BIG DATA ANALYTICS: CHALLENGES, TECHNOLOGIES AND KEY APPLICATIONS

Fanoon Raheem[1], Isuri Uwanthika[2]
[1,2]Postgraduate Institute of Science, University of Peradeniya
fanoonarfs@gmail.com, isuriuwanthika@gmail.com

ABSTRACT

Big data is the paradigm that refers to the data generated not only in terabytes, but in Exabyte 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 day, 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. In this review paper, it explains about the several challenges faced in the 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 including 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

Introduction

The amount of data that is generated in massive amount, daily, exponentially is defined as Big Data (Dogra D, 2018). 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 day. 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 grown by nine times in volume within just 5 years and it also indicates that the amount would grow even more, around 35 trillion gigabytes, by the year 2020 (Qiu, Wu, Ding, Xu, & Feng, 2016).

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 ("Big Data Analytics - What it is and why it matters | SAS," n.d.).

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.

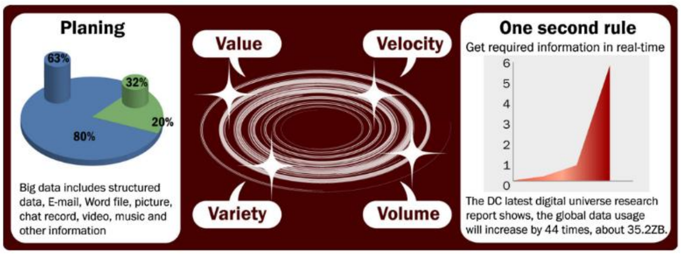
Planing
63%
32%
20%
80%
Big data includes structured data, E-mail, Word file, picture, chat record, video, music and other information
Value
Velocity
Variety
Volume
One second rule
Get required information in real-time
6
5
4
3
2
1
0
The DC latest digital universe research report shows, the global data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

Figure 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 (Jainendra Singh, 2014):

- 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 characterised by four key characteristics, 4Vs, (Jainendra Singh, 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.

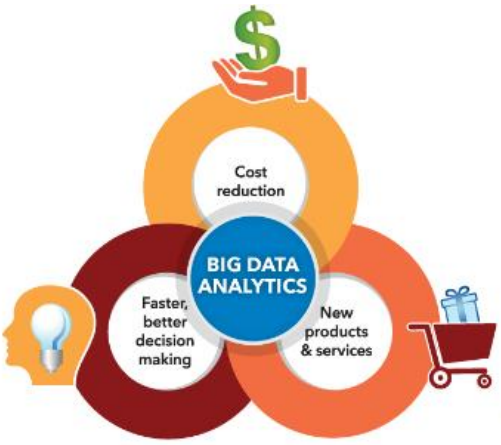

Figure 2. *4Vs of Big Data (Chen M et al, 2014)*

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 (Russom, 2011).

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 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 3. Section 4 explains the readers about the key application of big data. Finally, the paper concludes with few further recommendations related to big data analytics.

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 (Chen, Mao, & Liu, 2014);

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 may even cause ineffective data analysis. Data represented in an efficient manner will reflect data structure, class, and type, as well as integrated technologies, which will assist in enabling efficient operations on various datasets.

Reduction in redundancy and data compression

Datasets have a high redundancy in general. This reduced redundancy and compression of data are essential in order to reduce the indirect cost on the entire system on the premise that there is no effect on the potential data values.
e.g.: sensor networks produce huge amount of data which are highly redundant, and they are later filtered and compresses at orders of magnitude.

Management of data life cycle

Pervasive sensing and computing generate data at an unpredicted rates and scales when compared with the amount of data generated from the relatively slow advances of storage systems. The main issue in this regard is that the current storage system is unable to support such massive data. Therefore, it is very important to develop a principle on which data to be stored and which data to be discarded.

Analytical Mechanism

The analysis of big data involves the processing of mass amount of heterogeneous 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.

Data confidentiality

There are no effective maintenance or analysis of big data by most of the big data service providers or owners at the present due to their limited capacity, which make them rely on professionals or tools for data analysis purpose. This will increase 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.

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.

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.

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 comprehensive big data network architecture in order to provide access to scientists from various fields so that the analytical objectives could be achieved.

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 gig data applications. This section will give an idea on several technologies related to big data.

3.1 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 differ by two aspects. 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 fundamental infrastructure of smooth 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 capacity that are efficient. Kissinger, President of EMC, suggested that cloud computing could be a base technology for big data application (Purcell, 2016).

3.2 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 types of data collected. The most classical characteristics are heterogeneity, variety, unstructured feature, noise and high redundancy.

HP forecasts that IoT will be the most important part 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 adopt big data for IoT applications though the development of big data is lagging behind. Also, it has been recognized that these two technologies are inter-dependent and should jointly be developed (Balas, Solanki, Kumar, & Khari, 2019).

3.3 Data center

Considering big data, data center is a vital component. It is not only a platform for concentrated data, but also has various responsibilities, like data acquisition, data management, data organization and leveraging data values and functions. 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 key infrastructure that is most urgently required (kaushik Pal, 2015).

- Data centers provide a powerful backstage support required by big data. Here, data center provides an infrastructure that has many nodes, high-speed internal network, effective dissipation of heat, and effective backup of data. But the normal operation of big data could be ensured only if a highly energy efficient, stable, safe, expandable and redundant data center is built.
- Innovation of data centers is accelerated as a result of rapid 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 reduce the operational cost.

3.5 Hadoop and Big Data

Hadoop is the most widely used application in big data industry at present. There are also significant number of academic researches 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 products and services. Many companies, now a days, provide Hadoop commercial execution 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 mixed architectures, local nodes, and remote clusters based on Hadoop for analysis purpose. Clusters are implemented based on Hadoop for offline analysis of complex data ("What is Hadoop? | SAS," n.d.).

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.

4.1 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 (Bach, Krstič, Seljan, & Turulja, 2019).

4.2 Web Data Analysis

Web data analysis is one of the most emerging technologies in research areas which aims to automatically retrieve, extract, and evaluate information 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 (Zheng & Peltsverger, 2014).

4.3 Multimedia Data Analysis

Multimedia data is growing rapidly to extract useful knowledge and understand the uniqueness of data through analysis. Multimedia data are heterogeneous and most of these data consists of richer 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 (Pouyanfar, Yang, Chen, Shyu, & Iyengar, 2018).

4.4 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 classified into two categories as link-based structural analysis and content-based analysis (Kolaczyk, n.d.).

4.5 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 unique characteristics 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 (Abolfazli & Lee, 2017).

Conclusion

Big Data has established its roots in every fields including science and engineering. Today, enterprises are on the verge of exploring big data in order to discover the unknown facts and patterns in their business process, especially those depend on mass consumers. Application of advanced analytic techniques will help the organizations to understand the current state of their business and track still-evolving aspects 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 challenges the researchers face in big data analytics process. Also, the applied technologies related to big data are clearly explained in the paper. These discussions aim to provide a comprehensive overview and big picture to readers of this exciting area.

References

Abolfazli, S., & Lee, M. R. (2017). Mobile Data Analytics. *IT Professional*, Vol. 19, pp. 14–16. https://doi.org/10.1109/MITP.2017.38

Bach, M. P., Krstič, Ž., Seljan, S., & Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. *Sustainability (Switzerland)*, *11*(5). https://doi.org/10.3390/su11051277

Balas, V. E., Solanki, V. K., Kumar, R., & Khari, M. (2019). Internet of Things and Big Data Analytics for Smart Generation. *Springer*, *154*(January), 309. https://doi.org/10.1007/978-3-030-04203-5

Big Data Analytics - What it is and why it matters | SAS. (n.d.). Retrieved October 31, 2019, from https://www.sas.com/en_us/insights/analytics/big-data-analytics.html

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

Jainendra Singh. (2014). Big Data Analytic and Mining with Machine Learning Algorithm. *International Journal of Information and Computation Technology*, *4*(4), 33–40.

kaushik Pal. (2015). How Big Data Impacts Data Centers. Retrieved October 31, 2019, from https://www.techopedia.com/2/31217/technology-trends/big-data/how-big-data-impacts-data-centers

Kolaczyk, E. D. (n.d.). *Tutorial: Statistical Analysis of Network Data*.

Pouyanfar, S., Yang, Y., Chen, S. C., Shyu, M. L., & Iyengar, S. S. (2018). Multimedia big data analytics: A survey. *ACM Computing Surveys*, *51*(1). https://doi.org/10.1145/3150226

Purcell, B. M. (2016). *Big data using cloud computing Big data using cloud computing*. (October).

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016, December 1). A survey of machine learning for big data processing. *Eurasip Journal on Advances in Signal Processing*, Vol. 2016. https://doi.org/10.1186/s13634-016-0355-x

Russom, P. (2011). *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*. Retrieved from https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf

What is Hadoop? | SAS. (n.d.). Retrieved October 31, 2019, from https://www.sas.com/en_us/insights/big-data/hadoop.html

Zheng, G., & Peltsverger, S. (2014). Web Analytics Overview. *Encyclopedia of Information Science and Technology*, *Third Edition*, (January), 7674–7683. https://doi.org/10.4018/978-1-4666-5888-2.ch756