TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 14987 - 14991

A Study on Big Data Analytics in Cloud Computing Infrastructure

R. Priyadharsini[1], Dr. P. Logeswari[2]

[1,2]*Assist. Professor*, Department of Computer Scienc, Sri Krishna Arts & Science College Coimbatore.

Article Info
Volume 82
Page Number: 14987 - 14991
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 28 February 2020*

Abstract:
Wireless Sensor Network(WSN) become a major technology for sensing in different application areas. Secure routing is attacks in a networks. Routing is fundamental works in internet. Security routing in Wireless Sensor Networks (WSNs) are security important because of its usage in applications like monitoring, tracking, controlling, surveillance, smart home etc. The measures for secure routing including cryptography, keys, trust, reputation and secure localization. Routing protocols is designed for consideration of power consumption not only for security purpose. For many applications of WSN, security is important requirement.

Keywords:*Big data Analytics, Big Data in Cloud, Hadoop, Storm, Spark, Hadoop, Splunk, Cassandra, Mahout, Estimation, Prediction, Heterogeneity, Zookeeper, Master and Worker nodes.*

I. INTRODUCTION

In the current world, digitization leads to the generation of data from various sources. Big data is a recent breakthrough in many of the industries. Data is available in various formats and in huge volumes. One of the V's in IBM's definition of big data defines the veracity of the data i.e., availability of the data. One of the main applications is decision making. Based on the customer purchase pattern we can make sales or marketing decisions or based on the customer's interest you have to suggest products to them [4]. The Fig.1 represents main V's of big data which are volume, velocity, variety and Veracity. There has been an unprecedented increase in the quantity of data - Volume. Big data refers to the huge volume of data that has been generated worldwide. Velocity refers to the speed at which the data is generated. Variety means the different types of data. The term big data differs from place to place. For some companies, 10TB refers to big data others 100TB refers to big data. The data are generated in real-time and in a huge volume. The traditional analytical process includes extract, load, and transform data. Big data analytics refers to the storage, aggregation, and process of data [6].

Data is the building block of an organization. If they lost the customer or employee details or transaction details It would be great trouble. To store every detail such as the transaction history of the customer, employee details, and all their venture storage and processing of data at various levels is important to extract meaningful information [9].In this paper, we are going to analyse the meaning of big data, types of big data, it's evolution, techniques, methodologies, and challenges involving big data.

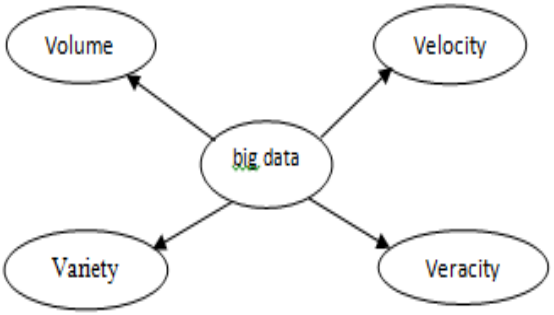

Fig.1 Big Data Proportions

January-February 2020
ISSN: 0193-4120 Page No. 14987 - 14991

Structured data is the data is arranged in a particular pattern. All the entries in it have the same attributes, format, and length. For instance, RDBMS. The Semi-structured entries have different structures or order. Graphical or scientifically processed data may be of this type.Unstructured data is coined as a Data with no proper structure. For instance, images and videos. Moreover, Multi-structured data is a combination of all the above. Data is becoming more valuable and important. Big data storage is a compute-and-storage architecture that collects and manages large data sets and enables real time statistics.

II.LITERATURE SURVEY

In 1919, The US commerce department partners with IBM to collect the data about more than 90000 enumerators and supervisors in one month to keep the federal employees busy using more than 200 million IBM supplied punched cards.In 1923, Criminal records are collected and analysed by the police department of Los Angles to improve criminal identification methods using IBM Tabulating Equipment.In 1934, The 405 Electric Tabulating Machine, a product by IBM is introduced to the tabulating industries before the invent of calculators in the 1940s. To process the punched cards and to print alphanumeric characters.In 1937, The US government partners with IBM to create and maintain the employment records of 26 million Americans and to implement the US Social Security Act,1935.In 1948, The 604 Electronic Calculating punch was introduced by IBM for faster data calculations beyond scientific calculation.In 1952, Drive Vacuum column, A magnetic tape was introduced by IBM. Magnetic tapes are used for data storage. It is the first digital storage. Even after 50 years, the cost-effective and high capacity storage medium is magnetic tapes.In 1956, the Random-Access Method of Accounting and Control (RAMAC) was introduced by IBM. The first magnetic hard disk for data storage. It can store about 2000 bits of data and costs around $10000 per megabyte [1].In 1969, The Apollo Mission which is designed by NASA and programmed by IBM led its way to the moon, The Apollo had so much information to relay and the computers represent in an electronic shorthand form.In 1970, IBM introduced relational databases. The information is stored in tabular form which is easy to understand. It transformed the older expensive process of finding the data. All the recently developed databases are based on relational databases. It entirely based on the relational model of data. Relational data have created a universally big model for storing data in the system.In 1988, IBM and the National Institute of Health formed a deal to increase their CPU's capacity and to work with the breakthroughs in the health industry.In 2002, The oxford university, IBM along with the UK government built a computer grid for the screening and diagnosis of breast cancer. It provides medical professionals with more help to treat the disease.In 2005, A landmark geographical project that maps the markers to modern people was found by IBM in collaboration with the National Geographic Society. This creates a picture of the movement of ancient humans to have a better understanding of human history [8]. In 2007, To identify patterns through the visual representation of data, IBM introduced Many Eyes to upload the data and to graphically represent it.In 2009, The marine institute of Ireland collaborated with IBM to develop The Smart bay project that collects environmental data through a network of sensors and to derive insights on pollution of the environment to better monitor it.In 2011, To understand natural language and to learn about the correlations that exist between the data, a computing system known as Watson is developed.In 2013, IBM partnered with top universities to include big data on the curriculum [5].Later Big Data became one of the top trends in all industries.

III. TOOLS USED FOR BIG DATA PROCESSING

Numerous tools are available for processing big data. Some of the emerging tools are Hadoop, MapReduce, Apache Spark, and storm. Hadoop follows batch processing. Apache Storm is used for

real-time data streaming. Using spark, the users can directly interact in real-time for analysis.

A. Apache Hadoop and MapReduce

Hadoop consists of Hadoop Distributed File System (HDFS), apache hive and pig. MapReduce follows parallel processing and it is based on divide and conquer method such as a map and reduce. Hadoop follows master-slave architecture. It works with two nodes such as master node and worker node. The master divides the complex problem into simpler subproblems in the map step. After finding solutions, it combines the output of every worker node in the reduce step [3].

B. Apache Mahout

Various classification, clustering, and regression techniques can be performed to provide better and scalable machine learning techniques. The companies include Google, IBM, Yahoo, Twitter, and Facebook use this tool [3].

C. Apache Spark

The components include driver, cluster manager, and worker. A Driver is the starting point of execution on a spark cluster. The cluster manager manages the resources on the spark cluster. Worker nodes perform all the tasks. Executors are responsible for executing the tasks. An open-source framework for big data processing [4].

D. Apache Storm

Apache Storm is similar to that of Hadoop. The users run different topologies on different storm clusters while Hadoop implements map-reduce. Hadoop follows batch processing while Apache Storm follows real-time data processing. The two kinds of nodes are the master node and worker node. The roles are nimbus and supervisor which is similar to job tracker and task tracker of map-reduce. Nimbus is responsible for assigning tasks and maintenance.The supervisor performs the tasks assigned to them by the nimbus. It starts and terminates the process according to the instructions of the nimbus [4].

E. Splunk

It is an intelligent platform that combines cloud technologies and big data. Cloud that enables the users to access shared pools of configurable systems resources and higher-level services that can be rapidly provisioned with minimal management effort.The goal of cloud computing is to allow users to take benefit from all of these technologies, without the need for deep knowledge about or expertise with each one of them. The cloud aims to cut costs, and helps the users focus on their core instead of being impeded by other obstacles.To provide support for business applications, generate real-time data searching and diagnose problems of information technology [4].

F. Open Refine

It is similar to Hadoop that follows batch processing. It is used to transform the data from one form to another. The data sources are web and external data [3]. Hence, we can clean it. It is dynamically baptized as data wrangling.

IV.TOOLS USED FOR BIG DATA STORAGE

A. Apache HBase

It is a Hadoop distributed database that is used to store data. It combines Hadoop Distributed File System with real-time data access and it runs on top of Hadoop. It is a NoSQL database. It stores the data as key-value pairs and has the capabilities of map-reduce [3].

B. Apache Cassandra

Cassandra is decentralized so there is no single point of failure. It provides high scalability and availability. Data is replicated and stored in multiple locations to become fault-tolerant. It provides high scalability and availability. The throughput increases linearly as new machines are added. Read and Write operations can be performed easily [3].Various other tools are available for data cleaning, filtering and visualization.

V.METHODS FOR BIG DATA PROCESSING

A. Classification

It classifies the data to a set of categories that aids in more accurate results. A data instance has to be classified into any one of the target classes which is already defined. For instance, whether a customer can be trusted or not in a credit card transaction

January-February 2020
ISSN: 0193-4120 Page No. 14987 - 14991

database given his purchase and payment history. It also provides systematic arrangement in to various groups and categories.

B. Estimation

Also known as regression. The output is numerical instead of categorical in classification. To determine a numerical value for the output attribute. For instance, Estimate the marks of a student who has done all the practice tests.

C. Prediction

It predicts the future outcome instead of the current behaviour. It is similar to classification. The output can be either categorical or numeric. For instance, Predict the next week's price of the material. It may be based on the facts.

D. Association Rules

The hidden rules from the larger transaction databases are found out. They are also known as association rules. For instance, consider the rule milk, butter -> biscuits represent the information that whenever milk and butter are purchased biscuit is also purchased. Placing these three items together can increase the overall sales of all three products. It also specifies about a large set of data items. This rule illustrate how can item sets occur frequently.

E. Clustering

To group the data into clusters based on similar characteristics. For instance, group the customers according to their performance or experience, etc [7]. We can break the data points and those data can be similar or dis-similar. clustering does not depend on predefined classes.

VI.CHALLENGE DIMENSIONS

Data is generating enormously which is hard for the computing resources to handle. The computing processors follow Moore's law and the chips are very small in size. We must find an effective way to handle the fast-growing data in today's world.

A. Data heterogeneity

Structured databases can be handled by SQL. When there is heterogeneity in data such as audio, video or webpage. the traditional tools won't work. The first step in data processing is to structure the data.

B. Data Timeliness

Based on the volume of the data, it will take time to analyse the data. The velocity of the data is a constant problem.

C. Value

The value of the data decays so the timeliness of data processing is essential. For instance, credit checking while banking is essential [10].

D. Data Discovery

Finding relevant and useful data from the vast collection of data that is available is a greatest challenge.

E. Data comprehensiveness

Extracting useful information that includes personal identification information of the customers without compromising their privacy is also a biggest challenge [6].

VII. CONCLUSION

We have discussed the meaning of big data along with its methods, tools, and challenges. The four V's that define big data are volume, velocity, variety, and veracity. Big data is one of the recent trends and major challenges in the current world. Extracting the useful information from vast data and identifying the relevant data from the vast data that is growing in today's world is also a major challenge. Apart from all the challenges, the properly processed useful analysis of data can help various industries to improve their business. A lot of organizations depend on big data analysis to improve their business.

VIII. FUTURE WORK

The data growth is expected to double every two years. The difficult task is to transform the data into knowledge. The data may involve uncertainty in many forms. A combination of two or more models will result in better results. The future work includes forming hybrid models for better analysis of data. To achieve faster processing, high performance, and throughput is also one of the future works. The immediate need is to express the data and access requirements of applications. The privacy of sensitive data will be the biggest challenge in the future. Demand for data scientists will be high and

January-February 2020
ISSN: 0193-4120 Page No. 14987 - 14991

Machine Learning will be the next big thing in big data. Investments in big data processing will be more.

REFERENCES

[1] Sasa baptistic, Paul Van Der Laken, "History, Evolution and Future of Big Data and Analytics: A Bibliometric Analysis of Its Relationship to Performance in Organizations", British Journal of Management, Vol. 30, 229–251 (2019), 2019.

[2] Jens Baum, Christoph Laroque, Benjamin Oeser, Anders Skoogh and Mukund Subramaniyan, "Applications of Big Data analytics and Related Technologies in Maintenance—Literature-Based Research", Machines 2018, 6, 54, 2018.

[3] Ms. Komal,"A Review Paper on Big Data Analytics Tools", International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES) , 2018.

[4] D. P. Acharjya, Kauser Ahmed P, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.

[5] Sabitha M.S, Dr.S.Vijayalakshmi, R.M.RathikaaSre , "Big Data – Literature Survey ", International Journal for Research in Applied Science & Engineering Technology (IJRASET) , 2015.

[6] Sruthika S, Dr.N.Tajunisha, "A study on Evolution of data analytics to big data analytics and its research scope", IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS'15 , 2015.

[7] Neha Midha and Dr. Vikram Singh, "A Survey on Classification Techniques in Data Mining", IJCSMS (International Journal of Computer Science & Management Studies) Vol. 16, Issue 01 Publishing Month: July 2015, 2015.

[8] Hari Kumar.R M.E (CSE), Dr.P.UmaMaheswariPh.d, "Literature survey on big data and preserving privacy for big data in cloud", International Journal of Technical Research and Applications e-ISSN: 2320-8163, 2014.

[9] Nada Elgendy and Ahmed Elragal, "Big Data Analytics: A Literature Review Paper", Conference Paper in Lecture Notes in Computer Science • August 2014, 2014.

[10] Hui Li, Zin Lu, "Challenges and Trendsof Big Data Analytics", 2014 Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 2014.