A Study on Big Data Analytics: Platforms and Tools, Challenges, Technologies and Key Applications

Fathima Sihnas Fanoon Abdul Raheem[1], Isuri Uwanthika Gatamanna Arachchige[2]

Postgraduate Institute of Science, University of Peradeniya
fanoonarfs@gmail.com

Abstract. Big data is the paradigm that refers to the data generated not only in terabytes, but in hexabytes and more. Big data has gained an enormous success in numerous application areas including social media, economy, finance, healthcare, agriculture, and many more. Now a days, data are being produced at lighting speed. Data produced in many media are either structured, unstructured or semi-structured, using which the researchers could find the unknown pattern behind these datasets. Big Data is one of the most important applications in parallel and distributed systems. Processes related to analysis are often carried out with a deadline and during the process, it is vital for analytics also to concern about the quality of the data. This paper highlights few platform and tools used in the data analytics process. Further, it explains about the several challenges faced in big data analytics. Data representation, data management, data confidentiality and few other are found to be the major challenges in big data analytics. Also, the paper aims to study the technologies applied for big data analysis in the fields of cloud computing, Internet of Things and Hadoop. Finally, the paper reviews on the key application areas of big data analytics by researchers.

Keywords: Big Data, Cloud Computing, Hadoop, Internet of Things

1. Introduction

The volume of data which is being produced in massive amount, every day, exponentially, according to researchers, is defined as Big Data. The present world is moving towards the development of the internet and online technologies which includes big and powerful data servers, and as a result of this, huge amount of data and information are being generated from various resources and services which are available now a days. The main reason for these data generation is the interactions by and about people and things. Social media is one such platform which contributes largely in the massive generation of data. According to a survey in 2011, the amount of data has enlarged by nine times in amount within just five years of time and it also indicates that the amount would grow even more, around 35 trillion gigabytes, by the year 2020 [1].

This paper aims to study on what actually big data analytics is and what technologies are used in analysing the data produced every day and what challenges the industries face in big data processing. The key application areas where big data could be applied is also studied in the paper.

Big data analytics is the long process of exploring large amount of data to study the hidden patterns, correlations and other insights that could help the industries to study their business trends. Compared to the traditional methods, the present big data analytic technologies are more efficient and speed, helping the organizations to make quick decisions with regard to their business. Also, this makes the organizations to work at a faster rate and stay agile [2].

© Faculty of Management and Commerce
South Eastern University of Sri Lanka (SEUSL)
University Park, Oluvil 32360. Sri Lanka

Received: 12th Aug, 2020
Revised and Accepted: 25th Sep, 2020

Tom Devenport, the research director of IIA, in his report, Big Data in Big Companies mentions the following areas, in which the 50 organizations he interviewed, got the advantages. The major benefits identified are illustrated in the figure 1 below.

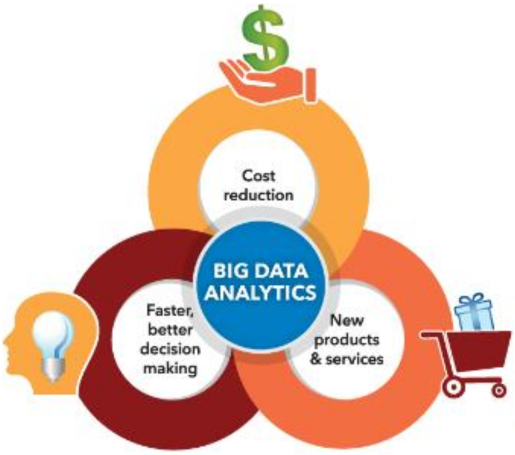

Fig. 1. Big Data Analytics Organizational Benefits

Cost Reduction: Hadoop and cloud-based analytic technologies shows significant cost reduction gains for storing large amount of data and thus able to identify efficient methods to carry out businesses.

Faster, better decision making: Hadoop technology being faster along with the in-memory analytic techniques, the information analysis becomes faster and help the organizations to make quicker business decisions based on the analysis results.

New Products and Services: Organizations are able to manufacture products that satisfy the need of the customers based on the analytics results.

Big data plays a vital role in every kind of organizations including healthcare services in the current era. Accordingly, the following types of data are defined as Big Data [3]:

- Traditional Enterprise Data – constitutes of data related to customer from CRM systems, transactional ERP data, web store transactions and general ledger data.
- Machine-generated/Sensor Data – contains information from Call Detail Records (CDR) weblogs, smart meters, manufacturing sensors, equipment logs, and training systems data.
- Social Data – these are data received from the customer feedback streams, micro-blogging sites (e.g. Twitter) and social media (e.g. Facebook).

Volume is the visible parameter of big data, though it is characterized by four key characteristics, 4Vs, [3] as given in the figure 2 below:

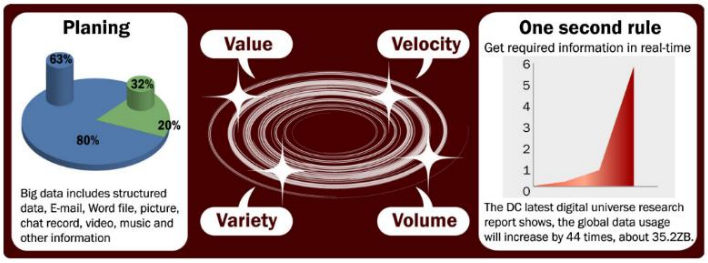
Planing
63%
32%
20%
80%
Big data includes structured data, E-mail, Word file, picture, chat record, video, music and other information
Value
Velocity
Variety
Volume
One second rule
Get required information in real-time
6
5
4
3
2
1
0
The DC latest digital universe research report shows, the global data usage will increase by 44 times, about 35.2ZB.

Fig. 2. 4Vs of Big Data (Chen M et al, 2014)

- Volume – large amounts of data are generated from machines when compared to non-traditional data.
- Velocity – large entry of thoughts and relationships which are useful for CRMs are generated through social media data streams.
- Variety – traditional data are demarcated by a data scheme and shows a slow transformation whereas non-traditional data show an unsteady rate of change.
- Value – there is a significant change in the values of different types of data economically. Though non-traditional data consists of useful information hidden in them, it is a challenge for the people to identify that particular data and extract the data for analysis.

The process one adopts to identify the unseen patterns, correlations which are unknown, trends in the market, and other useful information related to business from large amount of heterogeneous types of data generated from various data sources is defined as Big Data Analytics. These analytical findings help business people to achieve a high growth in a short period of time. Therefore, big data analytics is an area in which advanced analytic techniques are applied on big data sets. This process involves mainly of two things, combination of big data and analytics [4].

The organizations face a huge challenge in managing this large amount of data, which is heterogeneous, and therefore, a proper data management technique is essential in order to proceed with big data analytic actions. The actions will include event correlation, metric calculation and statistic preparation with analysis.

The organization of the paper is as follows. Section 1 introduces what big data analytics is and section 2 is on the platforms and tools related to big data analysis process. The next section goes on to explain the challenges faced in big data. The technologies adopted in the process of big data analytics is given in section 4. Section 5 explains the readers about the key application of big data. Finally, the paper concludes with few further recommendations related to big data analytics.

2. Big Data Analytic platforms and Tools

Management of massive amount of data does not require super computer or high cost as a result of the enhancement in the computing technology. Various types of big data analytic tools have been industrialized by computer professionals and the preference of a tool is dependent on the data set nature, the complexity of the problem to be analyzed, application of algorithm and the analytical solution, capability of the system, data security, performance and the scalability. Some of the widely used big data analytic platforms and tools are explained in detail in this section[5].

2.1. Apache Mahout

Apache Mahout is an open source framework for processing huge datasets for machine learning using MapReduce. The software consists of several Java libraries, providing algorithms which are optimized to convert the Java machine learning tasks into MapReduce jobs.

2.2. R

A programming language for advanced analysis of larger data sets using Hadoop. R has a complete set of classification models when compared to Mahout but limited to object-oriented programming language which may lead to memory management issues. Therefore, it is recommended to use R along with Mahout since R is much suitable for processing smaller datasets whereas Hadoop/Jaql can be applied for larger datasets processing.

2.3. Alteryx

An advanced data analytic platform which can be used for blending data and processes related to internal businesses, tools for third-party usage and data centers on cloud could be merged. Also, Alteryx allows the usage of data analytic tools in a single workflow.

2.4. Google Cloud Platform (GCP)

It is one of the leading cloud Application Programming Interfaces (API), though it was established recently. GCP could be used for developing simple websites to complex applications using the physical assets and the virtual resources available in the cloud datacenters.

2.5. H2O

H2O is a data processing and assessing tool, which in addition provides the users parallel processing, analytics, math and machine learning libraries and supports programming languages Java, R, Python and Scala. Also, it has a user-friendly interface which can be easily used by people with lack of programming knowledge.

2.6. MicroStrategy

MicroStrategy is an integrated platform which stores the data to be analyzed in Hadoop clusters and then later could be accessed from the desktop computers and mobile devices. It is a real-time visualization and an interaction platform to derive decisions.

2.7. RapidMiner

An analysis platform that does not require programming knowledge and offers an interface for designing data analysis models. Data files of various formats could be imported into this application and later be processed for analysis depending on the attributes. It offers an advantage of a better predictive and descriptive model of he data processed.

2.8. Datameer

Datameer Analytics Solution (DAS), is a platform that could be integrated with Hadoop for business purposes assisting business users with reports, charts and dashboards. Both structured and unstructured data could be processed using Datameer.

2.9. Microsoft

A platform that is capable of providing predictive analysis is provided by Microsoft known to be SSAS that is integrated along with the SQL server. This platform offers the easiness to integrate to Azure's cloud data platform and web deployment service and a way to simply utilize it by the data scientists.

3. Big Data Challenges

Overflow of data in big data causes huge challenges in the acquisition of data, storage of data, management and analysis of data. Relational Database management System (RDBMS) is the traditional method to manage and analyse data which is only applicable for structured data, other than semi-structured or unstructured data. Also, RDBMs require more utilization of expensive hardware and it is unable to handle data those are heterogenous and in larger volume. The researchers have proposed some solutions for maintaining and storing big data, but there are obstacles in their development. The key challenges in their development are [6].

3.1. Data Representation

Data show heterogeneity in type, structure, semantics, organization, granularity, organization, granularity and accessibility. The main aim of data representation is that it makes data more meaningful for computer analysis and user interpretation. A misrepresentation of data will result in the reduction of the data validity from its originality and sometimes, cause ineffectiveness in data analysis. Data represented in an efficient manner will reveal data structure, class, and type, as well as integrated technologies, which will assist in enabling efficient operations on various datasets.

3.2. Reduction in redundancy and compression of Data

Datasets have a great redundancy in general. This reduced redundancy and compression of data are essential in order to lessen the unintended cost on the whole system on the idea that there is no effect on the potential data values.

3.3. Management of data life cycle

Ubiquitous sensing and computing generate data at an unpredicted rates and scales when compared with the amount of data generated from the moderately slow developments of storage systems. The main issue in this regard is that the existing storage system is unable to maintain such massive data. Therefore, it is very important to develop a principle on what data to store and what data to discard.

3.4. Analytical Mechanism

The analysis of big data involves the processing of mass amount of diverse data within a given period of time. These tasks cannot be performed by the traditional RDBMs. In contrast, non-relational databases show an advantage in processing unstructured data and they have become the mainstream of big data analysis. But they too show few problems in relation to their performance and particular applications.

3.5. Data confidentiality

There are no effective maintenance or analysis of big data by most of the owners or service providers of big data at the present due to their limited capacity, which make them depend on professionals or tools for data analysis purpose. This will escalate the potential safety risks. But data analysis must be delivered to a third party only when there are proper preventive measures taken to protect sensitive data in order to ensure their safety.

3.6. Energy management

From the perspective of both economy and environment, consumption of energy of mainframe computing systems have gained more attention. As a result of increase in the data volume and analytical demands, processing, storage and transmission of big data consume more electrical energy. Therefore, a system-level power consumption control and management mechanism must be established.

3.7. Expandability and scalability

The system that is developed for data analytics should support both the present and future datasets. Also, the algorithms must be able to analyse even more large and complex data sets.

3.8. Cooperation

Big data analysis is an interdisciplinary field which requires researchers from different fields to cooperate to provide the potentiality of big data. This may require the establishment of a complete architecture for big data network in order to provide access to experts from various fields so that the analytical objectives could be achieved.

4. Big Data Technologies

Researchers on big data field have proposed some solutions in the process of data analysis. The development of these innovative technologies and platforms could be applied to develop several big data applications. This section will give an idea on several technologies related to big data.

4.1. Cloud Computing

Cloud computing is closely related to big data. Cloud computing aims to provide efficient big data applications with the usage of larger computing and storage resources that has a concentrated management. It also aims to provide applications which are computing capacity efficient. Issues related to storage and big data processing could be solved through cloud computing.

There are several overlapping technologies between cloud computing and big data, but they vary by two features. One is that they differ by their concept to a certain extent. IT architecture is transformed in cloud computing whereas big data is dependent on cloud computing as the ultimate infrastructure of even operation.

Secondly, it differs on the type of customers. Cloud computing targets Chief Information Officers (CIOs) while big data targets Chief Executive Officers (CEOs).

Cloud computing provides system-level resources with the features same as in computers and cloud computing supports the upper layer in order to operate operating systems and big data. This combined technology results in functions which are same as database and data processing dimensions that are efficient. Kissinger, President of EMC, suggested cloud computing could be a base technology for big data application [7].

4.2. IoT and Big Data

Networking sensors in enormous amount are embedded into the machines in as IoT paradigm. These sensors are deployed in a way that they can collect data in different fields. The characteristics of data generated from this technology differ as a result of the difference in the forms of data obtained. The most conventional features are heterogeneity, diversity, unstructured feature, noise and high redundancy.

HP forecasts that IoT will be the most important chunk of big data by the year 2030, although it is not much dominant now a days. Intel reports that the presence of abundant terminals generating masses of data, semi-structured or unstructured generation of data and data being useful only when it is analysed are the main features for IoT to become the big data paradigm.

It has been become a compulsion to embrace big data for IoT applications though the advancement of big data is falling behind. Also, it has been predicted that these two technologies are dependent on each other and should jointly be technologically advanced [8].

4.3. Data center

Considering big data, data center is a vital component. It is not only a platform for determined data, but also has various accountabilities, like data acquisition, data management, data organization and controlling data values and its operations. Data are the major concerns of data centers. The core for supporting big data in the emergence of physical data center network, but currently, it has become the significant infrastructure that is essentially utmost [9].

- Data centers deliver a dominant backstage support required by big data. Here, data center provides an infrastructure that has many nodes, high-speed inside network, effective dissipation of heat, and effective backup of data. But the normal operation of big data could be ensured only if a vastly energy proficient, even, safe, pliant and redundant data center is fabricated.
- Innovation of data centers is accelerated as a result of speedy growth in the applications of big data. Performance of data centers could be enhanced by processing and computing large volume of structured and unstructured data, and also by the sources of analytical data. While developing data centers, it is also mandatory to focus on how to diminish the operational cost.

4.4. Hadoop and Big Data

Hadoop is the most extensively used application in big data industry at present. There are also significant number of academic research being carried out using Hadoop. According to a survey, Yahoo runs Hadoop in 42,000 servers at four data centers in order to support its services. Many companies, now a days, provide Hadoop commercial implementation or support or both, including Cloudera, IBM, MapR, EMC and Oracle.

Sensors are deployed to collect data from industrial machineries and systems. CloudView is a framework proposed for data organization and cloud computing infrastructure, which uses miscellaneous architectures, local nodes, and remote clusters centered on Hadoop for analysis purpose. Clusters are implemented based on Hadoop for offline analysis of complex data [10].

5. Big Data Applications

Wide range of applications are involved in big data analysis, which are extremely complex. There are six most important data analysis fields and this section reviews the key applications of big data.

5.1. Text Data Analysis

Text is the most common format used to store information. Text analysis is a process which involves the extraction of useful information and knowledge from unstructured text. In contrast, text mining is an inter-disciplinary area where information retrieval, machine learning, statistics, computing linguistics and data mining activities are involved. Natural Language Processing (NLP) and text expressions are the basic concepts applied for text analysis. Data analysis, data interpretation and text generation are supported by NLP [11].

5.2. Web Data Analysis

Web data analysis is one of the most emerging technologies in research areas which intentions to spontaneously retrieve, extract, and assess facts from web documents and services in order to derive knowledge. Web analysis is inter-connected with several other fields such as database, information retrieval, NLP and text mining. Further, web analysis is classified into three fields as Web content mining, Web structure mining and web usage mining [12].

5.3. Multimedia Data Analysis

Multimedia data is growing rapidly to extract expedient knowledge and understand the uniqueness of data through analysis. Multimedia data are diverse and most of these data consists of more affluent information compared to structured or text data. Extraction of data from multimedia data is a challenging task. Many disciplines are inter-related with multimedia data analysis like multimedia summarization, multimedia annotation, multimedia index and retrieval, multimedia suggestion and etc [13].

5.4. Network Data Analysis

Now a days, large amount of data are generated as a result of the usage of social networking services including Facebook, Twitter and LinkedIn. The data generated here are text, images and other network multimedia data.

The existing research on social media contexts, according to the data-centred view, is categorized into two sets as link-based structural analysis and content-based analysis [14].

5.5. Mobile Data Analysis

Usage of mobile is growing at a rapid rate in the current era due to the advancing technologies. There are more than 65,000 mobile applications available and the monthly mobile data flow is increasing rapidly. Mobile data analysis has few challenges and it has distinctive features such as mobile sensing, moving flexibility, noise and a large amount of redundancy. There are various fields in which research on mobile data analysis is being carried out [15].

6. Conclusion

Big Data has established its roots in every fields including science and engineering. Today, enterprises are on the verge of exploring big data as to discover the unknown facts and patterns in their business process, primarily, those depend on bulk of clients. Application of innovative analytic techniques will help the organizations to comprehend the progress of their business and pursue the still evolving regards such as customer behavior. The paper reviews the background study of the big data and its state of art. The first section introduces about the background of big data which is then followed by the platform and tools used for analysis, then the challenges researchers face in big data analytics process. Also, the applied technologies related to big data are clearly explained in the paper, to provide the readers a big picture of this emerging area.

References

[1] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, “A survey of machine learning for big data processing,” *Eurasip Journal on Advances in Signal Processing*, vol. 2016, no. 1. Springer International Publishing, 01-Dec-2016.

[2] “Big Data Analytics - What it is and why it matters | SAS.” [Online]. Available: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html. [Accessed: 31-Oct-2019].

[3] Jainendra Singh, “Big Data Analytic and Mining with Machine Learning Algorithm,” *Int. J. Inf. Comput. Technol.*, vol. 4, no. 4, pp. 33–40, 2014.

[4] P. Russom, “Big Data Analytics Fourth Quarter 2011 Tdwi Re Se A Rch Co-sponsored by Big Data A N A Ly Tic S Fourth Quarter 2011 Tdwi Best Practices Report Introduction to Big Data Analytics,” 2011.

[5] S. Al-Shiakhli, “Big Data Analytics: A Literature Review Perspective,” 2019.

[6] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, 2014.

[7] B. M. Purcell, “Big data using cloud computing Big data using cloud computing,” no. October, 2016.

[8] V. E. Balas, V. K. Solanki, R. Kumar, and M. Khari, “Internet of Things and Big Data Analytics for Smart Generation,” *Springer*, vol. 154, no. January, p. 309, 2019.

[9] Kaushik Pal, “How Big Data Impacts Data Centers,” 2015. [Online]. Available: https://www.techopedia.com/2/31217/technology-trends/big-data/how-big-data-impacts-data-centers. [Accessed: 31-Oct-2019].

[10] “What is Hadoop? | SAS.” [Online]. Available: https://www.sas.com/en_us/insights/big-data/hadoop.html. [Accessed: 31-Oct-2019].

[11] M. P. Bach, Ž. Krstič, S. Seljan, and L. Turulja, “Text mining for big data analysis in financial sector: A literature review,” *Sustain.*, vol. 11, no. 5, 2019.

[12] G. Zheng and S. Peltsverger, “Web Analytics Overview,” *Encycl. Inf. Sci. Technol. Third Ed.*, no. January, pp. 7674–7683, 2014.

[13] S. Pouyanfar, Y. Yang, S. C. Chen, M. L. Shyu, and S. S. Iyengar, “Multimedia big data analytics: A survey,” *ACM Comput. Surv.*, vol. 51, no. 1, 2018.

[14] E. D. Kolaczyk, “Tutorial: Statistical Analysis of Network Data.”

[15] S. Abolfazli and M. R. Lee, “Mobile Data Analytics,” *IT Professional*, vol. 19, no. 3. IEEE Computer Society, pp. 14–16, 2017.