International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 25, 2017 ISSN: 2348-6600
Reference ID: IJCS-276 PAGE NO: 1789-1793
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

Big data: Issues, Challenges and Tools.

M.Priya M.E.,
Teaching Assistant,
Alagappa Institute of Skill Development.
Alagappa University.
Karaikudi, India.
priyamathymail@gmail.com.

Dr.C.Balakrishnan,
Assistant Professor,
Alagappa Institute of Skill Development.
Alagappa University.
Karaikudi, India.
balasjc@gmail.com

Abstract- Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. We discussed big data characteristics in introduction. The primary purpose of this paper is to present the issues and challenges of big data analytics. The challenges include quality of data, capturing, analysis, processing, storage, sharing, and privacy violations. The paper concludes with description of big data analytics tools Hadoop, Hive, PIG, WibiData, PLATFORA, SkyTree.

Keywords - Big data, Big data analytics Tools, Techniques, Hadoop, Hive.

1. INTRODUCTION

Data is growing at a huge speed making it difficult to handle such large amount of data. Data are generated from online transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, social networking interactions, science data, sensors and mobile phones and their applications. They are stored in databases grow massively and become difficult to capture, form, store, manage, share, analyse and visualize via typical database software tools.The main difficulty in handling such large amount of data is because that the volume is increasing rapidly in comparison to the computing resources. 5 exabytes (1018 bytes) of data were created by human until 2003. Today this amount of information is created in two days. Today we are in digital world of data. Data is created to 8 zettabytes (1021 bytes). It is predicted to double every two years [8]. IBM indicates that every day 2.5 exabytes of data created also 90% of the data produced in last two years [6].

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 25, 2017 ISSN: 2348-6600
Reference ID: IJCS-276 PAGE NO: 1789-1793
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

Big data can be defined with the following properties associated with it:

Variety

Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web Pages, Web Log Files, social media sites, e-mail, documents, sensor devices data both from active passive devices. All this data is totally different consisting of raw, structured, semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems.

Volume

The Big word in big data itself defines the volume. At present the data existing is in petabytes and is supposed to increase to zetta bytes in nearby future. The social networking sites existing are themselves producing data in order of terabytes every day and this amount of data is definitely difficult to be handled using the existing traditional systems.

Velocity

Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows. For example the data from the sensor devices would be constantly moving to the database store and this amount won't be small enough. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion.

2. RELATED WORK

According to [3], Big Data is a collection of large and complex data sets. It deals with this huge amount of data and executing it on multiple nodes there is a higher risk of having bad data and even data quality issues may exist at every stage of the processing. They found out three major issues such as increasing need for integration of huge amount of data available, instant data collection and deployment issues, real time scalability issues.

In [2] BrijeshB.Mehta and UdaiPratapRao discussed about privacy issues in various areas like, mobile data, healthcare, social media, and web usage data mining. They found the challenges with existing preserving techniques for unstructured big data and proposed a novel approach, which can extract a sensitive attributes from unstructured data using some of the machine learning techniques to preserve privacy of an individual. In [1] Ibrahim et al., presented a review on the rise of big data in cloud computing and proposed a classification for big data, a conceptual view of big data, and a cloud services model. This model was compared with several representative big data cloud platforms. They also reviewed some of the challenges in big data processing. The review covered volume, scalability, availability, data integrity, data protection, data transformation, data quality/heterogeneity, privacy and legal/regulatory issues, data access, and governance. Furthermore, the key issues in big data in clouds were highlighted. They also discussed the background of Hadoop technology and its core components, namely, MapReduce and HDFS.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 25, 2017 ISSN: 2348-6600
Reference ID: IJCS-276 PAGE NO: 1789-1793
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

3. BIG DATA ISSUES AND CHALLENGES

Recent years big data has been accumulated in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches. Here the challenges of big data analytics are discussed as follows.

Quality of Data: Collection of huge amount of data and its storage comes at a cost. More data if used for decision making or for predictive analysis in business will definitely lead to better results. Big data basically focuses on quality data storage rather than having very large irrelevant data so that better results and conclusions can be drawn. This further leads to various questions like how it can be ensured that which data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not to draw conclusions from it etc.

Heterogeneous Data: Unstructured data represents almost every kind of data being produced like social media interactions, to recorded meetings, to handling of PDF documents, fax transfers, to emails and more. Structured data is always organized into highly mechanized and manageable way. It shows well integration with database but unstructured data is completely raw and unorganized. Working with unstructured data is cumbersome and of course costly too. Converting all this unstructured data into structured one is also not feasible. Structured data is the one which is organized in a way so that it can be managed easily.

Storage and Processing Issues: The storage available is not enough for storing the large amount of data which is being produced by almost everything. Social Media sites are themselves a great contributor along with the sensor devices etc. Because of the rigorous demands of the Big data on networks, storage and servers outsourcing the data to cloud may seem an option. Uploading this large amount of data in cloud doesn't solve the problem. Since Big data insights require getting all the data collected and then linking it in a way to extract important information. Terabytes of data will take large amount of time to get uploaded in cloud and moreover this data is changing so rapidly which will make this data hard to be uploaded in real time. At the same time, the cloud's distributed nature is also problematic for Big data analysis. Thus the cloud issues with Big Data can be categorized into Capacity and Performance issues.

Scalability:The scalability issue of Big data has lead towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters. This requires a high level of sharing of resources which is expensive and also brings with it various challenges like how to run and execute various jobs so that we can meet the goal of each workload cost effectively. It also requires dealing with the system failures in an efficient manner which occurs more frequently if operating on large clusters.
Privacy and Security issue: It is the most important issue with big data which is sensitive and includes conceptual, technical as well as legal significance. The personal information of a person when combined with external large data sets leads to the

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 25, 2017 ISSN: 2348-6600
Reference ID: IJCS-276 PAGE NO: 1789-1793
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

inference of new facts about that person and it's possible that these kinds of facts about the person are secretive and the person might not want the Data Owner to know or any person to know about them. Information regarding the users (people) is collected and used in order to add value to the business of the organization. This is done by creating insights in their lives which they are unaware of.

4. TOOLS AND TECHNIQUES AVAILABLE

Hadoop:The name Hadoop has become synonymous with big data. It is an open-source software framework for distributed storage of very large datasets on computer clusters. Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. Hadoop is not for the data beginner. To truly harness its power, you really need to know Java. It might be a commitment, but Hadoop is certainly worth the effort – since tons of other companies and technologies run off of it or integrate with it. But Hadoop Map Reduce is a batch-oriented system, and doesn't lend itself well towards interactive applications; real-time operations like stream processing; and other, more sophisticated computations.

Hive:Hive is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. It amplifies the reach of Hadoop, making it more familiar for BI users.

PIG:PIG is another bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source.

WibiData:WibiData is a combination of web analytics with Hadoop, being built on top of HBase, which is itself a database layer on top of Hadoop. It allows web sites to better explore and work with their user data, enabling real-time responses to user behavior, such as serving personalized content, recommendations and decisions.

PLATFORA**:** Perhaps the greatest limitation of Hadoop is that it is a very low-level implementation of MapReduce, requiring extensive developer knowledge to operate. Between preparing, testing and running jobs, a full cycle can take hours, eliminating the interactivity that users enjoyed with conventional databases. PLATFORA is a platform that turns user's queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 25, 2017 ISSN: 2348-6600
Reference ID: IJCS-276 PAGE NO: 1789-1793
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

SkyTree:SkyTree is a high-performance machine learning and data analytics platform focused specifically on handling Big Data. Machine learning, in turn, is an essential part of Big Data, since the massive data volumes make manual exploration, or even conventional automated exploration methods unfeasible or too expensive.

5. CONCLUSION

Big Data is going to continue growing during the next years,and each data scientist will have to manage much more amount of data every year. This data is going to more diverse,larger,faster and is becoming the new scientific data research and for business applications. This paper described the characteristics of big data. We discussed the issues, challenges in big data analytics and available tools. Big data analysis helps business people to make better decisions and researchers to identify new opportunities.

REFERENCES

[1] Ibrahim AbakerTargioHashem, IbrarYaqoob , Nor BadrulAnuar , SalimahMokhtar , Abdullah Gani , SameeUllah Khan, "The rise of "big data" on cloud computing: Review and open research issues " Information Systems 47(2015) 98–115.

[2] BrijeshB.Mehta, UdaiPratapRao, "Privacy Preserving Unstructured Big Data Analytics: Issues and Challenges", International Conference on Information Security & privacy, 2016, 78, 120-124

[3] Naveen Garg ,Dr. Sanjay Singla, Dr.SurenderJangra , "Challenges and Techniques for Testing of Big Data", Procedia Computer Science 85 (2016) 940 – 948.

[4] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", IEEE, International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, 2012.

[5]Sam Madden, "From Databases to Big Data", IEEE, Internet Computing, May-June 2012.

[6] S. Singh and N. Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011.

[7] Martin Courtney, "The Larging-up of Big Data", IEEE, Engineering &Technology, September 2012.

[8] Intel IT Center, "Planning Guide: Getting Started with Hadoop", Steps IT Managers Can Take to Move Forward with Big Data Analytics, June 2012.