A Study on Big Data Analytics

Heena Girdher[1], Anshul Garg[2]
[1, 2]Dept of Computer Applications
[1, 2]Chandigarh Group of colleges, Landran (Mohali)

Abstract- *In the information era, data is increasing at exponential rate so it becomes difficult to capture, store, share and analyze such huge amount of data. Due to the rapid growth of data, traditional database management tools are unable to analyze large amount of data. This paper discusses methods and tools which can be applied to big data and applications of big data in various domains.*

Keywords- big data, analytics, HDFS, Map reduce, Hadoop, Cloud computing

I. INTRODUCTION

With the enhancement in technology over years, data has also been increasing at exponential rate. If we talk about 1970s or 80s, not many people were using computers so data fed into the computers was very less. But now a days, everyone is using smart appliances. These smart appliances are connected to each other via network generates lots of data. The main source of data is Social media. We people are human beings. We love to share our feelings, thoughts with others. Social media provide us the platform to share, generates very large amount of data. If we look at the statistics, Facebook users generate 34,722 likes every minute. 100 TBs of data is uploaded daily on Facebook [1]. Twitter generates 175 million tweets per day[2]. YouTube users upload 100 hours of new video per minute [3]. In 2011, 1.8 ZB of data were created as of that year, according to IDC [4]. In 2012, this value increased to 2.8 ZB. By 2020, enterprise data is expected to total 40 ZB, as per IDC [5]. So we are dealing with lots of data which termed as big data. Thus efficient tools and techniques are required to manage big data.

Big data Big data refers to large and complex datasets that cannot be processed and analyzed using traditional relational database tools and techniques. There are five main features that characterize big data i.e. volume, variety, velocity, veracity and value.

Volume: -refers to size of data. Data is in petabytes rather than in terabytes.

Variety: -means data is generated from heterogeneous sources. Data now includes structured as well as unstructured data like emails, text, audio, video and images etc. and semi-structured data like XML.

Velocity: -means the speed at which data is generated.

Veracity: -refers to quality of data. Data may contain missing values, inconsistencies, noise etc.

Value: -Finding correct meaning out of the data (identifying valuable data). It means we have to identify valuable data from large amount of data [6].

II. BIG DATA CHALLENGES

1. **Storage: -** Data is large in size [in petabytes]. To store such a large amount of data is a great challenge. The disk space of the computer infrastructure should be able enough to handle such large amount of data.
2. **Transfer: -**To transfer petabytes of data, large bandwidth of the communication channel is required.
3. **Processing: -**As data is increasing at a rapid rate, algorithm should be scalable to handle massive amount of data.
4. **Data Quality: -** Data may contain noise, incomplete, and inconsistent values. To find the right data from enormous amount of data is a big challenge [7].

III. BIG DATA ANALYTICS

New database techniques introduced to address the challenges of big data. This section will put a highlight on the technologies that make big data analytics possible.

Cloud Computing:- Cloud computing is the sharing of computing resources such as storage, servers, databases, networking and softwares over the internet. Cloud computing eliminates the expense of setting up IT infrastructure on a local system. In cloud computing you only have to pay for what you need, upgrades are automatic and scaling up or down is easy. Cloud computing provides three types of services:-

- Iaas (Infrastructure as a service)
- Paas (Platform as a service)
- Saas (Software as a service)

Iaas (Infrastructure as a service):- With Iaas, you can take IT infrastructure- servers, VM’s, storage, networks, Operating system on rent by paying for what you need.

Paas (Platform as a service):- It provides on demand platform for developing, testing, delivering and managing software applications.

Saas(Software as a service):- It is a method for delivering software applications over the internet, on demand and typically on a subscription basis. In such case, cloud providers manage software upgrades and security patching. Example- Gmail is the software and Google is the service provider. [8].

Hadoop: -Apache Hadoop is a framework that allow us to store and process large datasets in parallel and distributed fashion. Hadoop consists of two core components: -HDFS and MapReduce [9].

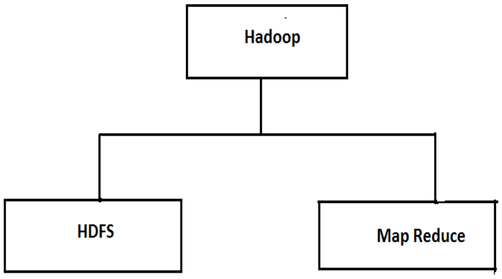

Fig 1.1: Hadoop Components

HDFS HDFS is a distributed file system in Hadoop for big data storage. HDFS consists of two main components: Name node and Data node. Name node is the master node that maintains all the nodes in the cluster. The actual data is stored in blocks on Data nodes. Name node record the metadata of the blocks in the cluster e.g. location of the blocks stored, size of the files. Name node receives the request from the client and checks its metadata to find out which is suitable data node for storing the data related to the client [10].

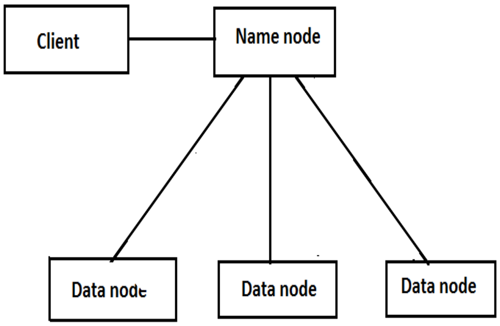

Fig 1.2: HDFS Components

Map Reduce is a programming framework that allows us to perform distributed and parallel processing on large datasets in a distributed environment. MapReduce consists of two phases- Map and Reduce. Mapper maps input key/value pairs to a set of intermediate key/value pairs. Map function partition large computational task into smaller tasks and assigns them to the appropriate key/value pair [11]. The output from each mapper would be input for Reducer. Reducer then performs the computation by combining all values which share the same key value [12].

IV. APPLICATIONS OF BIG DATA

1) **Healthcare** Big Data improves the quality of healthcare. Big data provides the detailed information of patients to doctors as opposed to the information gathered by one time visit to the hospital. Diseases can be predicted earlier before spreading based on this information. Through the use of wearable devices, patients may be alerted about their health conditions time to time. The quality of hospitals can be improved by examining the shortage of doctors and nurses on timely basis. Personal treatment can be provided to patients by monitoring the effect of medication continuously and drug dosage can be minimized to avoid side effect [13].
2) **Employment** Big data improves the employees' selection process by unbiased employment decisions. It help employers to take better decisions in hiring, performance evaluations and promotions.
3) **Infrastructure Maintenance** Buildings, roads, pipes can be maintained based on analyzing big data [15].
4) **Transportation** Bus services can be improved by measuring the time between buses at each stop together with the number of passengers waiting. More buses can be provided depending upon the number of passengers.
5) To improve water or power supply.

6) It helps to provide pension to senior citizens without any hinderance [15].

V. CONCLUSION

In this paper, we have discussed the innovative topic of big data which has gained lot of popularity due to its applications in various domains. Big data is increasing at exponential rate and decision makers have to handle such huge amount of data every year. We discussed in this paper several insights about big data, as well as its characteristics, importance and the core challenges for the future. Moreover, some of the big data analytics tools and methods were examined.

We believe that big data is becoming an important asset for decision makers. If this data is analyzed properly, it can reveal valuable insights to decision makers.

REFERENCES

[1] Facebook, Facebook Statistics, 2014, http://www.statisticbrain.com/facebook-statistics/.

[2] Twitter, “Twitter statistics,” 2014, http://www.statisticbrain.com/twitter-statistics/.

[3] YouTube, “YouTube statistics,” 2014, http://www.youtube.com/yt/press/statistics.html.

[4] IDC, “Analyze the futere,” 2014, http://www.idc.com/.

[5] S. SagirogluandD. Sinanc, “Big data: a review,” in Proceedings ofthe International Conference on Collaboration Technologies andSystems (CTS ’13), pp. 42–47, IEEE, San Diego, Calif, USA,May2013.

[6] Nada Elgendy and Ahmad Elragal, “Big Data Analytics: A Literature Review Paper”, September 2014.

[7] Ricardo Baeza- Yates, “Big Data or Right Data?”, vol-1087

[8] VinayakBorkar, Michael J. Carey, Chen Li, “Inside “Big Data Management”: Ogres, Onions, or Parfaits?”, EDBT/ICDT 2012 Joint Conference Berlin, Germany, 2012.

[9] EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)

[10] Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)

[11] Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011)

[12] EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)

[13] Yanglin Ren, Monitoring patients via a secure and mobile healthcare system, IEEE Symposium on wireless communication,2011

[14] Bill Hamilton, Big Data Is the Future of Healthcare, Cognizant white paper, 2010.

[15] Luis M. A. Bettencourt, “ The uses of Big Data in Cities”, 2013