Sindh Univ. Res. Jour. (Sci. Ser.) Vol.49 (004) 825-830 (2017)

http://doi.org/10.26692/sujo/2017.12.0066

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Big Data Analytics Solutions and Challenges

A. AHMED, R.U. AMIN*, M. R. ANJUM**, I. ULLAH, I. S. BAJWA**

Department of Computer Science and Information Technology, University of Balochistan, Quetta

Received 5thOctober 2016 and Revised 26thApril 2017

Abstract: Big data is considered as a serious problem since few years ago. However, advancements in analytics has enhanced the performance of analytics of large sized datasets. Lots of research is being carried out to improve various Big Data Analytics schemes (BDA) instead of cumbersome traditional data processing methods. Big data analysis is considered to be a pipeline of recording; cleaning, extraction, and annotation; aggregation, integration and acquisition, analysis, modeling and interpretation. This work describes what big data is and its importance, moreover, how it's changing the analytics by furnishing us with new tools, techniques and opportunities for harnessing large amounts of unstructured and structured data. We also present difference line between BDA and traditional analytics. Finally, we concluded the current challenges that are faced by big data analytics experts.

Keywords: Cloud Computing, Big Data, Analytics.

1. INTRODUCTION

Analysis of the Internet usage for last two decades shows a dramatic increase in volume of data. The entity behavior is logged and due to this reason amount of data is increasing day by day over the Internet. Due to growing of data at a huge speed like for example Exabyte's make it difficult to handle (Gray., 2003)(Courtney,2012)We can say that volume is increasing rapidly as compared to the resources. Sources from where these data are generated are emails, videos, transactions, images, audios, click streams, logs, posts, social networks, mobile phones, maps, traffic pattern, Radio Frequency Identification (RFID) tags (Hua, *et al.*,2009)and weather data, etc.

In early 2000s, data volumes started skyrocketing and CPU[1] and storage technologies were overawed by the various terabytes of big data on that point data scalability crisis was faced by IT.Butwith the help of Moore's law, we escape from defeat. During the technological developments of the past few decades, CPU and storage not only provided us with enormous speeds and access but they greatly increased the architectural capacity and system's intelligence by furnishing these on a reduced price.(Zaslavsky, *et al.*, 2013) Before this companies were not capable of managing or affording big data due to lavishing budget requirements on its analysis and collection.

Nowadays, enterprises are adapting and exploring big data to determine many new facts. This is one of the important technological achievements because economic recessions throughout the world have enforced profound changes in many industries having masses of consumers. Thanks to these advancements in data that now companies can comprehend present level of their business along with the consumer's behavior by employing modern analytical data strategies.

It is estimated that approximately 2.5 quintillion bytes of data is generated by human beings on per day basis and the data creation rate has increased so much that today 90 percent of the world data is generated alone during the past three years. Due to this reason, new technologies are needed to analyze massive data sets.

In this work, we present an analysis of the big data techniques used in analytics. Our main concern is to identify their core concepts, architectures and the challenges they face during the big data management and analytics. Following isthe hierarchy of this research: next section contains some basic concepts and analogies regarding big data. In section 3, we present big data analytics (BDA) and some of its corresponding technologies. We also present a framework for supporting our deliberations in the discussion. In section 4, some of the security vulnerabilities are identified and finally, a conclusion is added towards the end.

2. BIG DATA

Big data is one the uprising technologies that has brought several benefits in the data management techniques. Big data is used for huge datasets that have comparatively large and multifaceted structure therefore it is not easy to handle them by using on-hand and

Corresponding author: atiq.ahmed@uob.edu.pk,riazulamin@gmail.com

*Department Of Computer Science, Balochistan University Of IT, Engineering And Management Sciences, (BUITEMS) Quetta.

** Islamia University Bahawalpur

conventional database management techniques or contemporary data processing applications. Some of the main issues which are faced by big data are analysis, logging, storage space, transmission, searching and retrieval, distribution and visualization for further processing or result generation.

Big data is widely employed to express all kind of concepts, containing: social media analytics, enormous amounts of data, upcoming data management abilities, real time data and much more. Big data is characterized by its 4 main and few other components some of which are shown in **(Fig 1).**

Volume or refers to the huge amounts of data for which companies try to strap up to make better decision making across various enterprises. Data at scale e.g. data is larger than terabytes to Peta bytes. Variety depicts the heterogeneity of data that is coming from distinct sources. Handling the intricacy of different data types like for example, structured data, unstructured data and semi-structured are supposed to be managed here.

Velocity reflects the data in motion e.g., analysis of streaming data to enable decision making within the fractions of a second. Velocity impacts latency that is a lag time between when it is accessible and when data is created or captured. These days data is continually being generated at a very fast speed so that is not possible for the conventional systems to accumulate, analyze and capture that data. Veracity includes data uncertainty and rank of consistency which is combined with certain data types. One of the most important big data requirements and challenges is striving for improved data in terms of quality, but unluckily some of the best data cleansing techniques even fail to eliminate the intrinsic irregularity of some data e.g., predicting the economical factors or a future buying decision of a customer.

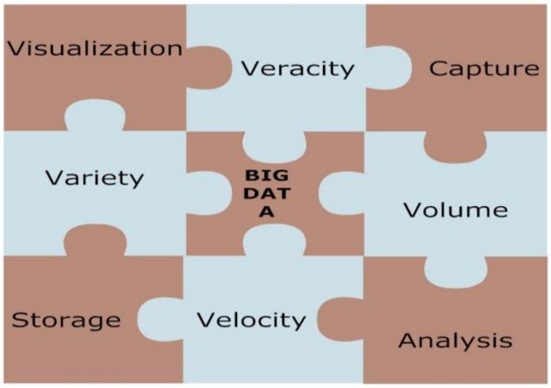

Fig.1. Big Data: The Big Picture

Nowadays, new big data technologies and tools continuously are developing. Many of new big data technologies depend upon MPP[1] databases that which can simultaneously distribute and process very huge datasets across multiple servers.

It is due to the advent of big data that organizations are now capable of gaining a comprehensive cognizance of their business, opponents, merchandises, consumers and many more in situations where big data is proficiently and successfully logged/captured, processed and analyzed. Due to this many improvements are made e.g. sales increased, lower costs, improve products and services etc. Widely used examples of big data are cited in **(Table 1):**

Table 1. Summary of Widely Used Big Data Spheres

Technique	Description
IT Logs	Using it to improve information technology troubleshooting and detection of security breaches, effectiveness, momentum, and upcoming event prevention.
Capacious past re-callings	Use its information for improving consumer satisfaction and interaction
Social media content	This is used to comprehend the consumer's psyche and improving the services, commodities and relations.
Fraud detection	Preventive measures in any commerce that involve online financial transactions like banking, insurance, investments, shopping and health related claims.
Financial market transaction	Take corrective action and Quickly assess risk.

Some of the factors driving the big data adoption are related with the technological developments in processing, storage and analysis. Rapid decline in the storage cost and CPU speed during the past few years is one of the factors. Secondly, elasticity and cost-effectiveness of cloud computing and data centers for flexible computing and storage spaces has played a vital role. Apart from these advent of novel big data frameworks like Hadoop (Cohen, *et al.*, 2009)(Barlow, 2013).that permit its users to have benefits of those distributed computing structures which were restricted to common users previously, by facilitating them in storing huge amounts of data via elastic and parallel processing. These developments have emerged numerous variances between (BDA) and traditional analytics.

3. BIG DATA METHODS

Loads of new heterogeneous data is reached in the Internet contributing in big data formation and all this has the prospective to furnish insights which can easily change any business statistics in no time. A whole new industry of supporting architecture is generated by big data such as MapReduce(Langheinrich, 2009).It is a programming structure paradigm which is created by Google Inc. for decentralized computing environments. MapReduce uses the divide and conquer technique where large and complex datasets are broken down to comparatively smaller components. Later on these components undergo parallel processing that usually takes place in two phases **(Fig 2)**

3.1.1 Map:

The master node data is split in multiple sub parts where each part represents the problem space. There is a worker node that deals with a number of smaller problem subset which is controlled by a JobTracker node. Its respective results are stored in the local file system where a reducer exists that is capable of accessing them.

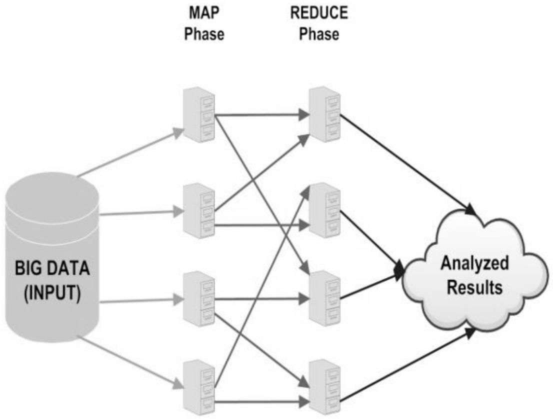

Fig 1. MapReduce Components

3.1.2 Reduce:

In this phase all the input data from participating nodes is merged after a profound analysis. For parallelize the aggregation there can be various reduce tasks and under the control of JobTracker these tasks are executed on the worker nodes.

3.2 Hadoop

It is one of the most common and well known tools for batch processing of large datasets. It is heterogeneous open source platform and java based framework. The Hadoop framework offers its developers HDFS[2] that enables them to store huge data files. MapReduce is also one of the architectural components whose programming aspects are employed in Hadoop for parallel and decentralized data processing of large datasets with the occurrences of processing problems or failures. Hadoopis designed in a way that it does not tackle the processing issues related with the real time or streaming data as it might add some more complexity in its architecture. Numerous tools are used that assist analysts to make multifaceted queries and execute machine learning algorithms over Hadoop including **(Fig 3)**

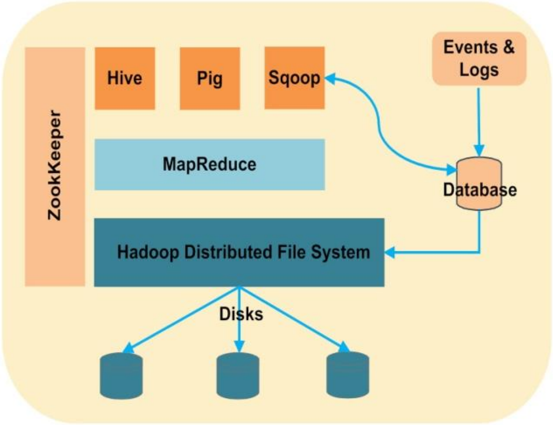

Fig 2. Hadoop Architecture

3.2.1 Pig:

It attempts to fetch Hadoop nearer to the realities of business and system developers. Pig is an open source and was developed by Yahoo (Ghazal, 2013). Pig is a scripting language for complex queries.

3.2.2 Hive:

It was developed by Facebook, allows the executions of complex queries against a Hadoop cluster. Initially it was open source but after becoming one of the important components of Hadoop structure, it permits to create and run queries by its users for large and complex datasets kept in the Hadoop clusters.

3.2.3 Mahout and RHadoop:

These include machine learning and data mining algorithm for Hadoop.

3.2.4 Spark 4:

It is a new framework that was deliberated to increase the effectiveness of machine learning and data mining algorithms that reprocess repetitively a working dataset. Consequently, improving the efficiency of sophisticated data analytics algorithms. For efficient storage space and query execution over big data, there are numerous supported databases have been designed including CouchDB, Cassandra, HBase, Greenplum Database, Vertica and MongoDB (Cohen, *et al.*, 2009).

[2] Hadoop Distributed File System

4. BIG DATA ANALYSIS AND ITS LIMITATIONS

BDA is the research procedure which is applied on the huge datasets in order to expose the concealed patterns and clandestine correlations within the data. Resulting valuable information becomes a deep asset for the businesses that assists them in obtaining more affluent and in-depth insights regarding the business to attain some benefits over the market competitors. This is the key reason that implementations and strategies regarding big data must thoroughly and precisely be analyzed before execution.

Big data analytics which are cost effective and precise in terms of time are proved to be the key constituents and have emerged as key to success in many disciplines like businesses, scientific and engineering and government activities. Social networking and web searching websites keep track of user activities by logging and capturing the user activities which are later on analyzed for the purpose of enhancing the site design, interactivity, user experience, on-site ad; and for keeping track of malicious code and spam detection from unknown sources and fraudulent activities on their pages.

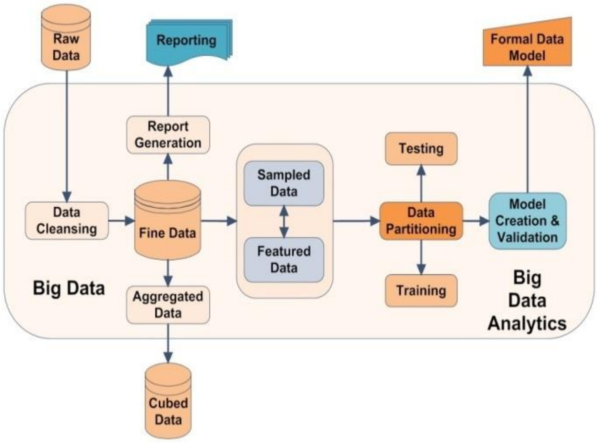

Fig 3. Big Data Analysis and its processes

Big data technologies are commonly split in two kinds depending upon the states that data is in. if analytics are performed on stagnant data then it is referred as batch processing and they are performed on moving data then the technique is called as stream processing **(Fig 4).** It is evident from the current research that synchronous processing does not necessitate existing in system's memory all the time. Some novel mechanisms like Drill and Dremel(Boyd, and Crawford.2012)which are interactive by their mechanism can analyze enormous datasets.

Data analytics include: predictive analytics, behavioral analytics and data interpretation. For this (Zikopoulos, and Chris. 2011).have proposed a MAD[3] mechanism for big data which is a three-fold procedure that expresses the user's expectations from the analytic system. First, magnetism which is used to attract all the possible data sources in spite of issues like lack of structure or possible unidentified data representation, presence of outliers and partial information which maintains many more valuable data sources out of usual data warehouses. Second is agility that performs the updation and synchronization processes in fast data evolving environments. Finally, depth refers to the deep analysis and machine learning processes of data that are needed to assist the analytics mechanism and which are more sophisticated than the traditional drill downs and roll-ups in complicated statistical analysis.

4.1 Data Centric Issues

Big data proceeds with a wide variety of analytical issues. It needs a fairly large amount of advanced skills to perform any type of analysis on gigantic datasets that can be semi structured, structured or unstructured. The required type ofanalysis that is needed to be performed on datasets greatly relies on the attained results i.e., decision making mechanism (Yang. and Fong,2013) It is usually done by employing one of the following methods; either by determining the upfront which big data is relevant to or incorporating the huge data volumes in analysis. The main analytical challenging questions are:

- What does happen when data is varied and its volume increase so much and dealing with this is not known?
- Does the whole dataset need to be analyzed?
- Does all of the data need to be stored and where?
- What benefits would be taken from the data?
- What methodology would be employed to search the significant factors of data?

Storing and processing of such huge amounts of data is also a critical issue (Fisher, 2012). (Alexandrov, 2014). Almost everything and everyone on this planet is generating data which needs to stored and processed accordingly. However, we do not have sophisticated mechanisms that can satisfy the needs. One option is to keep our large datasets over the clouds but to get this data processed; it will be called back which would increase the communication and computation costs. Cloud storages are distributed in nature which might give rise to performance and capacity utilization issues as well in the course of BDA.

[3] Magnetism, Agility and Depth

4.2 Governance and Social Privacy

Privacy maintenance mainly depends upon the technical confinements including abilities of analysis, correlation and extraction of useful data from likely larger and sensitive datasets. Developments in BDA offer us mechanisms to utilize and extract those data from the datasets, making defilements of privacy comparatively easier for its users. Data that is used for analytics usually includes intellectual property or synchronized information. System architects who are dealing with big data have to ensure that data utilized is in line with the policies and regulations and it is protected.

In terms of social aspects, businesses usually gather information about the user and from the users to enhance their market value employing the user's.Similarly, literate individuals or we better say the people who have knowledge about the BDA and its services could take all the advantages leaving deprived the others and their morality in the society would be worsened as a result of their unawareness. These issues must be resolved by simplifying the underlying BDA procedures and its resulting knowledge so as to eliminate the discrimination factors and the knowledge be equally accessible to every living being on this social sphere.

4.3 BDA and Security

Analytics landscape is changing by big data. Situational awareness and information security can be improved by BDA e.g., big data analytics can be employed for various tasks like log files analysis, financial transactions, and also monitoring the network traffic in order to see suspicious and anomalous behavior, and creating the coherent views by correlating multiple information sources (Ghazal, 2013). Currently, fraud detection can be categorized as one of the most noticeable utilities that can be done using BDA in many grounds like insurance, health care and etc. The following evolution is anticipated in the perspective of bid data analytics in terms of network intrusion detection (Khan, 2014) as mentioned in **(Table 2)** as well.

Log analysis, network events and packets for forensics analysis and network intrusion detection are some of the most important problems. Technologies that were traditionally used have failed to furnish the techniques that could sustain long term and large scale analytics for numerous causes:

Table 1: Data Analytics for Intrusion Detection

1st Gen	Intrusion Detection Systems	Need of layered security is realized by security architects because it's impossible that a system with 100 percent protective security
2nd Gen	Security Information and Event Management (SIEM)	Different intrusion detection sensor alerts are managed and in enterprise settings rules was the big challenge SIEM systems filter and aggregate security alarms that have originated from various sources and pragmatical data is presented for further security analysis.
3rd Gen	BDA in Security	Big Data tools provide a significant development in pragmatical security intelligence by cutting down the time required for consolidation, correlation and contextualization the varied security occurrence information. Moreover, correlation of huge past data for forensic reasons.

- Economically it was not feasible to store and retaining large amounts of data. As a result after a fixed retention period most of the activity logs and other captured computer activities were erased from the systems.
- Large structured datasets were inefficient of performing analytics and complex queries because tools that were traditionally used did not enforce the technologies involved in big data.
- Conventional tools have not been designed to manage and analyze the unstructured data. Therefore, customary tools had defined very inflexible schemas.
- Tools of big data like regular expressions and Piglatin can query data in flexible formats.

Cluster computing infrastructures are used by big data systems. Novel emerging big data innovations like databases associated with the Hadoop ecosystem and stream processing are facilitating the analysis and storage of diverse datasets at an extraordinary speed and scale. Thesetechnologies have the ability to change the security analytics by:

- Gathering data at a very big scale that comes from multiplicity of external and internal enterprise sources like vulnerable databases;
- Data analytics is performed deeply;
- Security related information is provided in consolidated view, and (Gani, 2016)
- Synchronous analysis of streaming data is achieved (Jeong, Yoon-Su, and Seung-Soo Shin. 2016)

It is important to know that the tools which are used for big data still require deep system analysts and architectural support for having a profound understanding of the system to accurately construct the big data analysis tools.

5. CONCLUSION

Big data analytics is an emerging type of knowledge work with adequately of opportunities for study and productivity enhancements. BDA is new important avenue to know about how people interact with computing. Big data analysis success is depends on following technologies i.e., cloud computing, machine learning, data mining, stream processing, time series analysis and visualization. There are many challenges that big data analysis faces. Variety, velocity and volume create many challenges in search, storage, retrieval and visualization issues. There are conflicting and inconsistent circumstances during big data analysis. One such challenge is to handle properly different types of inconsistencies during analysis and preprocessing. Additional challenges are security, provenance, privacy and modeling. Our analysis says that the BDA is a fast-growing, powerful practice and for the social business it's a key enabler. Beside all of this big data is an issue that needs to be discoursed more in the coming eras.

REFERENCES:

Alexandrov, A., (2014). "The Stratosphere platform for big data analytics." The VLDB Journal 23.6: 939-964,.

Barlow, M. (2013). Real-Time Big Data Analytics: Emerging Architecture. "O'Reilly Media, Inc.",

Boyd, D and K. Crawford.(2012) "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." Information, communication & society 15.5: .662-679.

Bernstein, D.(2014) "The Emerging Hadoop, Analytics, Stream Stack for Big Data," in IEEE Cloud Computing, vol. 1, no. 4, 84-86, Nov. A. Rabkin and R. H. Katz, "How Hadoop Clusters Break," in IEEE Software, vol. 30, no. 4, 88-94,.

Cohen, J., B. Dolan, M. Dunlap, J.M. Hellerstein, and C.Welton, (2009) MAD skills: new analysis practices for big data. Proceedings of the VLDB Endowment, 2(2), 1481-1492.

Christine L. B., (2015) "References," in Big Data, Little Data, No Data:Scholarship in the Networked World , 1, MIT Press, .416

Dean J. and S. Ghemawat, (2008) "Mapreduce: Simplified data processing on large clusters," Commun. ACM, vol. 51, no. 1, 107–113.

Gray, J. (2003) "What next?: A dozen information-technology research goals," J. ACM, vol. 50, 1, 41–57.

Ghazal, A, (2013). "BigBench: towards an industry standard benchmark for big data analytics." Proceedings of the ACM SIGMOD international conference on Management of data. ACM.

Grolinger, K., W. A. Higashino, (2013) ."Data management in cloud environments: NoSQL and NewSQL data stores," J. of Cloud Comp: Advances, Systems and Application.

Gani, A., (2016) "A survey on indexing techniques for big data: taxonomy and performance evaluation." Knowledge and Information Systems 46.2: 241-284,.

Hua, Y., H. Jiang, Y. Zhu, D. Feng, and L. Tian, (2009) "Smartstore: A new metadata organization paradigm with semantic-awareness for next generation file systems," in Proc. of the Conf. on High Performance Computing Networking, Storage and Analysis, ser. SC '09. ACM, 10:1–10:12.

Jeong, Yoon-Su, and Seung-Soo Shin. (2016) "An efficient authentication scheme to protect user privacy in seamless big data services." Wireless Personal Communications 86.1, 7-19, 2016

Kaisler, S. , F. Armour, J. A. Espinosa and W. Money, (2013) "Big Data: Issues and Challenges Moving Forward", IEEE, 46th Hawaii International Conf. on System Sciences.

Khan, N., (2014) "Big data: survey, technologies, opportunities, and challenges"The Scientific World Jou.

Langheinrich, M. (2009). "A survey of rfid privacy approaches," Personal Ubiquitous Comput., vol. 13, no. 6, 413–421.

Marz, N., and J. Warren. (2015) Big Data: Principles and best practices of scalable realtime data systems. Manning Publications Co.,

Olston, C. B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, (2008) "Pig latin: A not-so-foreign language for data processing," in Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '08. New York, NY, USA: ACM, 1099–1110.

Yang H. and S. Fong, (2013) "Countering the concept-drift problem in Big Data using iOVFDT," IEEE International Congress on Big Data,.

Zikopoulos, P., and E. Chris. (2011).Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media,

Zaslavsky, A., C, and D. Georgakopoulos. (2013) "Sensing as a service and big data." arXiv preprint arXiv:1301.0159.