WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Big Data: Challenges and Future

Surbhi Bahukhandi[1], Sahil khera[2]

[1, 2]Uttaranchal University, Dehradun

ABSTRACT

Data has become the requisite part of each and almost every industry, whether it is economy or is an organisation or for an individual. Every hour new data is generated over exabytes from digital technologies and information systems. For analysing such type of huge data, a lot of endeavour at different levels is needed. That is where Big Data comes in picture. The purpose of this paper is to highlight Big data problems and challenges, additionally the future of Big Data. Furthermore, It includes how to act with Big Data by using tools like Hadoop.

Key Words: Big Data, Hadoop, MongoDB, Challenges

Introduction

There has been a huge requirement to store and procedure enormous amount of data nowadays and will be in upcoming years also. Big data is accountable for storing and dealing with wide range of data sets. It provides opportunities to ameliorate research and helps in settlement support appeal like business and machinating also. Today's wide-reaching organization, CERN, bring about "200 PB of figures per year in large hadron collider project". The quota brought on statistics on the internet is about 2.5e+18 bytes per day.There are about 3.6 million google search that is performed each one minute every day. There are millions of data being shared on Instagram, face book or on any other social media platform every single day. When dataset evolves into massive data that its warehouse and clarification become a challenge, then that dataset is considered as big data. Cloud computing provides canonical support to abolish the challenges of storing huge amount of data resources. It provides data storage volume and synchrony. Both big data, and cloud computing are interrelated to each other. Big data is considered an issue, and cloud computing is a solution for it while conventional storehouse cannot fulfil the approach of big data.

Big data can be explained as the amount of binary data generated from distinct source of technology. Technologies can be detectors, OCR devices, cell phones, broadcasts, electronic mail, or any social networking sites like Facebook, Tumblr, LinkedIn, and Snapchat. Data categories in them include photos, recordings, messages, voice messages and union of each of them. To analyse the data and for recognition of design, it is important to accumulate and manage the data firmly on cloud. Big data can be represented as an advancement of person cognitive procedure. If the data cannot be hold by the traditional databases, then we can classify it as a big data.

The big data can be classified into some forms counting:

• ***Structured data:*** the term structured data normally indicates to the data with a defined length and configuration. Structured data instance include digits times, and sets of characters and digits called strings.

• ***Semi-structured data:*** the term semi- structured data are a formation of structured type data which don't obey the conventional form of data models correlated with the relational database or any different form of data table, however accommodate tags or any different marker for the separation of semantic components.

• ***Unstructured data*** : the term unstructured data is information which cannot be set out as stating to a PRE- set data schema, that's why it cannot be set aside in RDBMS. Many professional papers are considered unstructured.

The idea of big data is a crucial alternation in academic and firms. The model is observed as an attempt to acknowledge the big datasets that come up with an encapsulated statistics over an enormous data load. Additionally, it is observed by the researchers as a mean to reserve enormous dataset. This concept is a big fuss and will be trendy in the future.

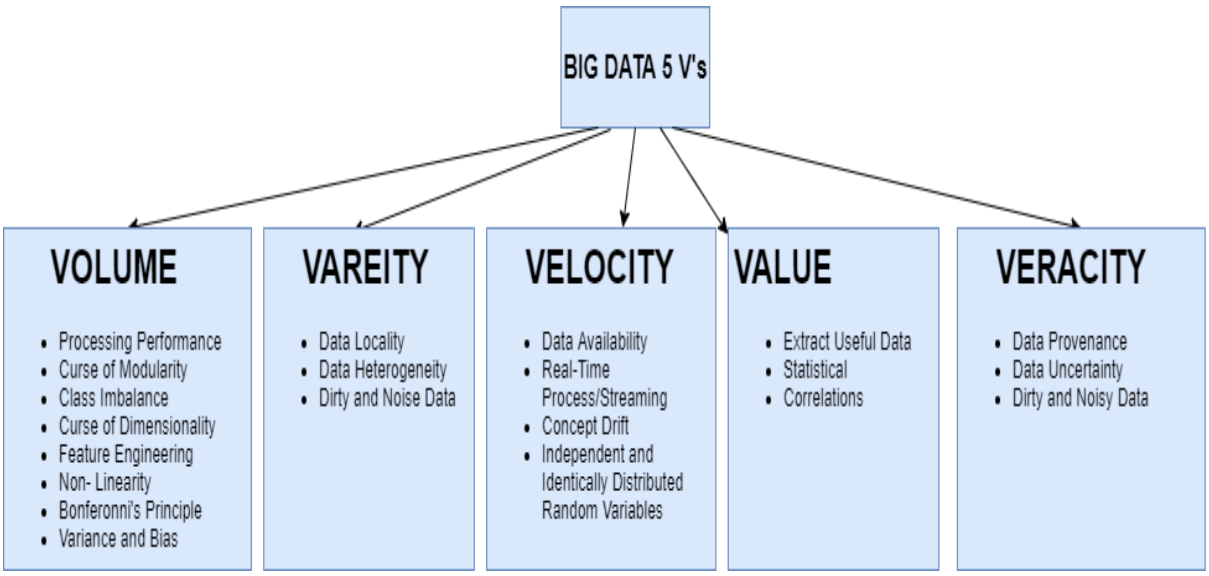
Attributes of big data

In Fig.1 the five distinct features of big data, entitled V's 5, are given:

• ***Volume:*** It outlines the dimensions of dataset which is managed by big data structure. It illustrates the data quantity produced from various origins that manifest the enormous amount of data in numbers.

• ***Velocity:*** It covers the distinct rates through which data streams enter and leave the system.it shows the data constancy speed from various origins example Twitter and Instagram.
• ***Value:*** it outlines the data's accurate value. It shows the big data value (significance of data following analysis) which is because of the certainty that data alone is valueless. The value is and end phase which approaches after the clarification of volume, variety, contrast, and other phases.
• ***Veracity:*** It mentions the correctness of data and labelling the concealment, probity, and proximity of data. The standard of data can differ much which can act on the correctness of analysis.
• ***Variety:*** it constitutes data types and expanding amount of cyberspace clients far and wide. It refers to distinct forms of data from many sources that deal with the framework of big data.

FIGURE 1: 5 V'S [4],[5]



Big data analytics tools

- **HADOOP**

It's a free software that depends upon framework of java. With the help of it we can store and process big data. The data is kept in a low-priced servers that passes as clusters. It controls numerous functions and never let the user think about the hardware collapse. As the refining ability of application servers has been growing rapidly, the databases have fall behind on account of their finite capacity and momentum. although, now a days, as numerous applications are giving rise to big data to be sorted, Hadoop takes part in a remarkable part in providing a necessary makeover to the database sphere.

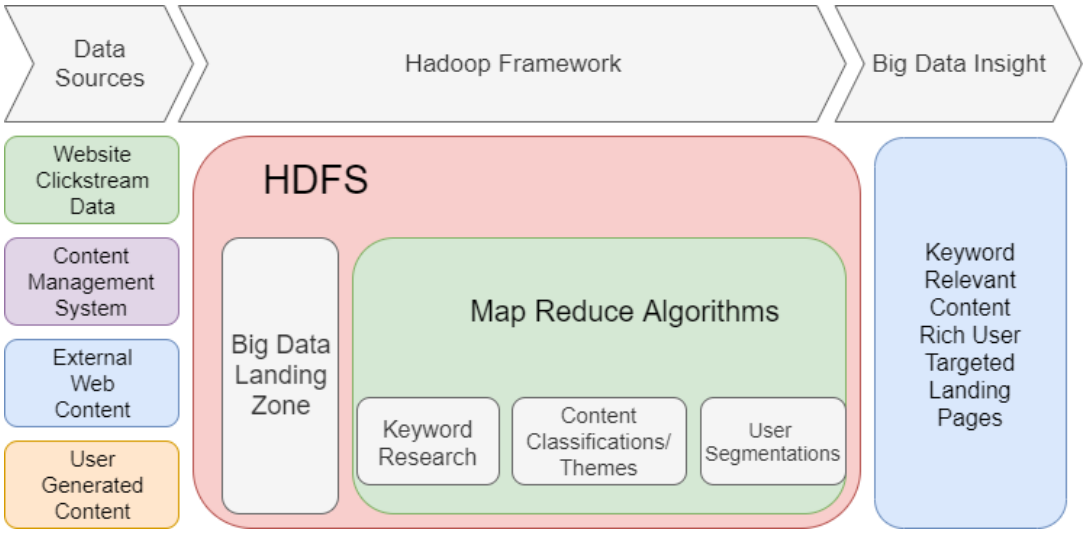


FIGURE 2: ARCHTECURE OF HADOOP [6]

- **MongoDB**

MongoDB is a current substitute to database. It's the finest Big Data Analytics tool for functioning on data sets that change regularly. MongoDB is used in storing data from mobile applications ,content administration ND production index. SIMILAR TO Hadoop, THE USER can't ACTIVATE MongoDB immediately. it is a NoSQL database management platform that can put distinct data prototype, and stores them in key-value sets. It was evolved for functioning with huge size of distributed database .

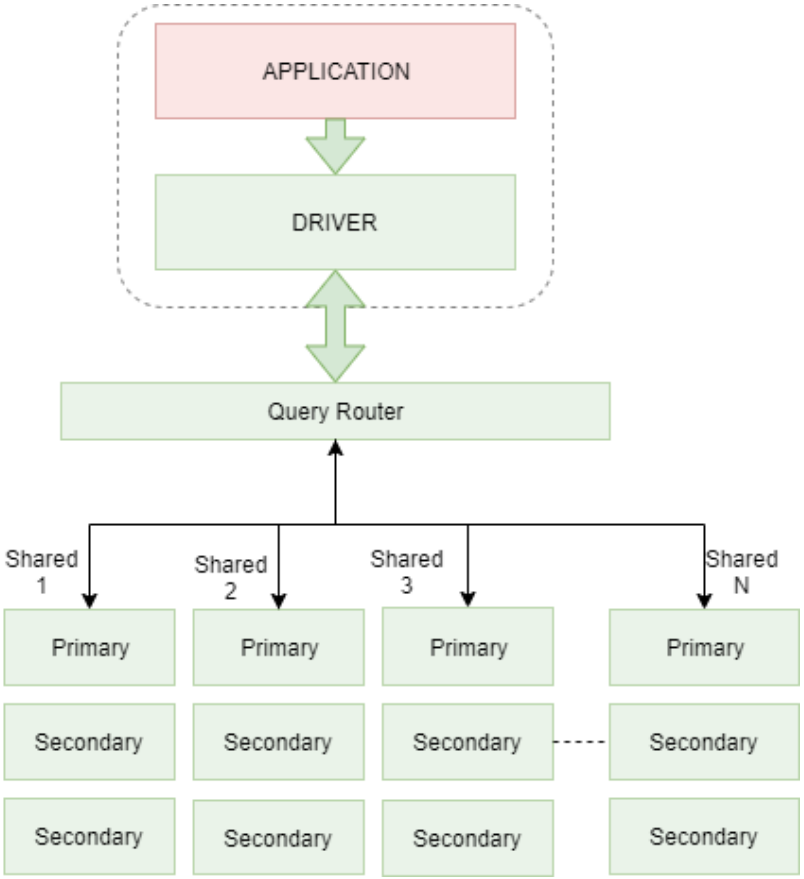


FIGURE 3: ARCHITECTURE OF MONGODB [7]

- **MAPREDUCE**

The processing of huge datasets on cluster with aligned as well as distributed algorithms is possible by MapReduce. It is applied to put down applications which can process extensive amounts of data parallelly and very quickly on huge clusters. It is basically having two functions, namedmap() and second one is reduce(). Filtering as well as classifying is handled by mapper function and encapsulating the tasks is done by reducer function.We can process the input data by defining these functions and using our own logic.MapReduce ease gruelling and strenuous tasks to process an enormous amount of data that is already processing distributed environment with the help of "parallel I//O scheduling".[issues paper].According to [10], execution of iterative algorithms is the major drawback of MapReduce as the data is passes to the succeeding iteration after being written on the disk.

CHALLENGES

Big data analysis is a next level and advanced analysis process and viewing techniques on large data sets to reveal hidden patterns and anonymous relationships in order to make better and more effective decisions. Discrepancy, inadequacy, intricacy, extent, and privacy of data are some of the challenges of big data mining.

- **Inadequacy of data**

With incomplete data it is quite unpredictable to do the data analysis and during that time only, it should be manipulated to give a proper and appropriate result. Performing the action of making incomplete data into complete data is a big challenge. Data generally go incomplete or get missed when the source from where we are taking or collecting the data is broken or is crashed due to malfunction of sensor data. To handle this, data mining algorithms has been made which makes forecast on the missing values by building model from the observed data. Those model gives the appropriate forecasting from observed data by grabbing the pattern and then filling the data with highest probability

of happening or being there.

- **Intricacyof data**

New day new data are created every second which clearly tells how rapid the data is increasing. And with more data, it also becomes a challenge to manage it. As the data could be unstructured along with huge size, older software are not designed that way to handle data in such situations. And, when it comes to retrieving such huge data, making a model out of it also a challenge.

- **Security and privacy**

Big data sizes are increasing every day constantly. According to [8], data sharing methods have been created like Dwork, formulated on "69 sound privacy definition". The cause of huge cloud computing is on account of accessing wide network. The techniques for security are being used to protect big data on cloud [8].

Thus, the privacy and security of data is a huge apprehension from the perspective of Big Data. Hence, to protect data, and privacy, generally, two approaches are followed. Either restrict access by certification and access control, so the information which is sensitive is accessible by limited users only. Confidential data cannot be resolved by single record if data fields are not being specified. It can be achieved by injecting randomness into the data.[3]

FUTURE SCOPE OF BIG DATA

In the latter case, different challenges, and related approaches exist where overhead retention can be reduced and the cost of disk reading and writing can be reduced. These challenges in processing systems have emerged in the picture when they provide error detection. Quick and accurate error detection before causing damage or operating delay and system detection are in a normal state within a limited time period due to minor malfunction and high utility usage.According to [9], The connection of data warehousing and cloud is require because of its ease of use and low-budget.
The given methods will be listed below to address the above-mentioned challenges:
• Integration of error reporting and repetition methods is required in retention systems whenever error tolerance occurs. This combo will predict errors and as soon as errors are detected, it will start duplicating hard to create a small number of copies at minimal cost and data delivery.
• An integrated approach is also required that includes deleting codes and duplicating storage systems. The erase code saves very well compared to duplication and saves network traffic and disk I / O, so the integrated approach to both will provide the best solution for future error tolerance in future research.
• A flexible hearbeat-based system that measures operating load, processing time and setting the shut-off value of the fly is also required to improve the detection of permanent errors. However, the method of detecting a heartbeat is limited when it detects temporary errors. Indirect performance can be modified to monitor the continuity of processes and to quickly terminate slow-moving tasks / processes, so that the app is free from widespread use.
• A wide range of new diagnostic and recovery methods may also overcome the tolerance of errors. The location of the data fragments and backup copies of the processing systems. Then, instead of relying on the master node cluster for handling error detection, it will be redirected to each location that will handle data no longer acting as partner nodes and set up peer-to-peer communication with another node that will use processing functions. in the first piece of data from time to time, we will exchange hearts. And whenever a malfunction or node is detected, all processes will be managed by a partner node. Using this method, all error detection will be distributed between nodes instead of being managed only by the master node. The total delay from the acquisition phase to the recovery will also be improved.

CONCLUSION

The augmentation of data is rapid on account of expansion of social network platforms, stocks, media news and so many. Big Data is a very hot topic of research not only for scientific data but for business data as well. It has become requisite for mechanized and pre-programmed uncovering of intelligence that is the part of the patterns that occurs periodically. Big Data not only help the companies with superiorresolution, but also assists them to predict or speculate the substitute and then to create new chances. In this paper we consider about the issues, challenges and future correlated to big data mining and the analysis tools such as Hadoop which could help organizations to know their customers and marketplace more effectively and of higher quality. It assists in gaining and acquiring convenient knowledge out of those big data using the tools for their work.

REFERENCES

[1] Venkatesh H, Shrivatsa D Perur, Nivedita Jalihal. A Study on Use of Big Data in Cloud Computing Environment.

[2] Pedro Caldeira Neves, Bradley Schmerl, Jorge Bernardinoand Javier Cámara1. Big Data in Cloud Computing: features and issues.

[3] Nabeel Zanoon, Abdullah Al-Haj, Sufian M Khwaldeh. Cloud Computing and Big Data is there a Relation between the Two: A Study.

[4]https://www.sciencedirect.com/science/article/pii/S1877050915006973/pdf? md5=35c8b4ef1ebf6837fdaf529137e0ab24&pid=1-s2.0-S1877050915006973-main.pdf
[5]https://www.researchgate.net/figure/Big-Data-characteristics-with-associated-challenges_fig1_316448042

[6]https://www.semanticscholar.org/paper/A-Big-Data-Hadoop-Architecture-for-Online-Analysis-Naik/3272b7daec3854a1bf6efad66180d7de570560d1

[7]https://siliconangle.com/2013/08/21/10gen-boosts-hadoop-ecosystem-with-upgraded-connector/mongodb-architecture/

[8] Parth Goel, Radhika Patel, Dweepna Garg, Amit Ganatra. A Review on Big Data: Privacy and SecurityChallenges.

[9] ERUM MEHMOOD, AND TAYYABA ANEES. Challenges and Solutions for ProcessingReal-Time Big Data Stream: a systematicliterature review.

[10] TRUNG NGUYEN, RAYMOND G. GOSINE, AND PETER WARRIAN. A Systematic Review of Big Data Analyticsfor Oil and Gas Industry 4.0.