Big Data Technologies: Brief Overview

Yojna Arora
PhD Scholar
Gyan Vihar School of Engineering & Technologies
Suresh Gyan Vihar University, Jaipur, India

Dinesh Goyal, PhD
Gyan Vihar School of Engineering & Technology
Suresh Gyan Vihar University, Rajasthan, India

ABSTRACT

In the current scenario, big data is the biggest challenge for the industries to deal with. It is characterized by Huge Volume, Heterogeneous unidentified sources, High rate of data generation, inability to extract value information from irrelevant data. There are many approaches been put forward for dealing with this Big Data, some of them are RDBMS, Hadoop, Cloud Computing etc. This review article includes an elicitation of definitions of Big Data from some previous work, its characteristics, applications, various implementation techniques proposed for dealing with Big Data. It also discusses about some of the benchmarks which are proposed by companies.

Keywords

Big Data, Hadoop, Map Reduce

1. INTRODUCTION

1.1 Big Data Definitions

Big Data is defined as amount of data just beyond technology's capability to store manage and process efficiently [1]

Big data that is too fast, too big or too hard for existing tools to process [2]

Big Data is a term defining data that has three characteristics. First is the great volume of data, second the data cannot be structured into tables and third is velocity which means data is generated rapidly and thus is need to be processed and analyzed fast [3]

Big data concerns l arge volume, complex growing data sets with wit multiple autonomous sources. Stated by Xindong Wu in [4]

In [5] Big data is defined as large amount of data which requires new technologies and architecture so that it becomes possible to extract value from it by capturing and analysis process

Big Data: A massive volume of both structured and unstructured data that is so large that it is difficult to process using traditional database and software techniques [6]

Big Data concerns massive, heterogeneous, autonomous sources with distributed and decentralized control [7]

Data sets that are growing exponentially and that are too large, too raw, or too unstructured for analysis using relational database technique [8]

Sources of big data include: transactions, scientific experiments, genomic investigations, logs, events, emails, social media, sensors, RFID scans, texts, geospatial data, audio data, medical records, surveillance, images, and videos [30]

Data can be categorized into three forms Structured, Semis structured and unstructured such as data coming from E commerce websites, web server logs and social network respectively. [7]

1.2 Characteristics of Big Data

Big data possess various characteristics. The 3D big data characteristics were first introduced in [9] [29].

Further amendments to the definition include the addition of a fourth V, veracity, by IBM [32], emphasizing the aspect of data quality.

Then later 5 V's of Big Data were introduced as explained in [7].

- **a.** **Variety:** It includes the data coming from heterogeneous sources from all the three categorizations i.e. structured, semi structured and unstructured sources
- **b.** **Volume:** By volume it means the enormous amount of data which sis been generated every day.
- **c.** **Velocity :** The extremely fast rate at which the data is generated is termed as velocity, one of the characteristic of Big Data
- **d.** **Value:** The relevance of the important information which is taken out by applying queries over data , means the value
- **e.** **Variability:** The inconsistencies which arise during the data flow

Furthermore, Targeted services, Products, solutions and Applications Data Presentation, Usability and interpretation are other characteristics explained in [10]

[4] Illustrated Complexity: It is quite an undertaking to link, match, cleanse and transform data across systems coming from various sources. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly spiral out of control

1.3 Applications of Big data

The large number of big data applications as explained in [4]

- a. Healthcare
- b. Public sector administration
- c. Retail
- d. Manufacturing
- e. Personal location data
- f. Fact based decision making
- g. Improved customer experience

h. Improved sales

i. New product innovation

j. Reduced risk

k. Higher quality product and services

l. Most efficient operations

2. BIG DATA IMPLEMENTATION TECHNIQUES

2.1 Hadoop [43]

Hadoop is an open source software library that allows for distributed processing of large data sets across clusters of computers using simple programming model explained in [3] [7] .Hadoop includes following modules :

a. **Hadoop Core :** Common utility that support other modules

b. **Hadoop Distributed File System :** A distributed file system that is used for storing files on clusters

c. **HBase :** A distributed database for random Read/Write access

d. **Pig:** A high level data processing system for analyzing data sets that occur a high level language.

e. **Hive:** A data warehousing application that provides a SQL like interface and relational model.

f. **Sqoop:** A project for transferring data between relational databases and Hadoop.

g. **Avro:** A system of data serialization.

h. **Oozie:** A workflow for dependent Hadoop jobs.

i. **Chukwa:** A Hadoop subproject as data accumulation system for monitoring distributed systems.

j. **ZooKeeper:** A centralized service for providing distributed synchronization and group services

2.1.1 Hadoop Distributed File System [7] [12] [13]

HDFS is a very large distributed file system that provides fault tolerance and also has high throughput. HDFS stores files as a series of block and replicates the data blocks for fault tolerance. It stores big data sets and provides global access to files in clusters. HDFS stores metadata on a dedicated server, called “Name Node”. Application data is strpred on another network called “Data Node” All servers are fully connected and communicate with each other using TCP-based protocols. HDFS architecture is composed of four parts

a. **Name Node:** The Name Node is responsible of managing all metadata and file system actions. It handles the file system namespace operations like open, close, and renames both file and directory Also; it makes all decisions regarding replication of blocks. NameNode maintains the tree of namespace and maps the file blocks to DataNodes (i.e. the physical location of file’s data). A single NameNode is considered a bottleneck for handling requests in scientific application environments

b. **Data Node:** The DataNode stores data in the Hadoop file system, Each DataNode stores data blocks on behalf of local or remote clients. Each block is saved as a separated file in the node’s local file system. On startup, DataNode connects to the NameNode and performs a handshake. The purpose of the handshake is to verify the namespace ID and the software version of DataNode. If NameNode does not match DataNode, the DataNode automatically shuts down. After the handshake is successful, the DataNode registers with the NameNode. DataNodes persistently store their unique storage IDs. The storage ID is an internal identifier of the DataNode which makes it as recognizable even if it is restarted with a different IP address or port. The storage ID is assigned to the DataNode, when it registers with the NameNode on the first time and never changes later. The DataNode then responds to the requests that coming from the NameNode, for the file system operations. The DataNodes service the read, writing and file replication requests based on the direction from which NameNode coming

c. **Job Tracker:** The JobTracker talks to the NameNode to determine the location of the data. JobTracker schedules individual maps reduces or intermediate merging operations to specific machines. It monitors the success and failures of these individual tasks. Also, it works to complete the entire batch job. If a task fail, the JobTracker will automatically relaunch the task, possibly on a different node, up to a predefined limit of retries

d. **TaskTracker :** The JobTracker is the master overseeing the overall execution of a MapReduce job. The TaskTrackers manage the execution of individual tasks on each slave node. Although there is a single TaskTracker per slave node. Each TaskTracker can spawn multiple Java Virtual Machines (JVMs) to handle many maps or reduces the tasks in parallel. The TaskTrackers also transmit heartbeat messages to the JobTracker, usually every a few minutes, to reassure the JobTracker that is still a live

2.1.2 Hadoop Drawbacks

From the architecture of Hadoop and its workflow of data computation, there are many drawbacks of Hadoop. These drawbacks are:

a. Hadoop needs high memory and big storage to apply replication technique.

b. Hadoop supports allocation of tasks only and do not have strategy to support scheduling of tasks.

c. Still single master (Name Node) which requires care.

d. Load time is long.

2.2 Grid computing [[3]

Grid Computing is a kind of distributed computing. A grid centre is represented by a number of servers that are interconnected by a high speed network, each of the servers plays one or many roles. The two main benefits of computing in a Grid Centre are the high storage capability and the processing power. A grid centre is composed of :

a. **Computing Elements :** CE manages the resources of the Grid node and manages the jobs launched

b. **Storage Elements :** The SE offers the storage and data transfer services

c. **Worker nodes :** These are the servers that offer the processing power

d. **User Interface (UI)**

2.2.1 Steps followed

a. A user accesses the UI via SSH (Secure Shell) and he receives a Proxy Certificate (PyC).

b. The user than sends the job written in Job Description Language (JDL) and the PyC to the WMS .

c. The WMS checks the PyC and if the needed resources for the job are available. It, then, sends the job and the PyC to the CE.

d. The CE checks the authenticity of the user again and then sends the job to be processed by a WN.

e. The WN computes the job and then sends the results to the WMS and the state of the job to the CE.

f. The users gather the results using the UI and he can store them on the SE.

2.3 Mobile Agent Based [13]

Mobile Agent is software that can migrate during execution across a heterogeneous or homogenous network. The architecture combines the advantages of Mobile Agent & Hadoop. Some of the advantages are:

There are seven reasons for using mobile agents as follows:

1. Reduce the network load
2. Overcome network latency,
3. Encapsulate protocols,
4. Execute asynchronously and autonomously,
5. Adapt dynamically,
6. Naturally heterogeneous and robust, and
7. Fault-tolerant [1].

2.3.1 MRAM Workflow

The workflow of the proposed MRAM framework is shown in Figure 2. It has the following steps:

a. Input text files to the platform.

b. Server portioning the file to blocks with the same size.

c. Application server assigns a data block to each computing node, but in our approach the server takes a task as the other nodes.

d. The computing node runs map on the input data and producing intermediate data pair for every word. It then sends its intermediate data pairs to application server directly to perform the reduce operation.

e. The reduce operation counts the number of occurrences of each word using the values and emits it as a key-value pair and save the result in a file

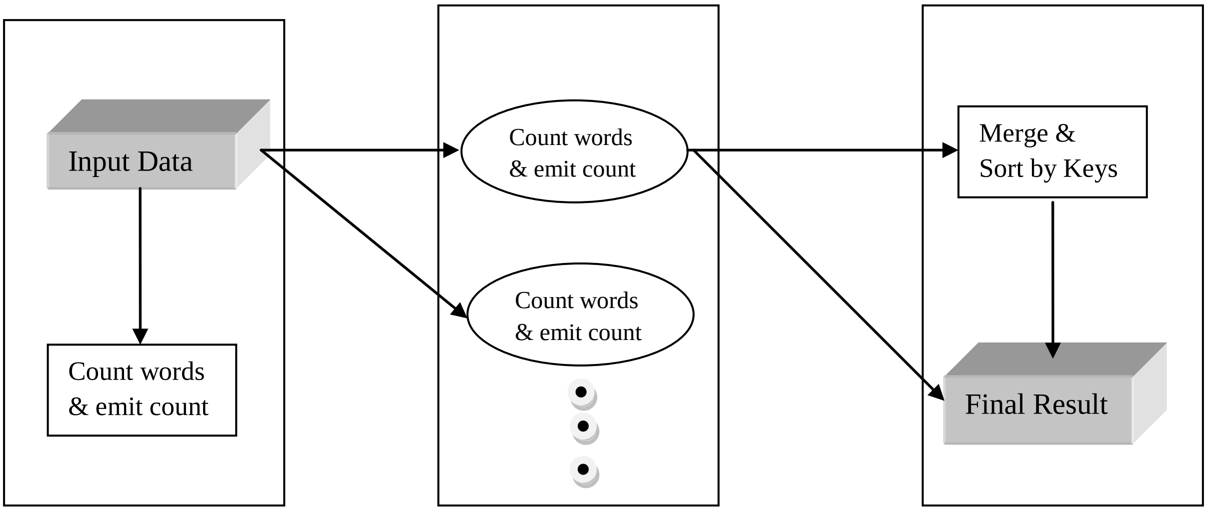


Fig 1. MRAM workflow [13]

The advantages of MRAM are:

a. Support allocation and scheduling tasks

b. Provides fault tolerance and don't need high memory or big disk to support it.

c. Load time for MRAM is less than that of Hadoop.

d. Solve single master (centralized node) problem

e. Using features of mobile agent.)Improve execution time because of no need to huge processing to replication data.

2.4 RDBMS [14]

a. Selection and joining fast

b. Volume increase RDBMS better

2.5 Map Reduce& DFS [2] [11]

In MapReduce [10], the first step is the map job which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job then takes the output from a map as input and combines those data tuples into a smaller set of pairs. The map function can run independently on each key/value pair, exposing enormous amounts of parallelism. Similarly, the reduce function can run independently on each intermediate key, exposing significant parallelism as well. Similar to other distributed systems, MapReduce also constitutes a master and a set of workers. The master is called JobTracker, while the workers are called TaskTrackers

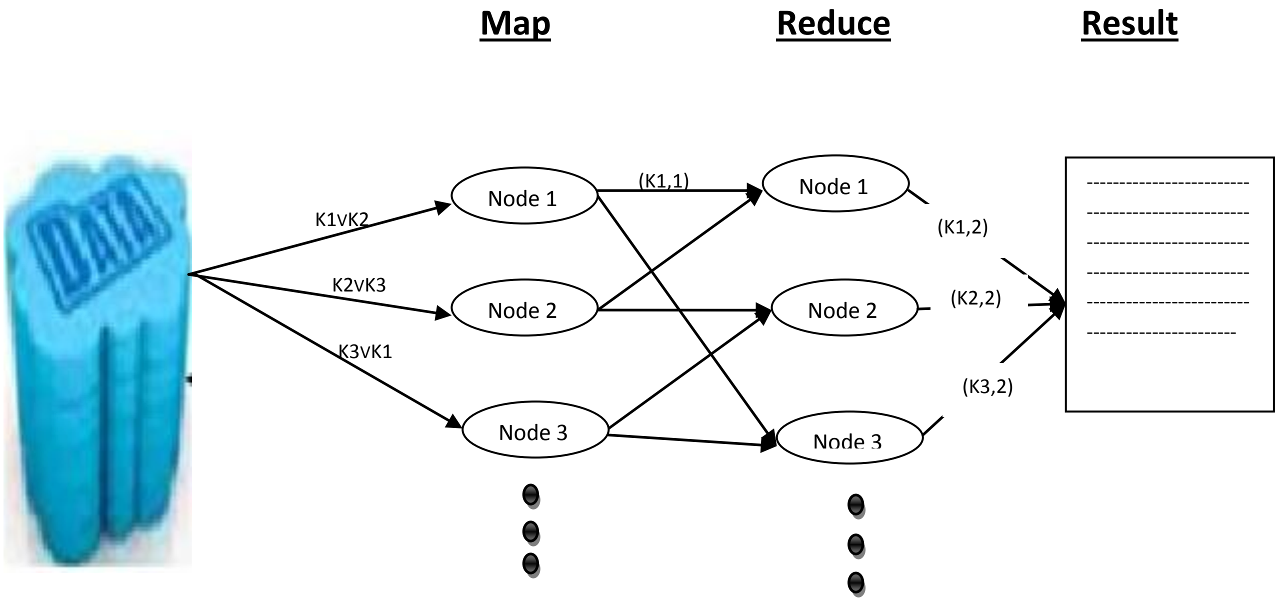
Features :

a. Volume a problem

b. Support advanced analysis of unstructured data

c. Good for large data variety and velocity

d. Setup cost (Purchase + migration + Consulting + Training



2.6 HPCC(High Performance Computing Cluster)

The three main HPCC components are [4]

i.HPCC Data Refinery (Thor) is a massively parallel ETL engine that enables data integration on a scale and provides batch oriented data manipulation.

ii.HPCC Data Delivery Engine (Roxie) is a massively parallel, high throughput, ultra fast, low latency, allows efficient multi user retrieval of data and structured query response engine.

iii.Enterprise Control Language (ECL) is automatically distributes workload between nodes, has automatic synchronization of algorithms, develop extensible machine

Learning library has simple usage programming language optimized for big data operations and query transactions.

2.7 Analytics methods using Data Mining as explained in [16] are

a. Classification: Classification is the process of arranging a specific object in set of categories, based on the each concept attributes. Data classification consists of two-step process. The first step building a data model and determining data classes in advance using training sample. This step is also referred to as supervised learning. The next step classifies the dataset using model or classifier as determined by class label attributes. [28]

b. Prediction: Classification and prediction are the most commonly used mining techniques and have a lot in common. The main difference between two methods can be described as follows: Prediction predicts continuous value whereas classification predicts discrete value like class label

c. Clustering: Clustering also can be explained based on classification. Classification is done by learning the samples in advance, whereas clustering defines a class without a prior learning to classify data. Clustering is the process of grouping data which have multiple properties based on similar attributes without predefined criteria

d. Association: Association rule discovery is the process of finding interesting relationship or correlation among databases. Interesting relationship means useful patterns or frequent item sets. Typically used algorithm is the Apriori algorithm. Because the word ‘useful’ is a subjective meaning, useful value differs depending on user's purpose or environment and so on. This is commonly used in service fields. Also it is often described by beer and diaper correlation.

2.8 Knowledge Extarction

Data ------ Information ------ Knowledge ------- Meta knowledge------Expertise

a. **Noise :** Items that carry no content of knowledge

b. **Data :** Value drawn from some domain of discourse

c. **Information :** Meanings of data values as understood by those who use them

d. **Knowledge:** Specialized information about some some domains that allow one to make decisions.

e. **Expertise:** Specialized operative knowledge that is inherently task specific and inflexible.

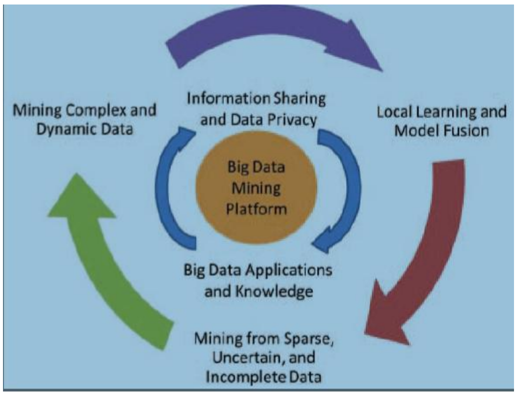

Information Sharing and Data Privacy
Mining Complex and Dynamic Data
Local Learning and Model Fusion
Big Data Mining Platform
Big Data Applications and Knowledge
Mining from Sparse, Uncertain, and Incomplete Data

Fig 3: Big Data Processing Network

In [17], two dimensions of big data are considered i.e. Service generated Big Data or Big Data as service. It takes care of service logs, service QOS and service relationship, along with providing Big Data infrastructure, Platform and Software as service.

3. CONCLUSION

The paper gives a detailed description of Big Data, its characteristics, applications and implementation techniques. Big Data has various definations given in various sources but all talks about five characteristics of Big Data i.e Volume, Variety, Velocity, Value and Variability. Big data has its application in almost all the spheres ranging fro financial sectore to healthcare. Many techniques are implemented to solve the Big Data problem. It includes open source software Hadoop, Grid Computing, Mobile Computing, Cloud Computing, Data Mining and many more. But yet there is a scope for further research as it is surrounded with great issues on a broad classification it is issues related to Storage, Processing and Management of Big Data. The volume of Data been generated every day structured, unstructured or semi structured need sufficient storage to be kept and taken care of. Further the issues related to fast generation of results are still pertaining. Also, management issues related to Big Data are still under considerations.

4. REFERENCES

[1] Stephen Kaisler, Frank Arrmour, J. Alberto," Big Data: Issues and Challenges Moving Forward",46th Hawaii International Conference on System Science, IEEE,2012

[2] Sam Padden, "From database to Big Data,", in IEEE Computer Society, 2012

[3] Dan Garlasu, "Data Implementation Based on Grid Computing",

[4] Avita Katal, Mohammad Wazid and R H Goudar, "Big Data : Issues, Challenges, Tools and good Practices", in IEEE 2013

[5] Seref Sagiroglu and Duygu Sinang, "Big Data : A Review",IEEE, 2013

[6] Yuri Demchenko, Paolo Grosso andCees de Laat, "Addressing Big Data Issues in Scientific Data Infrastructure", in IEEE 2013

[7] Parth Chandarana and M Vijayalakshmi, "Big Data Analytics Framework", in International Conference on Circuits, System, Communication and Information Technology Applications",IEEE, 2014

[8] Rich Adduci, Dave Blue and Guy Chiarello, "Big Data : Big Opportunities to create Business value", in EMC2

[9] Doug Laney, "3 D Data Management : Controlling Data Volume, Velocity and Variety", in Application Delivery Stratergies, Meta Group, 2001

[10] First Tekiner and John A keane, "Big Data Framework", in IEEE international conference on Systems, Man and cybernetics, IEEE, 2013

[11] Janusz Weilki, "Implementation of Big Data Concept in organizations- possibilities, impediments and challenges", proceeding of 2013 Federated conference on computer science and information systems, pp985-989, IEEE, 2013

[12] Edmund Kohlwey, Abel Sussman, Jason Trost and Amber Maurer, "Leveraging the cloud for Big data Biometrics", in World Congress in Services, IEEE,2011

[13] Youseef M Essa, "Mobile Agent Based New Framework for improving Big Data Analysis", in International Conference on Cloud Computing and Big Data, IEEE, 2013

[14] Katharina Ebner, Thilo Buhnen and Nils Urbach, "Think Big with Big Data: Identifying Suitable Big Data Strategies in Corporate Environment", 47th Hawaii International Conference on System Science", IEEE, 2014

[15] Dr Daniel Fasel, " Potentials of Big Data for Governmental Services",

[16] Sung Hwan Kim, Nam UK Kim and Tai Myoung Chung, "Attribute Relationship Evaluation Methodology for Big Data Security", in IEEE, 2013

[17] Zibin Zheng, Jiemming Zhu and Michael R Lyu, "Service generated big data and big data as a service : An Overview", in IEEE International Congress on Big Data, 2013

[18] Wen Xiong, Zibin Yu, and Zhendong Bei, "A characterization of Big Data Benchmarks", in IEEE international conference on Big Data", IEEE,2013

[19] Xindong Wu, Xingquan Zhu, Gong Qing Wu and Wei Ding, "Data Mining with Big Data", in IEEE transactions in knowledge and data engineering, Vol 26, Number 1, January 2014

[20] Lei Wang, Jainfeng Zhan and Chunjie Luo, "Big Data Bench: A Big Data Benchmark suite from Internet Services", in IEEE, 2014

[21] Xueli Huang and Xiaojiang Du, "Achieving Big Data Privacy via Hybrid Cloud", in Infocom workshop on Security and Privacy on Big Data, IEEE, 2014

[22] Barna Saha and Divesh Srivastava, "Data Quality : The other face of Big Data", in IEEE, 2014

[23] Carol J Romanowski and Rajender K Raj, " Catching the wave : Big Data in the classroom", in IEEE 2013

[24] Du Zhang, " Inconsistencies in Big Data", in Proceeding of IEEE international conference on Cognitive Informatics and Cognitive Computing" IEEE, 2013

[25] Marcus R. Wigan and Roger Clarke. “Big Data’s big unintended consequences”, in IEEE Computer Society, 2013

[26] Jinsong Zhang, Yan Chen and Taoying Li, “Opportunities of Innovation under challenges of Big Data”, in 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE, 2013

[27] Lin Gu, Deze Zeng and Peng Li, “Cost Minimization for Big data Processing in Geo Distribeuted Data Centers”, in IEEE transactions on Emerging topics in Computing, 2014

[28] J Han, M Kamber and J Pei, “Data Mining Concepts and techniques”, Morgan Kaufmann, 2006

[29] B.Gerhardt, K. Griffin and R. Klemann, "Unlocking Value in the Fragmented World of Big Data Analytics", Cisco Internet Business Solutions Group, June 2012,

[30] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, Analytics: the real-world use of big data: how innovative enterprises extract value from uncertain data, Executive Report, IBM Institute for Business Value and Said Business School at the University of Oxford, 2012.

[31] W. GAO, et al. “BigDataBench: a Big Data Benchmark Suite from Web Search Engines”. The Third Workshop on Architectures andSystems for Big Data (ASBD 2013) in conjunction with ISCA 2013.

[32] Morgan, T., Ibm Global Technology Outlook 2012, Warwick, 2012.