Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 6, June 2021:138-145

Research Article

Research Paper On Big data and Hadoop

Abhinav Krishna Srivastava[1], Rakshit Sharma[2], Prince[3]

Abstract

In Today's era the word Big Data has a very useful meaning in every aspects. The meaning of Big data is to use and analysis of the data that is big in size. Now-a-days Big Data[1] is becoming very popular and there is so many research and analysis going on in this field. Researcher uses the big data to analysis the practical data and to find out its use and application based on the analysis.

Data analysis of the given data is done through various modes that is known as an unstructured data as we can easily find out now a days on the Google in the form of posts that has to be taken from social media sites[2] and images. Our research paper will give a an idea about Big Data and its application in daily life. We have also includes the challenges that has been faced by every data science engineer while analysis of the data. Our paper also talks about the Hadoop and gives a basic understanding of Hadoop.

Keywords: *Map reduce, Hadoop, Data Mining, Big data, HDFS*

[1] Student of Computer Science, Galgotias University, Greater Noida, India, abhinavkrishnasrivastava@gmail.com
[2] Student of Computer Science, Galgotias University, Greater Noida, India, rakshit.sharma49@gmail.com
[3] Student of Computer Science, Galgotias University, Greater Noida, India, princechaudhary5678@gmail.com

Introduction

Big Data is a very popular topic now-a-days, but everyone is talking about the different definition of Big Data. There is no one such definition which everyone has to follow. Data that is very big in size comes from different sources to analysis and compute which is only possible through Big Data.

Big data is very important that is not a structured or semi structured but is is only maintained by computation. Data analysis[3] of the given data is done through various modes that is known as an unstructured data as we can easily find out now a days on the Google in the form of posts that has to be taken from social media sites and images.

With the increasing demand in the technology and the population we have to store and analysis of data very fast. But it is not possible by data management[3][4], so we have introduces Big data that can easily compute and retrieve the data easily. Let me explain you with the help of an example Today everyone have android mobile phone and to run it efficiently we need RAM or ROM. Now our data is stored in ROM that is known as secondary memory and the size of the data is very large like 50GB or 120GB which is not possible to analysis by the traditional methods. In view to overcome this problem we have introduces Big Data concepts that will easily helps the user to compute and analysis the data easily.

There are mainly four types of characteristics of Big Data that is explained below-

- ***Volume*—**It defined the amount of data that is generated every second by social media or on the web page. Today data generation is increasing day by day. On facebook we can seen the analysis of volume as 4B times a day like button is hitted daily.
- ***Velocity*—**It means the speed of the data on which the given data is analysed of compute[5]. It is the most important characteristic of Big data. As no one is waiting for the data for a long everyone wish that their work is done as soon as possible.
- ***Variety*--** It is also important part of Big Data as it gives the information about the type of data that we are going to analysis and compute weather the data is of text, video, or audio. On facebook we daily see that there are about 150 Million people active everytime and about 100 Million posts were uploaded.
- ***Veracity*—**It deals with the data originality. As the data is coming from different sources so it may be deleted or modified somewhere. So veracity helps the user to find the accuracy of the data.

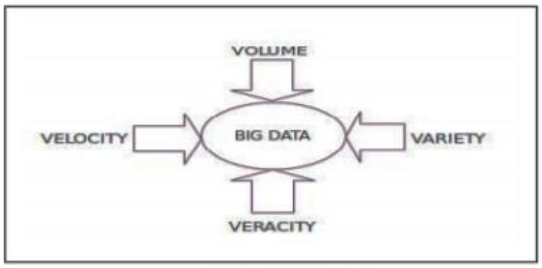

Fig. 1. 4 V's of BIG DATA.

Challenges And Opportunıtıes Introductıon

There will be so many websites and data that will provide you the information of Big Data challenges and opportunities. As the cloud becomes very popular so Big Data will also become very popular after cloud. There are so many opportunities in the field of Big Data as we can see Big data around us as earth, Hospital ,Finance and business[6], as the computation of the data of these field is very difficult so there are some tools of big data that will helps to easily analysis and compute the data that is of big volume. We think to model some big data challenges and opportunities and work on the data that we have got from different sources.

A. *Challenges In Big Data--Selecting a Temp*

- **Incompleteness and Heterogeneity—** As we see that if the user who is working on the Big Data wants to analysis of the given data set then it should be in the well format , but sometime it may not be it will be in unstructured form. Heterogeneity will be the most important and big challenge[7] that has to be faced in the data analysis. Let me explain you with an example consider a scenario of a hospital in which all types of patients were there. Now we have to make the record of patient medicine as well as his or her stay in the hospital. So it is very difficult to make the record of all patient accurately without any mistake. So incompleteness and heterogeneity is one of the challenge faced by Big data.
- **Timeliness—**There is also one other challenge that is known as Timeliness , which deals with speed of the data analysis. As it is very common that small data will take less amount of time to analyze while big data took more time to respond. But there are some conditions where we need to get the result or the output of the data as soon as possible. If there is some task happens of the transaction that has to be done without the permission of the user then it need to be cancel or stop before it was successful.
- **Privacy--** Now-a-days everyone is worried about their data privacy. It becomes a very big problem for everyone that their data is safe and secure. There are so many countries around the globe that makes and issued some laws enforcing that no data will be taken of anyone, without his permission. As we see social media will never leak the emotional news on twitter it will first warns the user to open that then only the post is available.
- **Scale—**By the name of the Big data we can understood that it deals with the large data set. Our system will gets slow while processing the big data so to overcome this problem be have make the powerful processor, but due to the large amount of data we can suffering with the same problem. The high rate of data is increasing day by day as cloud technology[8] is moving its data daily on the server.

B. Opportunities in Big Data—

Big Data gives various opportunities in every field like health, finance, corporate, and government. It deals with the large data and helps the user to meet their needs by providing them the profit at higher rates.

- **Technology--** As the technology increases day by day, the demand of Big data is also increasing day by day. Daily on Whatspp, facebook, twitter we see many posts these all will be handled by the big data. Facebook daily hitted the like button on the post of the

user by 500M times. There are so many opportunities in the field of technology in Big data.

- **Media—**Media also plays a very vital role in the field of big data by taking down the interest of the user by promoting their products over the website and web pages. Example for the same is when there are so many tweets on the twitter trending for the particular hastags then it will be easier of the data analyst to analyze the data of the tweet easily.
- **Government—**With the increasing demand of the big data government were also get benefit from big data. Now Big data is applicable in every sector of the society which election during the session 2014 in BJP government took benefit from big data.
- **Healthcare--** As we see in the hospital most of the data is unstructured. Healthcare department taking full benefits from the big data to analyze the report of every patient and maintain their record properly. Many technology has been introduced now to overcome the problem of data in the hospital.
- **Science and Research—**Now-a-days everyone is talking about the research in the field of Big data. So many data analyst take Big data as their research topic. Every month so many researcher publish their review paper on big data.

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

Hadoop Framework In Bıg Data

The process of big data that will be used as an open source in various field id known as Hadoop[6]. Every data analyst and the researcher used the hadoop software to analyze and compute the data easily. The main important factor that influence the structure and functioning of the hadoop is Map reduce, second one is Google's architecture and finally file system of the Google. It helps the analyst to easily compute the data on computing that is distributed over the city or country. There are some other parts of hadoop that are HFDS[9], kernel, base and Apache hive.

There are mainly two components of Hadoop in big data listed below—

- Storage(Hadoop file system).
- Processing(Map Reduce).

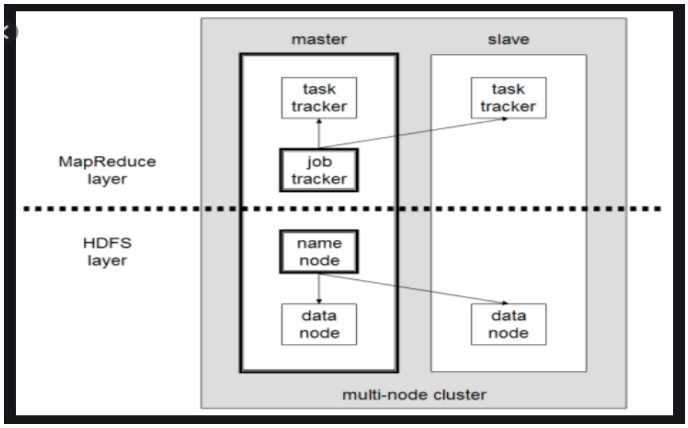

Fig 2. Hadoop architecture

HDFS Architecture—

HDFS which is also known as Hadoop Distributed File System contains a system of tolerant storage to store the data. The main benefit of using HDFS is to store large amount of data in the storage and if something happens to the system then we can easily retreive our data without any loss. Hadoop contains so many types of machines in it system software that will helps to work with all the machine in a simple way. Computers that are not expensive will helps to build up the clusters[10].

Now if any of the clusters fails to operate or work properly then Hadoop never let their user to loss the data it continues with the next available cluster. HDFS architecture also helps the user to break their work into one or more part, that will known as blocks to work effectively, now it will helps to store these blocks on different servers. Finally it helps to copy the same file in three different servers so that there will be no loss of the data of the user.

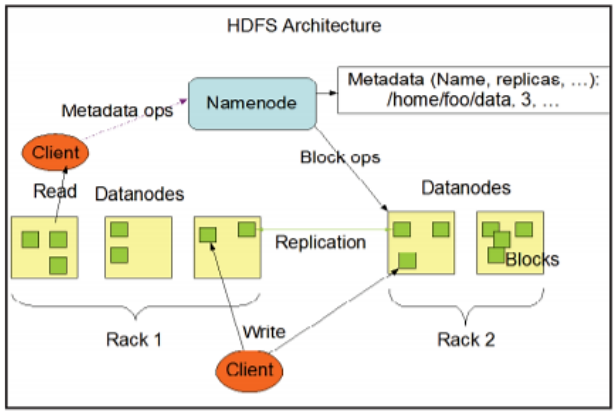

Fig. 3: HDFS Architecture

Map Reduce Architecture—

Map reduce is known for the processing part of the Hadoop system. Map reduce[12] helps the user for various operation to be performed on the large data of the user that has been taken from various sources. It helps to break the data and started to run the data in parallel. If we are an analyst the we will see as by many dimensions.

Let us give you an example if there is a dataset of some large organisation in the society then it will be very difficult to process that data at the same time. So, map reduce will break those

dataset into small parts to run them parallel and equally at the same speed or velocity. This will be very beneficial in the field of java while using hadoop. The programming languages which are Pig and Hive will be used for this purpose. The result that the user will get from these map reduce data processing will stored in HDFS architecture[11].

Map reduce basically perform and focus on two parts--

• map – means mapping the keys and values in the dataset.

• reduce – merging of all keys values that will link with intermediate values.

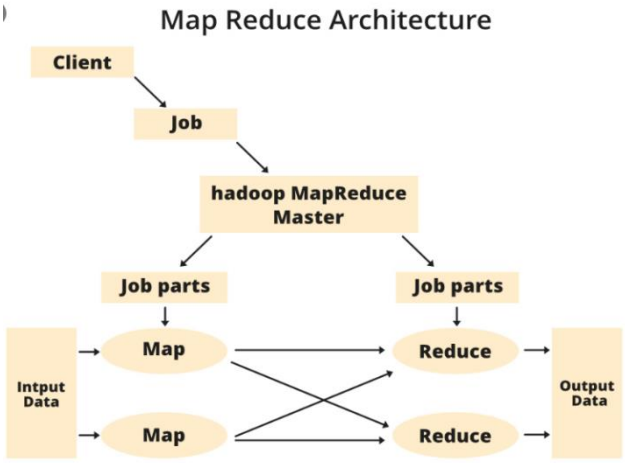

Fig 4. Map Reduce Architecture

Applıcatıon In Data Mınıng

Big data also plays a very important role in Finance field to make up the business more profitable in any organisation. It will also help the user and the researcher to research on its algorithm and pattern to discover new models. They also find out the patterns that is followed in the data set. Data mining[14] is mostly related to data mining because in data mining we have to extract the information that is useful to the user which has to be done only by the data analyst. Data analysis of the given data is done through various modes that is known as an unstructured[13] data as we can easily find out now a days on the Google in the form of posts that has to be taken from social media sites and images.

Big Data use so many of data mining theorems some are listed below--

- **Classification Analysis:** By classification analysis user can easily get the information about the data. This can also be done by clustering the data into group.
- **Cluster Analysis:** Basically cluster means same group of data. So, cluster analysis helps the user to identify the data that is similar to the other data of same type in the cluster. We can see this with an example that customer buying products on e-commerce websites like Amazon, Flipkart then there will be some recommendation given to the customer based on their previous purchase.

- **Evolution Analysis:** Evolution analysis is known for mining the data from the DNA sequences. This data mining technique is mostly used in banking sector to identify and analyze the data of the user on stock exchange of the previous year.
- **Outlier Analysis:** In some fields of data mining there is no use of making cluster and pattern for identifying the data set by the analyst.

Conclusıon

Now-a-days Big Data plays an important role in every field. In this research paper, we have discuss the brief about Big Data and Hadoop[15]. We have also discussed about the four characteristics of the BIG DATA. In every organisation and sector their will a large dataset of the working employee and company previous year records. So, to analyse the data easily technology have introduces the Big data concepts so the user and researcher can easily compute with the data without any loss to the data. The methods used for computing is not so costly and easily available. Data that is very big in size comes from different sources to analysis and compute which is only possible through Big Data.

References

S.Vikram Phaneendra, E.Madhusudhan Reddy,"Big Datasolutions for RDBMS problems- A survey", In 12thIEEE/ IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

Hewlett-Packard Development Company. (2012). Big Security for Big Data. L.P.: Hewlett-Packard Development Company

Aveksa Inc. (2013). Ensuring "Big Data" Security with Identity and Access Management. Waltham, MA: Aveksa.

Xu, X. (2012). From cloud computing to cloud manufacturing. Robotics and computer-integrated manufacturing, 28(1), 75-86.

Kaisler, S., Armour, F., Espinosa, J. A., Money, W. (2013). Big Data: Issues and Challenges Moving Forward. InternatioConfrence on System Sciences (pp. 995-1004). Hawaii: IEEE Computer Soceity

Katal, A., Wazid, M., Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. IEEE, 404-409.

Amit Verma et al.,"Care 2000: Analysis and Re- Engineering ", National Conference on Emerging Trends in Communication held at SVIET-BANUR(District Patiala, PUNJAB) pp. 73, on 20th -21st February, 2009.

Amit Verma et al.,"Cross Layer Feedback Design: Optimization For Energy Efficient Mobile Devices Protocols Stacks Over Wireless Sensor Networks", National Conference on Emerging Trends in Communication held at SVIETBANUR(District Patiala, PUNJAB) pp. 90, on 20th -21st February, 2009.

Marr, B. (2013, November 13). The Awesome Ways Big Data is used Today to Change Our World. Retrieved November 14, 2013, from LinkedIn: https://www.linkedin.com/today /post/ article/20131113065157-64875646-the-awesome-ways-bigdata-is-used-today-tochange-our-world

Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" in International Journal of Advance Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014

Mucherino A. Petraq papajorgji P.M.Paradalos 1998. A survey of data mining techniques alied to agriculture CRPIT.3(3): 555560.

SMITHA T, V. Suresh Kumar "Application of Big Data in Data Mining" in International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 7, July 2013).

Smitha.T, Dr.V.Sundaram, "Classification Rules by Decision Tree for disease prediction" International journal for computer Application, (IJCA) vol 43, 8, No-8, April 2012 edition. ISSN0975- 8887; pp- 35-37

Divyakant Agrawal, Challenges and Opportunities with Big Data, A community white paper developed by leading researchers across the United States.

Puneet Singh Duggal, Sanchita Paul, Big Data Analysis: Challenges and Solutions in International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.