A Bird's Eye View on Big Data Analytics

R.Anandhi[#1], G.Sekar[*2]

[#] Assistant Professor, Department of Computer Applications
D.G. Vaishnav College, Arumbakkam, Chennai, TN, India
[1] anandhi78@yahoo.com
[*] Assistant Professor, Department of Computer Science
Government Arts College, Vyasarpadi, Chennai, TN, India
[2] gsekarg@yahoo.co.in

***Abstract*— Big data analytics is the process of examining big and variety of data sets to get rid of unknown relations, market trends, hidden patterns, useful information that can help the firms to take wise decisions on business matters. This will help organization to earn more revenue opportunities, set effective marketing trends, and make good customer services with more efficiency and performance. The big challenges lie before the analytics of big data are data scattered across various units in various formats, lack of a clear ideas for implementation, ineffective co-ordination of big teams, lack of support and sponsorship for carrying out research projects, dependency of legacy systems for processing data etc.**

Keywords-Big Data, MapReduce, Hadoop, HDFS.

I. INTRODUCTION

There is no option for human that he should live amidst flood of data. During the last 30 years, many data management policies and procedures like logical and physical data independence, logical and physical schema, queries, optimizing those queries to reduce cost and time were adapted and later tuned a lot. The main aim of all research conducted was to store and retrieve data in effective and efficient manner. It is also obvious that if we turn backwards, we have improved a lot in field of Information Technology like Internet, Smart phones, Android technology, YouTube, social media like twitter, Facebook, Whatsapp, Instagram, WeChat. In any sort of communication between two medium, the main parameter to be considered is the data they shared [1]. Since devices and methodologies are increased for communication, obviously the data transferred will also be increased. Now many questions lie before us:

- What is the nature of shared data?
- Where to store the data?
- How to process the data on request?

The concept which was evolved in trying out to give best answer is the Big Data. It earns the name as "Big Data" because the size of the flooded data is very very large (ie) immeasurable in nature. Path travelled by the data size is kilobytes, megabytes, gigabytes, terabytes, petabytes and now it is in exabytes and zettabytes. A study shows that data available in internet would exceed the entire brain capacity of the living organisms on earth and 90% of that data was created only in the last two years. It is very surprising to know that for a day, quintillions of data are created and they reside at unknown servers in unutilized and unstructured format without knowing the validity of such useful data [2]. As an example for a minute, there is 350 GB of data on Facebook, 2, 77,000 tweets in twitter, creation of 570 new websites, 2 million searches in Google, 72 hours of videos uploaded in YouTube, 100 million e-mails sent, etc.

A. DEFINITIONS

- **Manyika** defines Big Data is the amount of data beyond the ability of technology to store, mange and process efficiently.
- Big Data is a term which defines the hi-tech, high speed, high-volume, complex and multivariate data to capture, store, distribute, manage and analyze the information is the definition given by **TechAmerica Foundation** in 2014.
- **Gartner and Gursakal** coined the term Big data as high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.
- **Gantz and Reinsel** said that Big Data Technologies are new generation technologies and architectures which were designed to extract value from multivariate high volume data sets efficiently by providing high speed capturing, discovering and analyzing methodologies.
- The cluster of methods and technologies in which new forms are integrated to unfold hidden values in diverse, complex and high volume data sets is the definition given by **Hashem**.
- **Laney** described as Big Data is high volume, high velocity and high variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Hence Big Data is the concept of new and innovative technology used to get, store, analyze, design, capture, process, and maintain zettabytes of data with high speed. The aim is to optimize the big data in such a way to retrieve the hidden knowledge to end users. Usually it involves various data sets which are differing in their size, structure, growth. The technology used to process the big data should give the result in a timeline within which the data is useful and valid [3].

B. AREAS USED

Big Data is used enormously in various fields and not limited to [4]:

- Social Media
- Online Services
- Telecommunication Domains
- Health and Medical Industry
- Transport Sectors
- Banking and Financial Services
- Research and Education
- Defense and Security
- Automation Domains
- Public and Private Sectors

C. DEFINITION PARAMETERS

Big data is initially characterized and defined by 3V's called as Volume, Variety and Velocity. But now the 3V's model is Big Data has been extended to 5V's model (Fig. 1) by adding two more parameters called as Value and Veracity [5].

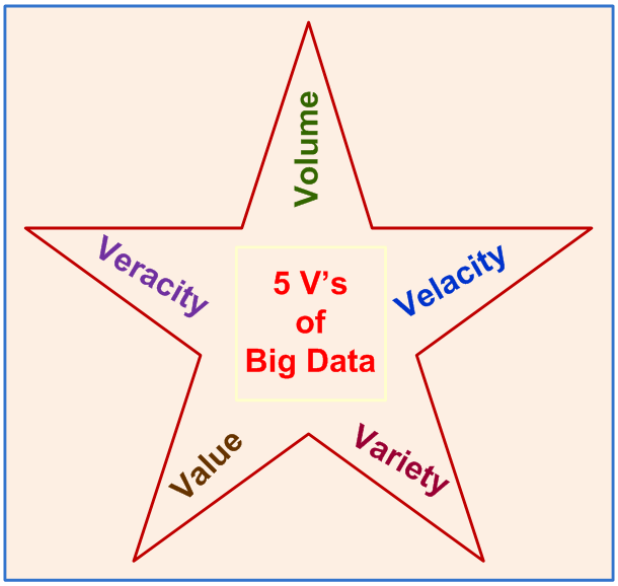

Fig. 1. 5 V's Model

- **Volume:** Volume indicates the amount of data. Data is growing day by day. It is inevitable as it is extracted from multiple devices. The data size also challenges the storage. The huge capacity of data is flooded from all connected sources involved in communication. Hence the data is measured in units of zettabytes. It has been predicted that by the year 2020, the existing data will be multiplied by 50 times.
- **Variety:** Variety indicates the heterogeneous nature of data. The data may be text, audio, video, log files, weblogs, legacy data and sensor data and they may also be in different formats. The data disposed by domains may be valuable, invaluable, structured, unstructured, semi-structured, probabilistic, multifactor, linked or dynamic in nature.
- **Velocity:** Velocity indicates the speed of data transfer, creation of files, access to data and delivery of data. If a transaction takes one minute to complete, it is considered to be too late in Big Data campus. Since the contents of data are constantly changing, velocity surely affects the performance of the system.
- **Value:** Value indicates valuable information extracted from datasets associated with Big Data. Among 5V's, value is the most important and critical parameter to be considered since each domain should be trusted one to produce useful and correct information to users.

- **Veracity:** Veracity indicates the integrity (ie) accuracy of data. Since lot of uncertainties like ambiguity, inconsistency, replication, tapping, spam and legacy issues may affect the data, veracity is hard to achieve. So Big Data should address trustworthiness, authenticity, reputation, availability and accountability of data.

II. PROCESSING STEPS

Big Data employs some existing techniques with improvement and tuning for its storage [6]. Big Data has to be passed through a pipeline for processing and tuning its validity and is referred as Big Data Ecosystem. There are many sources from which the data is first captured and recorded. Recording can be done by a series of filtering and compression. Since most of the data is not of interest, filtering must be done which should not get rid of important data. It should be noted that recording the information about the data (metadata) during its creation itself is not necessary since the pipeline will keep on changing in different ways.

Simply collecting and storing the information will not be helpful in any way as retrieval of data should be given more importance. Collected data is not in analysis format. But efficient retrieval obviously depends upon the efficient storage techniques used. More data so far collected will not be in a structured and uniform format. Here comes the concept of data cleaning-doing the process continuously with correctness will give the valid data.

Analysis of data does not simply include locating, identifying, understanding and citing data. The differences in data structure and semantics are to be resolved to make the problem understandable to computer algorithms. Data is analysed, designed and represented in the form visualized by human will be carried out by devising tools, revamping, implementation of techniques like cluster analysis, crowd sourcing, machine learning, text analytics and rules learning by a team of database experts.

Querying and mining Big Data is entirely different from existing relational database systems since the size of the sample is very large here. The nature of Big Data will be dynamic with more inter relationships and may be state and unauthorized data. So development of well-defined semantics and intelligible querying mechanism is done which are expected to produce correct interpretations of user request.

It is not enough to just provide results but also the additional information about the derivation of results to prove its correctness is necessary. A tight understanding of query languages is very much essential for correct interpretation of user requests.

III. TECHNIQUES

Since the datasets grow in size, experts found very hard to scale with traditional relational database management systems. The various data forms are [7]:

- **Structured**: Data available in the databases are structured in nature. They may be alphabetical, numerical, and alphanumeric. Structured data can be represented in schema. They are organized as entities and we can establish relationships among them. Here all entities will have same number of attributes.
- **Semi Structured**: The line which separates structured and semi structured data is blurred. Here entities may not have same attributes in the group.
- **Unstructured**: Here the data may be of any type without any format or sequence. Audio, images and video will fall under this category.

A. Map Reduce

MapReduce (Fig. 2) is a processing model developed by Google in Java for distributed computing [8]. Data aware caching (Dache), quick response time, parallel processing with flexible programming model makes this model popular while real time processing, shuffling of data are real challenges for MapReduce framework. As the name implies, this technique involves two functions: Map, which takes a set of inputs and produces another set of outputs. Here the master node will take the input and divides the big task into simpler ones and distributes to slave nodes (ie) the input data split into even-sized data blocks for load sharing. Each data block is given to a slave node and asked to perform the map function for the base problem. When all the map function is over, the runtime system will collect and submits the result back to the master node. The master node will map similar values and gives them to reduce function [9]. Reduce, which captures the output of map process as input, identifies the same key value and merges and announces the result. Hence MapReduce architecture works around key/value pairs. The framework will turn each input into a key/value pair and supplied to map function. The output will also be a key/value pairs which are sorted out by key. The reduce function will be applied for each common key once. MapReduce architecture employs task tracker to manage the map and reduce functions on a node in a cluster and job tracker to monitor job submissions and distribution of tasks among the nodes in a cluster. There will one job tracker in a cluster and more than one task tracker in a cluster [10].

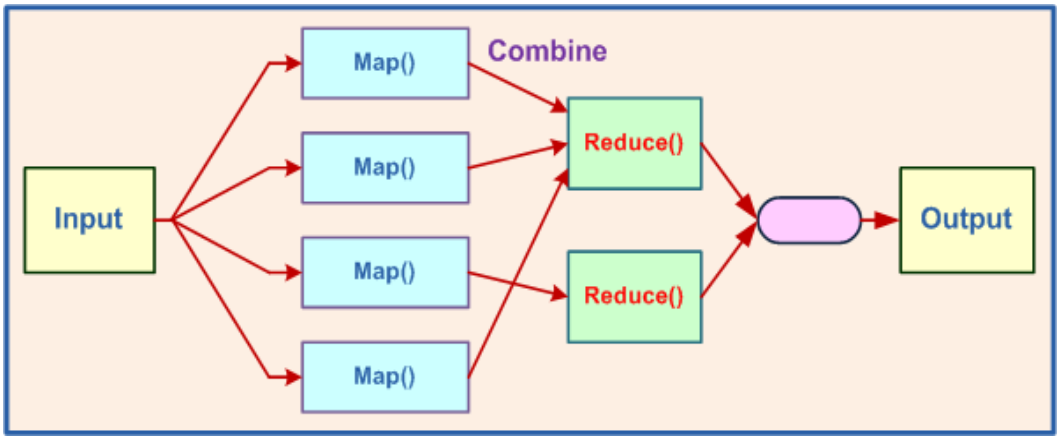

Fig. 2. MapReduce Architecture

B. Hadoop

Hadoop, product of Google's MapReduce is a programming framework used to process the biggest data sets in the distributed environment [11]. Apache Hadoop consists of Hadoop kernel, HDFS, MapReduce and related projects like hive, zookeeper, HBase (Fig. 3). Hadoop takes the unstructured or semi-structured request, process using MapReduce technique and locate the relevant information. Hadoop encloses a fault tolerant memory storage called as Hadoop Distributed File System, popularly referred as HDFS.

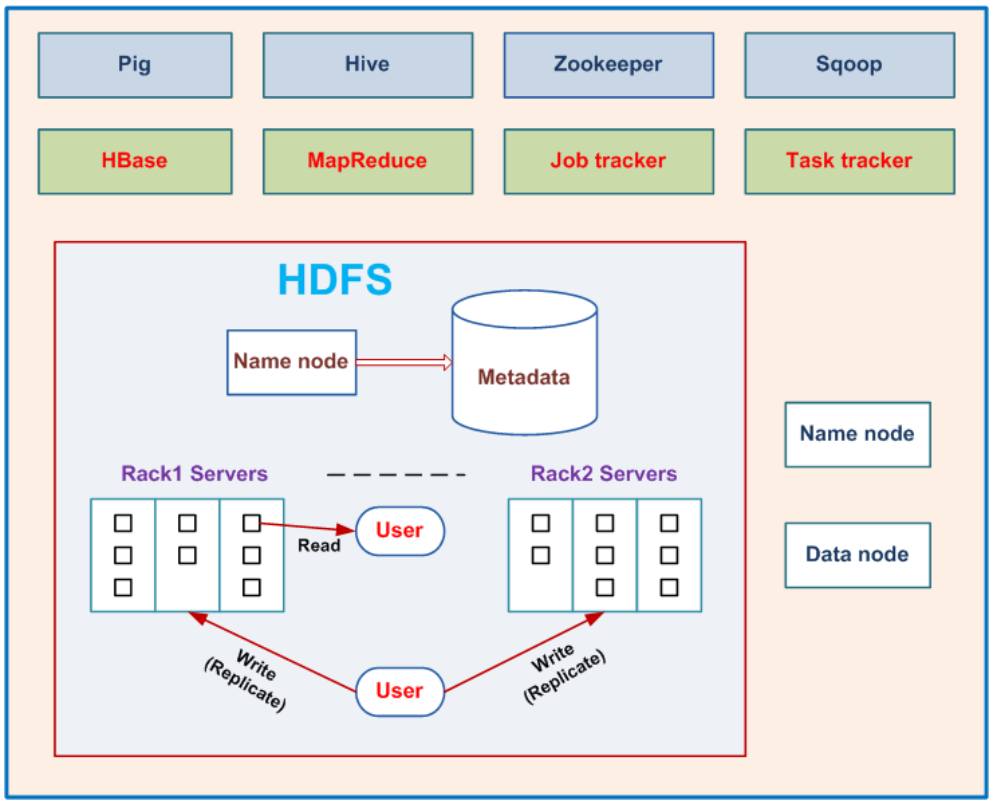

Fig. 3. Hadoop Architecture

HDFS has the capacity of scaling, tolerating failure without data loss and capturing huge information. HDFS is built by cluster of inexpensive computers. The data storage technique adopted by Hadoop is called as multiple clustered network attached storage (NAS) [12]. The nodes are categorized into Name node to manage metadata and control information, secondary name node and data node stores the blocks of HDFS. There will be one name node in HDFS with multiple data nodes. HDFS splits the incoming big file of information into pieces called as blocks and stores across the pool of servers [13]. The salient features of Hadoop are extensive storage, efficient computational capabilities, incrementally scalable, high throughput optimization, fault tolerant to failures by load sharing etc. But Hadoop suffers from single point of failure of master process. Other components of Hadoop to be mentioned are [14]:

- **Hive** is an application lies on the top of Hadoop used to provide conclusion about the carried out analytics. It is a decentralized agent used for applications by network local resources. Apache Hive data warehousing is a cloud based Hadoop hosting with a query language called as HiveQL that interprets SQL or similar queries into MapReduce forms. Hive is more secure with well tuned implementation but performs less than pig.
- **Pig** is high-level platform used to process and analyze data records written in high level language.
- **HBase** is a distributed, open source, non-relational (NOSQL) database system developed in Java that runs above HDFS. It is responsible for managing input and output of MapReduce architecture.
- **Zookeeper** gives distributed synchronization and centralization of services along with the maintenance of configuration records.
- **Chukwa** lies above HDFS and MapReduce framework used to collect and analyze massive amount of incoming logs.
- **Avro** provides functionality of data serialization and service of data exchange.
- **Sqoop** is a command-line interface application that provides platform which is used to converting data from Hadoop to relational databases and vice versa.
- **Oozie** queues the Hadoop jobs in a sequence and is a web application developed by servlets.

An improvement in Hadoop named HaLoop was developed by Yingi Bu which allows iterative applications for assembling Hadoop programs without major modifications and proves to have significiant efficiency in performance by iterative caching techniques monitored by loop aware scheduler.

IV. TECHNICAL CHALLENGES

There is several challenges available based on heterogeneity, scale, timeliness, privacy, human collaboration [15,16,17].

- Big Data is purely **heterogeneous** in nature in which mostly natural language dominate. But machine analysis algorithms require homogenous data. So structuring the Big Data is a tedious task to make the algorithms work. Even after performing necessary data cleaning, still some incompleteness and errors likely to exist due to this heterogeneity.
- Big Data will increase in **size** time by time. Managing the large volume of data was handled in past by increasing the resources like CPU power, memory as stated by Moore's Law. But it is not possible and wise way to adapt for Big Data also.
- **Speed** of data transfer is inversely proportional to the size of data. But users will be keen only in the response time of their query. Picking the necessary data from the ocean of data is practically impossible to be completed in seconds. So index based on constraints may be maintained to improve the performance. So velocity of data must also be increased (ie) Big Data speed should be make as directly proportional to Big Data size by employing efficient procedures.
- Maintaining the **privacy** of data (health records, defense records, and research) is also to be addressed as it may create sociological and technical problem in all levels. Strict laws governing the sensitive data should be enforced.
- Since algorithms alone cannot detect all patterns of user expectation, the system should also allow having experts to interact now and them along with their developed technique. Big Data systems should also allow humans to give their input and share their explorations.

V. CONCLUSION

As the data sharing increases, the requirement media for its storage is also to be increased. Thus now we are in the era of Big Data with many challenges and issues. So this paper explains the 5V's model of Big Data along with the pipeline technology to tune the Big Data. It also discuss the issues like heterogeneity, scale, lack of structure, error handling, privacy, timeliness, visualization and provenance at all pipeline stages of Big Data processing. This paper also describes the popular tool for Big Data Hadoop and its MapReduce technique for capturing Big Data. We must extend our great support for the research work carried out in tuning the storage and retrieval process of Big Data.

REFERENCES

[1] Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. Big Data Analytics. Big Data Technologies and Applications (pp. 13-52). Springer International Publishing, 2016.
[2] Rahm, E. Big Data Analytics. IT-Information Technology, 58(4), 155-156, 2016.
[3] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), 9,2016.
[4] Sonuga-Barke, E. J. Can Medication Effects Be Determined Using National Registry Data? A Cautionary Reflection on Risk of Bias in "Big Data" Analytics. Biological Psychiatry, 80(12), 893-895, 2016.
[5] Suthaharan, Shan. "Big data analytics." In Machine Learning Models and Algorithms for Big Data Classification, pp. 31-75. Springer US, 2016.
[6] Chandrasekaran, Balaji, and Ramadoss Balakrishnan. "Attribute Based Encryption Using Quadratic Residue for the Big Data in Cloud Environment." Proceedings of the International Conference on Informatics and Analytics. ACM, 2016.

[7] Jha, M. A., Dave, M., & Madan, S. Quantitative Analysis and Interpretation of Big Data Variables in Crime Using R. International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE), 5(7), pp-45,2016.
[8] Bhavani, R., and Sudha G. Sadasivam. "Gene expression data classification using MapReduce version of KNN hybridized with PSO." RESEARCH JOURNAL OF BIOTECHNOLOGY 11.7: 37-41,2016.
[9] He, Ting-Qin, et al. "Queuing-Oriented Job Optimizing Scheduling In Cloud Mapreduce." International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. Springer International Publishing, 2016.
[10] Liu, Yang, et al. "A MapReduce Based High Performance Neural Network in Enabling Fast Stability Assessment of Power Systems." Mathematical Problems in Engineering , Hindawi, Volume 2017, 2016.
[11] Gao, S., Li, L., Li, W., Janowicz, K., & Zhang, Y. (2017). Constructing gazetteers from volunteered big geo-data based on Hadoop. Computers, Environment and Urban Systems, 61, 172-186.
[12] Pradhananga, Yanish, Shridevi Karande, and Chandraprakash Karande. "High performance analytics of bigdata with dynamic and optimized hadoop cluster." Advanced Communication Control and Computing Technologies (ICACCCT), 2016 International Conference on. IEEE, 2016.
[13] Kune, R., Konugurthi, P. K., Agarwal, A., Chillarige, R. R., & Buyya, R. XHAMI–extended HDFS and MapReduce interface for Big Data image processing applications in cloud computing environments. Software: Practice and Experience, 47(3), 455-472, 2017.
[14] Meng, Bing, et al. "A novel approach for efficient accessing of small files in HDFS: TLB-MapFile." Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2016 17th IEEE/ACIS International Conference on. IEEE, 2016.
[15] Flood, Mark D., H. V. Jagadish, and Louiqa Raschid. "Big data challenges and opportunities in financial stability monitoring." Banque de France, Financial Stability Review 20, 2016.
[16] Hu, Jiankun, and Athanasios V. Vasilakos. "Energy Big data analytics and security: challenges and opportunities." IEEE Transactions on Smart Grid 7.5 ,pp: 2423-2436,2016.
[17] Song, Houbing, et al. "Next-generation big data analytics: State of the art, challenges, and future research topics." IEEE Transactions on Industrial Informatics , 2017.

AUTHOR PROFILE

Dr. R.Anandhi got M.C.A., M.Phil, in Computer Science from Bharathidasan University, Tiruchirapalli, TamilNadu (2001) and Bharathiar University, Coimbatore, TamilNadu (2005) respectively. She cleared NET and SET in Computer Science. She got Ph.D. in Computer Science Applications from SCSVMV University, Kanchipuram, Tamilnadu (2015). She has so far contributed 7 papers in International Journals and 6 papers in International and National Conferences. Her research areas are Cloud Computing and Distributed Computing. She is currently working as an Assistant Professor, Department of Computer Applications, D.G. Vaishnav College, Arumbakkam, Chennai, Tamilnadu..

Dr. G.Sekar got M.Sc., M.Phil, in Computer Science from Bharathidasan University, Thiruchirapalli, TamilNadu (1997) and Manonmanian Sundaranar University, Tirunelveli, TamilNadu (2002) respectively. He got Ph.D. in Computer Science from Bharathiar University, Tamilnadu (2017). He has so far contributed 7 papers in International Journals and 7 papers in International Conferences. His research areas are Distributed Computing and Databases. He is currently working as an Assistant Professor, PG & Research Department of Computer Science, Dr. Ambedkar Government Arts College, Vyasarpadi, Chennai, Tamilnadu. He is a life member of Computer Society of India (CSI).