International Journal of Engineering & Technology, 7 (4.5) (2018) 485-488

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

The importance of big data technology

Samson Fadiya [1] *, Arif Sari [1]

[1] Department of Management Information Systems, Girne American University, Canterbury, Kent, United Kingdom
**Corresponding author E-mail: greaterachiever@yahoo.com, arifsarii@gmail.com*

Abstract

The adoption of Web 2.0 technologies, Internet of Things, etc. by individuals and organization has led to an explosion of data. As it stands, existing Relational Database Management Systems (RDBMSs) are incapable of handling this deluge of data. The term Big Data was coined to represent these vast, fast and complex datasets that regular RDBMSs could not handle. Special tools or frameworks were developed to deal with processing, managing and storing this big data. These tools are capable of functioning in distributed industry- standard environments thereby maintaining efficiency and effectiveness at a business level. Apache Hadoop is an example of such a framework. This report discusses big data, it origins, opportunities and challenges that it presents, big data analytics and the application of big data using existing big data tools or frameworks. It also discusses Apache Hadoop as a big data framework and provides a basic overview of this technology from technological and business perspectives.

Keywords: *Big Data; Big Data Analytics; Relational Database Management Systems; Apache Hadoop*

1. Introduction

For the first time in the history of modern technology, do not computers change, but the information, which does, they process. The importance of data in today's business is difficult to overstate because no meaningful decision can be made without the analysis of relevant data. A few years ago, the world faced a new challenge - an incredibly fast growth in volumes and flows of digital data. Rapid accumulation of data occurred earlier, but it managed to cope with it - the old data storage and processing tools kept pace with the growth of their volumes. The data analysis isn't only drive the decision making but also takes an active part in develop- ing strategies and methods that ensure the existence and success of organizations. Earlier, entrepreneurs used to call data analysis "business intelligence," which perfectly characterizes the essence because data could provide a competitive advantage to those who used and interpreted them properly. Nowadays, a new term for referring to business intelligence was coined: "big data" This name also makes a good sense because since the times of "busi- ness intelligence" the volumes of data became incredibly large.

The current rapid increase in data sets is due to the extraordinary popularity of social networks and the importance of their pro- cessing: hundreds of millions of enthusiasts publish a huge num- ber of texts, images, videos, audio recordings. So, every minute users of the YouTube resource download 35 hours of video, every day they open the videos 2 billion times - 20 times more. Analyz- ing data published in social networks and news feeds, we can get invaluable marketing information about customers, catch changes in their moods, understand what is happening in key markets, what steps competitors plan to take, and much more. As the result, more effort should be applied to deal with them and make it useful for analytics professionals.

According to Akamai (2017), several notable events have occurred from a technological perspective. They include increased connection to the Internet all over the world i.e. increasing Internet penetration across the globe, the smart phone supplanting the personal computer as the primary computing device for most users, the use of social media as a primary method of communication and the disruption of said communication channel by authorities. These just a few notable technological developments were mentioned. However, a major phenomenon has emerged from the confluence of the Internet and connected parties. This is Big Data. Mannen's (2012) favorite definition of big data is that it is data, which is too large, too difficult or too fast for regular or existing data analysis tools to handle or process. He goes on to say that too large refers to currently existing scenarios where companies are immersed in a deluge of data to the tune of several petabytes. Too difficult refers to scenarios where the data just doesn't fit into the constraints of existing data analysis tools while too fast refers to scenarios where the data needs to be processed incredibly quickly but existing data processing tools are incapable of such functions. As the definition above implies, special tools and deployment platforms are required to process big data to reap its significant benefits. For instance, a company like Google, which has plenty of users who concurrently insert, update, delete and manipulate deals with big data. The sheer volume of user interactions with their system yields such large volumes of data that regular relational database management systems no longer suffice. Clearly, such companies had to find new tools and platforms to analyze and manage their data. IBM (2017) defines big data as large data sets that are beyond the ability of traditional relational databases to capture, manage, and process the data with low-latency. It goes on188

To say that big data one or several of three critical characteristics namely; high variety, high volume or high velocity. They round up by saying that big data comes from a myriad of sources which include the following and more; "sensors, devices, video/audio, networks, log files, transactional applications, web, and social media - much of it generated in real time and in a very large scale" IBM (2017).

For companies such as Google and Facebook amongst others, big data is truly priceless. Upon analysis, big data can yield invaluable information about the users of the services from these companies. In fact, in the month of December 2015, Facebook on average had

approximately 1.04 billion daily active users connecting to their network in different languages via different devices and engaging in different digital social actions (Facebook, 2015). That is a tremendous amount of data to deal with. However, the analysis of this data has provided a means to categorize and research cultural and social dynamics. In fact, Bhadani & Jothimani (2016) posit that the proper analysis of big data can provide firms with excel- lent actionable insights which can lead to the development of competitive advantage. Also, big data from these companies can act as input entry points for artificial intelligence. This is not pecu- liar to companies such as Google and Facebook. Big data analytics is a major endeavor in the public and private sector. In fact, it can be inferred the advent of big data analytics heralded the current age of artificial intelligence that we currently witness.

Big data and big data analytics were initially fraught with challenges for the entities that wanted to harness them. At the time, the existing database management systems were inadequately positioned to handle that volume of data at the speeds that were being witnessed and the complexity that was required. This led to the development of platforms and languages that are better suited to handle such big data operations. Currently, there is large number of big data tools and platforms. However, the focus of this report will be on Hadoop as an open source framework for the distribut- ed storage of big data.

This report will explore the advent of big data, the challenges that it created, and the tools that were developed to deal with the new paradigm and the problems that still exist in the big data sphere.

2. Related works

Madden (2012) described how open source RDBMs such as MySQL and PostgreSQL failed to deal with big data operations adequately. He posited that it was because those database systems had to slowly take in the data in a format that is a native represen- tation. Hence, they could not handle fast data processing actions. Also, these open source systems performed poorly with respect to their ability to scale as operations escalated. He compared the existing RDBMS systems with frameworks such as Hadoop and MapReduce and declared that frameworks like Hadoop and MapReduce are better equipped to process big data. They outper- formed existing RDBMS systems with respect to their ability to scale large amounts of data. However, he noted that with respect to managing big data, the frameworks performed just as poorly as the RDBMS systems. Therefore, he posited that frameworks such as Hadoop are optimized to process data but not to manage it. A critical part of the big data paradigm is big data analytics and vis- ualization. According to IBM (2017), the term big data analytics refers to employment of advanced analytic techniques against expansive, varied data sets that include different types such as structured/unstructured and streaming/batch, and different sizes from terabytes to zettabytes.

Analysing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously Inaccessible or unusable. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing, businesses can analyze previously untapped data sources independent or together with their existing enterprise data to gain new insights resulting in significantly better and faster decisions. Bhadani & Jothimani (2016) point out that big data has presented businesses with new challenges. They point to storage related issues and the limited ability of existing RDBMs to properly process and manage the data. However, they highlight the fact that big data also comes with quite a few opportunities for companies that can properly analyze the data and derive useful and actionable insights from the processed data. In other words, big data can be a double-edged sword depending on how it is employed by firms. Once again, this perspective highlights the need for the development of robust big data analytics systems with the ability to handle big, fast and com- plex datasets. Samson Oluwaseun Fadiya, Serdar Saydam & Van- dule Vany Zira (2014) said, regarding health data about patients, diseases and the data produced by various medical devices will be massive. And, data generated from different machines in the pro- duction sectors like transport, war ammunitions, finance and many more are similarly a source of massive data, with each being stored for a different reason for future use.

More so, organizations can process this data, analyze it and store it for intelligent decision-making, to gain a highly competitive benefit over their contemporary. However, to engage in effective and efficient big data analytics, the need for appropriate tools and frameworks cannot be understated. This need consequently led to the development of distributed storage frameworks, robust data languages, NoSQL frameworks that handle unstructured and complex data and effective big data analytics frameworks such as IBM Watson Analytics. Some of the popular big data tools that are used by successful analytics developers are described below.

2.1. Cassandra

Apache Cassandra is a scalable open-source NoSQL database, perfectly suited for managing many unstructured, semi-structured and structured datasets in several data centres or in the cloud. This tool is widely used today because it provides an effective man- agement of large amounts of data. A database offers high availability and scalability without compromising the performance of commodity hardware. Cassandra provides continuous availabil- ity, linear scalability. It can be used on multiple servers without a single point of failure. The task, which is solved by the developed architecture of Cassandra, is the ability to process petabytes of data and thousands of parallel user operations per second. The Cassandra architecture allows any authorized user who connects to any node of any data centre to access data using the CQL language (SQL similar language).

Cassandra provides the user with a mechanism for automatic data distribution, so the programmer or administrator does not need to perform additional operations or write additional code to distribute data on the cluster. Cassandra also includes a built-in custom replication mechanism. The replicated data is stored in the nodes of the Cassandra ring. Thus, in case of failure of one of the nodes, the cluster will still contain the nodes from which these data can be obtained. Replication can also be configured to work within a datacentre, multiple datacentres, or multiple cloud areas.

2.2. Cloudera

Cloudera is an American company, the developer of middleware, which produces a commercial version of the software framework Apache Hadoop. It was developed in 2008 and still is the most popular provider. The business model of the company is compared to the Red Hat business. Cloudera creates distributions of software products for organizations on the basis of free software and makes Profit by providing technical support for the delivered solutions. With the boom of technology for "large data", Cloudera is repeatedly noted as one of the most promising companies able to solve the problems of the corresponding class.

2.3. Google refine

Google offers an open source project that deserves attention - Google Refine. This is an excellent tool for cleaning data sets and for performing complex data operations, such as converting from one format to another, converting data. Simply saying, Google Refine will help to organize the data in the database that was noth- ing but a mess. As the result, the users can begin to process the data with the computer.

2.4. Hadoop

Hadoop is an open source project run by the Apache Software Foundation. Hadoop is used for reliable, scalable and distributed

computing, but can also be used as a general-purpose file store capable of hosting petabytes of data. Many companies use Hadoop for research and production purposes. Hadoop consists of two key components:

a) Distributed file system Hadoop, which is responsible for storing data on the Hadoop cluster;
b) MapReduce system, designed for computing and pro- cessing large amounts of data on a cluster.

2.5. Hadoop

Neo4j is an open-source graph database, whose history began with investments of Neo Technology in 2003. Since 2007, it has become publicly available. Neo4j has all the characteristics of databases, including compliance with ACID, maintaining clustering and recovery from a system failure. Neo4j is a high-performance, NoSQL database based on the principle of graphs. In it there is no such thing as a table with strictly prescribed fields, it operates with a flexible structure in the form of nodes and connections between them. Today is the leader among database graphs.

2.6. Rapid miner

A big data specialist needs this open source data science platform, which functions through visual programming. Now, there are many companies that need analytics systems, but the high cost and excessive complexity of this software in most cases forces us to abandon the idea of building our own analytical system in favour of a simple known excel. Also, additional costs for training em- ployees, supporting expensive storage systems, etc. And then Open Source solutions can come to their help - there are not so many of them, but there is very worthy software, one of which is RapidMiner.
Rapid Miner is a tool created for the mining date, with the basic idea that the miner should not program while performing his work. Data mining is needed for data, so it has been equipped with a good set of operators to solve a wide range of tasks for obtaining and processing information from a variety of sources (databases, files, etc.), and it is safe to say that this is also a good-full tool for ETL. In addition, it allows manipulating, analysing, model, creat- ing models, and integrating the data into business processes.

3. An importance of the development

Rustom (2013) described organizations that could not adapt quickly to the vast volumes of data flooding through their technology infrastructure as the proverbial "deer in the headlights". In other words, they were overwhelmed by the sheer volume, speed and complexity of the data (structured and unstructured) coming in. He went on to say that the right combination of Hadoop can effective jolt an organization out of its "analysis paralysis" and allow it to properly store, process and manage massive datasets. Essentially, without big data tools like Hadoop amongst others, it would be very difficult to process and manage this data. The importance of such tools/frameworks can be seen across different spheres of both the public and the private sector. Several uses cases documented by Moise & Pournaras (2017) indicate that use of frameworks like Hadoop have been integral in helping organizations engage in big data analytics.
The benefits of such endeavors have been the creation of systems that can engage in predictive modeling, fraud detection, recommendation engines and much more. In fact, Moise & Pournaras indicate that in 2008, Yahoo! Inc. employed Linux based Hadoop cluster to run its Yahoo! Search Web map. As stated earlier, this is not limited to the private domain. Moise & Pournaras go on to show how NASA is using big data analytics for scientific purpos- es such as "graph-based automated method for identifying and tracking Mesoscale Convective Complexes that can be categorized as a severe weather event" (2017).

4. Methodologies behind big data

We intend to explore Hadoop from the IT context and the business context. In other words, we intend to examine the technical as- pects of Hadoop and how the use of Hadoop has facilitated organ- izational effectiveness and efficiency across the globe.
According to Intel (2013), Apache Hadoop is a Java-coded distributed software platform for processing and storing data. They go on to say that with Hadoop, you can safely and consistently store extremely large datasets on industry-standard servers with direct-attached storage while scaling. Furthermore, they posit that this scaling can be done in a cost-efficient manner. It is no wonder that Hadoop is a favorite for companies like Yahoo! and Face- book. Alteryx & Hortonworks (2013) say that, "Apache Hadoop is an open source project from the Apache Software Foundation that has rapidly emerged as the best way to handle massive amounts of data, a.k.a, Big Data". They go on to say that, a primary defining attribute of Hadoop is that it surmounts the limitations of older data management models and does this on low cost hardware. According to them, adopting Hadoop as a tool for big data analyt- ics can achieved via two approaches:

- Employing Hadoop as a refinement tool which then loads the data into a data warehouse
- Employing Hadoop as the data store.

Alteryx and Hortonworks (2013), say that for an organization to successfully and effectively employ IT as a tool to facilitate business success, three principle priorities must be addressed: 190

- Quick delivery of analytic solutions for the resolution of time-sensitive problems
- Use of relevant data to answer questions in the correct context
- Ease of use by the consumers within and without the organization

Several organizations such as Facebook, NASA, Yahoo!, and Google just to name a few have implemented Big Data Hadoop and simi- Lar frameworks to ensure that these key priorities are met.

5. Conclusion

In conclusion, big data analytics will continue to remain rele- vant for quite a while. As such, big data tools/frameworks and platforms will continue to be relevant for the foreseeable future. In fact, as the data generated from social networks, Internet of Things, sensor-based networks, etc continues to grow, tools like Hadoop will evolve with the technological and business environment. Whether it will continue to remain relevant will depend on its ability to evolve fast enough to meet the growing demands of data-driven organizations across the world.
Big Data is not a speculative reflection, but a symbol of the overtaking technological revolution. The need for analytical work with large data will significantly change the face of the IT industry and stimulate the emergence of new software and hardware platforms. Already today, for the analysis of large amounts of data, the most advanced methods are used: artificial neural networks - models built on the principle of the organization and functioning of biological neural networks; methods of predictive analytics, statistics and Natural Language Processing (directions of artificial intelligence and mathematical linguistics, studying the problems of computer analysis and the synthesis of natural languages). Meth- ods are also used that attract expert people, or crowdsourcing, A / V testing, sentiment analysis, etc. That's why in this research pa- per, we have compiled a list of the top six big data tools that are used by successful analytics developers.

Acknowledgement

This research was supported by [Girne American University, Canterbury, Kent, United Kingdom]. We thank our colleagues from

[Girne American University, Girne, And Northern Cyprus] who pro- vided insight and expertise that greatly assisted the research.

References

[1] Akamai Technologies. (2017). Akamai's State of the Internet.
[2] Apteryx & Hortonworks. (2013). the Business Analyst's Guide to Hadoop.
[3] Bhadani, A. K., & Jothimani, D. (2016). Big Data: Challenges, Opportunities and Realities. Delhi, India.
[4] Facebook. (2015). Retrieved 2017, from Statistic Brain: http://statisticbrain.com/facebook-statistics.
[5] IBM. (2017, October). IBM Watson Analytics. Retrieved 2017, from IBM: https://www.ibm.com/us-en/marketplace/watson- analytics.
[6] Intel. (2013). Extract, Transform, and Load Big Data with Apache Hadoop.
[7] N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov and O. A. Kutnenko. A quantitative measure of compactness and similari- ty in a competitive space. Journal of Applied and Industrial Mathematics, 2011, Vol. 5, № 1, pp.144-154.
[8] N. G. Zagoruiko, I. A. Borisova, O. A. Kutnenko, V. V. Dyu- banov. A construction of a compressed description of data us- ing a function of rival similarity. Journal of Applied and Indus- trial Mathematics, April 2013, Volume 7, Issue 2, pp 275-286.
[9] Madden, S. (2012). From Databases to Big Data. IEEE Internet Computing, p. 4.
[10] McMahon GT, Gomes HE, Hohne SH, Hu TM, Levine BA &Conlin PR (2005). Web-based care management in patients with poorly controlled diabetes. Diabetes Care 28, 1624–1629. http://www.tadviser.ru/index.php/http://lpgenerator.ru/blog/201 6/04/01/obzor-5-instrumentov-dlya-sozdaniya-udivitelnyh-onlajn-grafikov.
[11] Moise, I., & Pournaras, E. (2017). Big Data Analytics.
[12] Pavlovskiy E.N. master is Program “Big Data Analytics” In Novosibirsk State University. International conference on Clouds, Big Data and Trust (ICCBDT- 2013), 13-15 November 2013, Bhopal, India.
[13] Russom, P. (2013). Integrating Hadoop into Business Intelli- gence and Data Warehousing.
[14] Samson, O. F., Serdar, S., Vanduhe, V. Z. Advancing big data for humanitarian needs. Humanitarian Technology: Science, Systems and Global Impact 2014, HumTech (2014).
[15] Thakurdesai PA, Kole PL & Pareek RP (2004). Evaluation of the quality and contents of diabetes mellitus patient education on Internet. Patient Education and Counseling 53, 309–313.