayesian network/Neural network, Decision Tree and Ensemble methods), model evaluation [4, 10].

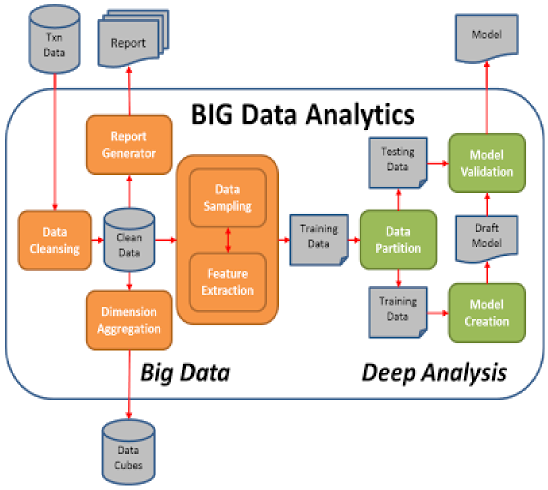

Fig. 4. Big Data and Deep Analysis

Sub section has to be in sentence case with no spacing above or below the star of it.

3 BIG DATA IMPLEMENTATION APPROACH

Since different models have their strengths and weaknesses, and it is often difficult to prejudge which type of model and its various versions will perform best on any given data set, an ensemble approach can be employed to build multiple solutions [10]. Here, literally hundreds of different algorithms can be applied to a dataset to determine the best or a composite model or explanation), a radically different approach to that traditionally used wherein the analyst selects an appropriate method based on their knowledge of techniques and the data [24]. In other words, Big Data analytics enables an entirely new epistemological approach for making sense of the world; rather than testing a theory by analyzing relevant data, new data analytics seek to gain insights 'born from the data'[15].

The explosion in the production of Big Data, along with the development of new epistemologies, is leading many to argue that a data revolution is under way that has far-reaching consequences to how knowledge is produced, business conducted, and governance enacted [19].

3.1 Big Data Implementation Architecture

Big Data architecture is premised on a skill set for developing reliable, scalable, completely automated data pipelines[1,2] . That skill set requires profound knowledge of every layer in the stack, beginning with cluster design and spanning everything from Hadoop tuning to setting up the top chain responsible for processing the data. The following diagram shows the complexity of the stack, as well as how data pipeline engineering touches every part of it. There are many possible ways to implement the data pipeline described. Here is one common implementation that works well in many projects [17].

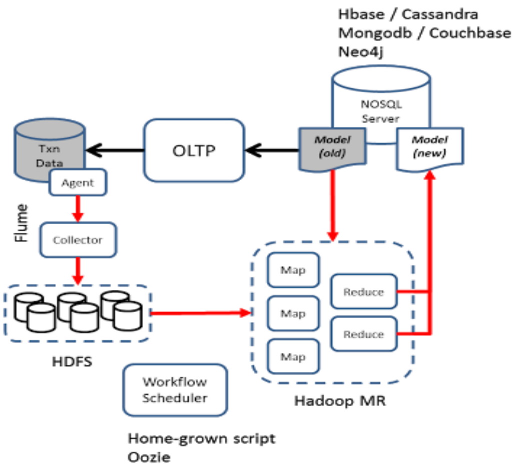

Fig. 5. Big Data Implementation Architecture

In this architecture, "Flume" is used to move data from OLTP system to Hadoop File System (HDFS). A workflow scheduler (typically a cron-tab entry calling a script) will periodically run to process the data using Map/Reduce. The data has two portions: (a) Raw transaction data from HDFS (b) Previous model hosted in some NOSQL server. Finally the "reducer" updates the previous model which will be available to the OLTP system [18].

4 BIG DATA CHALLENGES

Big Data Technologies are maturing to a point in which more organizations are prepared to pilot and adopt Big Data as a core component of the information management and analytics infrastructure. Big Data, as a compendium of emerging disruptive tools and technologies, is positioned as the next great step in enabling integrated analytics in many common business scenarios [1]. As Big Data wends its inextricable way into the enterprise, information technology (IT) practitioners and business sponsors alike will bump up against a number of challenges that must be considered before any Big Data program can be successful [3].

4.1 Scale

With Big Data you want to be able to scale very rapidly and elastically, whenever and wherever you want across multiple data centers and the cloud if need be. You can scale up or down. But most NoSQL solutions like MongoDB or HBase have their own scaling limitations that need to be considered.

4.2 Performance

In an online world where nanosecond delays can cost you sales, Big Data must move at extremely high velocities no matter how much you scale or what workloads your database must perform. The data handling hoops of RDBMS and most NoSQL solutions put a serious drag on performance [16].

4.3 Continuous Availability

When you rely on Big Data to feed your essential, revenue-generating 24/7 business applications, even high availability is not high enough. Your data can never go down. A certain amount of downtime is built-in to RDBMS and other NoSQL systems [17].

4.4 Workload Diversity

Big Data comes in all shapes, colors and sizes. Rigid schemas have no place here; instead you need a more flexible design. You want your technology to fit your data, not the other way around. And you want to be able to do more with all of that data – perform transactions in real-time, run analytics just as fast and find anything you want in an instant from oceans of data, no matter what form that data may take [18].

4.5 Data Security

Big Data carries some big risks when it contains credit card data, personal ID information and other sensitive assets. Most NoSQL Big Data platforms have few if any security mechanisms in place to safeguard your Big Data [18].

4.6 Manageability

Staying ahead of Big Data using RDBMS technology is a costly, time-consuming and often futile endeavor. And most NoSQL solutions are plagued by operational complexity and arcane configurations [18].

4.7 Cost

Meeting even one of the challenges presented here with RDBMS or even most NoSQL solutions can cost a pretty penny. Doing Big Data the right way doesn’t have to break the bank [13].

These seven (7) are some few challenges out of the many challenges associated with Big Data technology.

5 CONCLUSIONS AND FUTURE RESEARCH

The arrival of Big Data in society has prompted business and government to take actions to exploit its value and application. This paper described the characteristics of Big Data and presented architecture for Big Data analytics. Big Data technology deviates from traditional data management SQL-based RDBMS approaches as it deals with data with high volume, velocity and variety. The new paradigm moves towards NoSQL databases, massively parallel and scalable computing platforms, open-source software, and commodity servers. The capability for organizations to collect and process Big Data about individuals or groups as the data comes from various sources causes various privacy concerns. Further study could address the ethical issues that may arise from Big Data and the measures that society can take to mitigate such concerns.

6 REFERENCES

[1] Adamczyk, A., & LaFree, G. (2015). Religiosity and reactions to terrorism. Social science research, 51, 17-29.

[2] Ammu, N., & Irfanuddin, M. (2013). Big Data Challenges. International Journal of Advanced Trends in Computer Science and Engineering, 2(1), 613-615.

[3] Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., & Zhao, J. L. (2014). Transformational issues of big data and analytics in networked business. MIS quarterly, 38(2), 629-631.

[4] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly, 36(4), 1165-1188.

[5] Chen, L., Zheng, D., Liu, B., Yang, J., & Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. Nucleic acids research, 44(D1), D694-D697.

[6] Chen, M., Mao, S., & Liu, Y. (2014). Big data: a survey. Mobile Networks and Applications, 19(2), 171-209.

[7] Dahl, E. J. (2015). Book Review: A Practitioner's Way Forward: Terrorism Analysis by David Brannan, Kristin Darken, and Anders Strindberg (Salinas, CA: Agile Press, 2014). Homeland Security Affairs, 11.

[8] Demirkan, H., & Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. Decision Support Systems, 55(1), 412-421.

[9] Dixit, P. (2016). Securitization and Terroristization: Analyzing States' Usage of the Rhetoric of Terrorism State Terror, State Violence (pp. 31-50): Springer.

[10] Feldman, D., Schmidt, M., & Sohler, C. (2013). Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. Paper presented at the Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms.

[11] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

[12] Goh, B. (2015). Prosperity and Security: A Political Economy Model of Internet Surveillance.

[13] Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward. Paper presented at the System Sciences (HICSS), 2013 46th Hawaii International Conference on.

[14] Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12), 2032-2033.

[15] LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. MIT sloan management review, 52(2), 21.

[16] Lynch, C. (2008). Big data: How do your data grow? Nature, 455(7209), 28-29.

[17] Markus, M. L. (2015). New games, new rules, new scoreboards: the potential consequences of big data. Journal of Information Technology, 30(1), 58-59.

[18] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data. The management revolution. Harvard Bus Rev, 90(10), 61-67.

[19] Meng, X., & Ci, X. (2013). Big data management: concepts, techniques and challenges. Journal of Computer Research and Development, 50(1), 146-169.

[20] Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. Jama, 309(13), 1351-1352.

[21] Ohlhorst, F. J. (2012). Big data analytics: turning big data into big money: John Wiley & Sons.

[22] Riggins, F. J., & Wamba, S. F. (2015). Research Directions on the Adoption, Usage, and Impact of the Internet of Things through the Use of Big Data Analytics. Paper presented at the System Sciences (HICSS), 2015 48th Hawaii International Conference on.

[23] Russom, P. (2011). Big data analytics. TDWI Best Practices Report, Fourth Quarter, 1-35.

[24] Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. Paper presented at the Collaboration Technologies and Systems (CTS), 2013 International Conference on.

[25] Suthaharan, S. (2016). Big Data Analytics Machine Learning Models and Algorithms for Big Data Classification (pp. 31-75): Springer.

[26] Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. Big Data, 1(2), 85-99.

[27] Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. Knowledge and Data Engineering, IEEE Transactions on, 26(1), 97-107.