Survey Paper on Big Data

Sandeep Singh[1] Prithvipal Singh[2]
[1]Assistant Professor [2]Research Scholar

Abstract— This paper gives the impending of how Big Data can uncover additional value from the unstructured data(weblogs, text messages, videos, picture, social updates) generated by sensors, smart phone's, satellites, laptops, computers, organizations, social networking sites like Face book, Twitter, Yahoo, Google, YouTube etc. Big data term is used for extracting useful information from huge volume data which ranges in Exabyte, Zettabyte and beyond. Big data is an advantage over traditional systems. The technologies used by Big Data are Hadoop, Map Reduce, Hive, Pig, HDFS, Hbase. This paper reveals how various technologies deal with huge data.

Key words: Big Data, Hadoop, HDFS, Hive, Pig, Hbase, Map Reduce

I. INTRODUCTION

Big Data" is a term encompassing the use of techniques to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies"[8]

McKinsey & Co [9] defines big data is "the next frontier for advancement, competition and productivity". Big Data is not only for contest and expansion for all companies, but the precise use of Big Data also can increase productivity, innovation, and competitiveness for whole sectors and economies.

It said that 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This big volume of data comes from all over the place: sensors used to collect climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This massive amount of the data is known as "Big data"[10]. Big data is a catch-phrase, uses to describe a giant amount of both structured and unstructured data that is so complicated to process using traditional database and software techniques.

Big data helps companies, regulatory authority, university and research centres organizations to improve operations and make faster, more intelligent decisions.

Big data is actually an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Big data can be of any size i.e. petabytes, exabytes, Zettabyte, etc. Big data is an large-scale term for large collection of the data sets so this huge and complex that it becomes difficult to operate them using traditional data processing applications. When dealing with big datasets, organizations encountered with difficulties to create, manipulate, and supervise big data. Big data is mostly a trouble in business analytics because standard tools and procedures are not designed to explore and analyze massive datasets. An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible[11].

Biggest challenges of big data can be characterised by the following five V's are: Volume, Variety, velocity, value, veracity[12].

1) Volume: Data can be of any type MB, PB, YB, ZB, KB, TB of information. The data results into large files. Extreme amount of data with results as well is the actual problem of storage and it can be solved by reducing storage cost. It believes that growth of data volumes can be 50 times by 2020.

Organization	Volume of Data
YouTube[13]	(i) 300 hours of videos are uploaded to YouTube every minute. (ii) Each month, more than 1 billion unique users access YouTube (iii) Every day people watch hundreds of millions of hours on YouTube and generate billions of views. The number of hours people are watching on YouTube each month is up 50% year over year
Facebook[14]	(i)In every 20 minutes 3 million message sent (ii)1 million links are shared in every 20minutes. (iii)total number of monthly active users1,310,000,000 (iv)Average number of photos uploaded per day205.
Twitter[15]	(i)The site has over 645,750,000 users. (ii) The site generates 175 million tweets per day
Google+[16]	1 billion account has been created
Instagram[16]	Users share 40million photos per day
Linkedin[17]	(i) Total number of Linked users 313,000,000

Table 1: Volume of unstructured data in various companie

2) Variety: Data comes from the mixture of sources i.e. It can be any format of any type and it may be structured or unstructured such as text, audio,videos, log files and etc. The varieties are eternal, and the data came in the network without having been quantified or qualified in any way.
3) Velocity: The data comes at high speed. Sometimes 1 minute is too late so big data is time sensitive. Some organisations data velocity is main challenge. The social media messages and credit card transactions done in millisecond and data generated by this putting in to databases.
4) Value: Which addresses the need for valuation of enterprise data? It is a most important v in big data. Value is main buzz for big data because it is important for businesses, IT companies to store large quantity of values in database.

All rights reserved by www.ijsrd.com

5) Veracity: The increase in the range of values typical Of a large data set. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be polluted data. Big data and analytics technologies work with these types of data.

This paper is comprises of four sections. In section 2 literature survey has been discussed. In section 3 the various Big Data techniques has been discussed. In section 4 future scope has been discussed and the last section gives a conclusion of this paper.

II. LITERATURE SURVEY

J.Archenaa [4] in 2015 explained the large amount of heterogeneous data by healthcare and government agencies. With the help of Big Data Analytics using Hadoop plays an efficient role in performing meaningful real-time analysis on huge volume data and able to predict the emergency situations before it happens. It also providing an effective data that effectively helps the citizen care management. It describes about the big data use cases in healthcare and government.

Ranjana Bahri [3]in 2015 author describes the concept, challenges and management tools. Challenges like heterogeneity, incompleteness, scale, timeliness and data security. The tools and techniques are used for Big data management are Google BigTable, Simple DB, Not Only SQL (NoSQL), Data Stream Management System (DSMS),MemcacheDB, and Voldemort. For Big Data, some of the most commonly used tools and techniques are Hadoop, MapReduce, and Big Table.

Rohit Pitre [7] in 2014 authors talks about Mining platform, Privacy and Design of mining algorithms. Discuss the Challenging issues and its related work in data mining with Big Data.

Sangeeta Bansal [1] in 2014 author draw an analogy for data management between the traditional relational database systems and the Big Data. Author also discusses the Challenges in traditional data management using relational databases in an enterprise and International Data Corporation (IDC) believes organizations that are best able to make real-time business decisions using Big Data solutions will thrive, while those that are unable to embrace and make use of this shift will increasingly find themselves at a competitive disadvantage in the market and face potential failure.

Sameera Siddiqui [6] in 2014 discussed the summary of various surveys done by companies like TCS and IDC Enterprise on Big Data. Also discuss how big data is important to increase productivity growth in the entire world since it is affecting not only software-intensive industries but also public domains like education, health field ,education and administrative sector.

Yaxiong Zhao [5] in 2014 proposed data aware caching (Dache) framework that made minimum change to the original map reduce programming model to increment processing for big data applications using the map reduce model. It is a protocol, data aware cache description scheme and architecture. The advantage of this paper is, it improves the completion time of map reduce jobs.

Jian Tan [5] in 2013 author talks about the theoretical assumptions, that improves the performance of Hadoop/map reduce and purposed the optimal reduce task assignment schemes that minimize the fetching cost per job and performs the both simulation and real system deployment with experimental evolution. It also improves the performance of large scale Hadoop clusters.

Thuy D. Nguyen [5] (2013) author solve the multilevel secure (MLS) environmental problems of Hadoop by using security enhanced Linux (SE Linux) protocol. In which multiple sources of Hadoop applications run at different levels. This protocol is an extension of Hadoop distributed file system (HDFS).

Xin Luna Dong [5] in 2013 explained challenges of big data integration (schema mapping, record linkage and data fusion). These challenges are explained by using examples and techniques for data integration in addressing the new challenges raised by big data, includes volume and number of sources, velocity, variety and veracity. It also identifying the data source problems to integrate existing data and systems.

III. BIG DATA TECHNOLOGIES

However, the new Big Data technology improves performance, facilitates innovation in the products and services of business models, companies, regulatory authority, university and research centres and provides decision making support. Properly managed Big Data are accessible, reliable, secure, and manageable. Hence, Big Data applications can be applied in various complex scientific disciplines, including atmospheric science, astronomy, medicine, biology, genomics, and biogeochemistry.

A. Hadoop

Hadoop is written in Java and is a top-level Apache project that started in 2006.Doug Cutting developed Hadoop as a collection of open-source projects on which the Google MapReduce programming environment could be applied in a distributed system. While using Hadoop, organization can handle data that was earlier complex to administer and analyze. Now, Hadoop can do enormously bulky amount of data with altering structures . Hadoop is composed of HBase, HCatalog, Pig, Hive, Oozie, Zookeeper, and Kafka. Table2 sows the hadoop components and its functionality in brief. [18]

Hadoop Component	Functions
(1) HDFS	Storage and replication
(2) MapReduce	Distributed processing and fault tolerance
(3) HBASE	Fast read/write access
(4) HCatalog	Metadata
(5) Pig	Scripting
(6) Hive	SQL
(7) Oozie	Workflow and scheduling
(8) ZooKeeper	Coordination
(9) Kafka	Messaging and data integration
(10) Mahout	Machine learning

Fig. 1:

All rights reserved by www.ijsrd.com

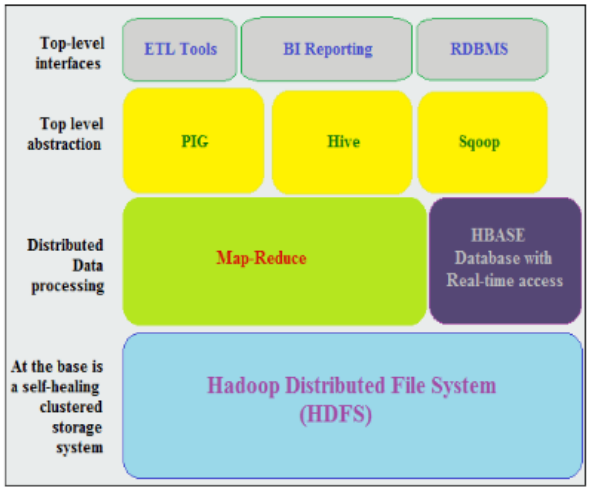

Fig. 2: Architecture of Hadoop[5]

B. Map Reduce

Map-Reduce was introduced by Google in order to process and store large datasets on commodity hardware. It is used for processing big amount of data records in clusters. The Map Reduce programming model is comprised of two functions i.e. Map() function and Reduce() function. Users can create their own processing logics having well clear Map() and Reduce() functions. Map function working Starts from master node that takes input and that divides into smaller sub modules and distribute among slave nodes. Slave node additional divides the submodules again that form the hierarchical tree structure. However, slave node processes the base problem and passes the result reverse to the master Node. The Map Reduce system arrange together all intermediate pairs based on the intermediate keys and refer them to Reduce() function for producing the final output. Reduce function works as the master node collects the results from all the sub problems and combines them together to form the output.

Map(in_key,in_value)-->list(out_key,intermediate_value)
Reduce(out_key,list(intermediate_value))-->list(out_value)

A Map Reduce framework is based on a master slave architecture and one master node handle a number of slave nodes . Map Reduce first dividing the input data set into blocks of even-sized data for equalizing load distribution. Each data block is then assigned to one slave node and is processed by a map task and result is generated. The slave node interrupts the master node when it is idle. The scheduler then assigns new tasks to the slave node. The scheduler takes data locality and resources into consideration when it disseminates data blocks. [5]

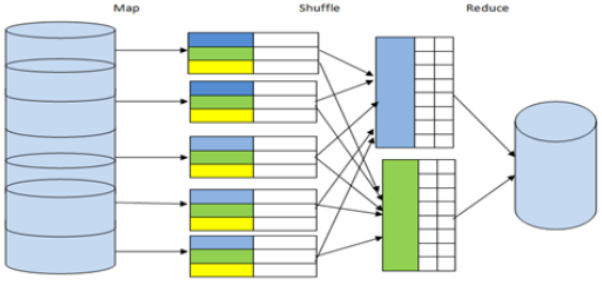

Fig. 2: Architecture of Map Reduce

C. Map Reduce Components

1) Name Node: deals and manages HDFS metadata
2) Data Node: stores blocks of HDFS
3) Job Tracker: schedules, allocates and monitors job execution on slaves—Task Trackers.
4) Task Tracker: runs Map Reduce operations.[5]

D. HIVE

Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. While initially developed by Facebook, Apache Hive is now used and developed by other companies such as Netflix. Amazon maintains a software fork of Apache Hive that is included in Amazon Elastic Map Reduce on Amazon Web Services in Fig-1. Apache Hive supports analysis of large datasets stored in Hadoop's HDFS and compatible file systems such as Amazon S3 file system[11]. It provides an SQL-like language called HiveQL with schema on read and transparently converts queries to map/reduce, Apache Tez and in the future Spark jobs. All three execution engines can run in Hadoop YARN. To accelerate queries, it provides indexes, including bitmap indexes. By default, Hive stores metadata in an embedded Apache Derby database, and other client/server databases like MySQL[11] can optionally be used. Currently, there are four file formats supported in Hive, which are TEXTFILE SEQUENCEFILE, ORC and RCFILE. Other features of Hive include:

- Indexing to provide acceleration, index type including compaction and Bitmap index as of 0.10, more index types are planned.
- Different storage types such as plain text, RCFile, HBase, ORC, and others.
- Metadata storage in an RDBMS, significantly reducing the time to perform semantic checks during query execution.
- Operating on compressed data stored into Hadoop ecosystem, algorithm including gzip, bzip2, snappy, etc.
- Built-in user defined functions (UDFs) to manipulate dates, strings, and other data-mining tools. Hive supports extending the UDF set to handle use-cases not supported by built-in functions.
- SQL-like queries (HiveQL), which are implicitly converted into MapReduce jobs.

While based on SQL, HiveQL does not strictly follow the full SQL-92 standard. HiveQL offers extensions not in SQL, including multitable inserts and create table as select, but only offers basic support for indexes[19]. Also, HiveQL lacks support for transactions and materialized views, and only limited subquery support. There are plans for adding support for insert, update, and delete with full ACID functionality[2].

E. HBASE

Apache HBase began as a project by the company Powerset out of a need to process massive amounts of data for the purposes of natural language search.Facebook elected to implement its new messaging platform using HBase in November 2010.

HBase is a column-oriented database management system that runs on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases.

All rights reserved by www.ijsrd.com

Unlike relational database systems, HBase does not support a structured query language like SQL; in fact, HBase isn't a relational data store at all. HBase applications are written in Java much like a typical MapReduce application. HBase does support writing applications in Avro, REST, and Thrift. An HBase system comprises a set of tables. Each table contains rows and columns, much like a traditional database. Each table must have an element defined as a Primary Key, and all access attempts to HBase tables must use this Primary Key.

An HBase column represents an attribute of an object[11]. In fact, HBase allows for many attributes to be grouped together into what are known as column families, such that the elements of a column family are all stored together. This is different from a row-oriented relational database, where all the columns of a given row are stored together. With HBase you must predefine the table schema and specify the column families. However, it's very flexible in that new columns can be added to families at any time, making the schema flexible and therefore able to adapt to changing application requirements.

Just as HDFS has a Name Node and slave nodes, and MapReduce has Job Tracker and Task Tracker slaves, HBase is built on similar concepts. In HBase a master node manages the cluster and region servers store portions of the tables and perform the work on the data. In the same way HDFS has some enterprise concerns due to the availability of the NameNode (among other areas that can be "hardened" for true enterprise deployments by InfoSphere BigInsights), HBase is also sensitive to the loss of its master node[2].

F. HPCC

HPCC is an open source platform that gives the service for managing of enormous big data . HPCC data model is defined by the user end according to the requirements. HPCC system is designed to manage the most complex and data-intensive analytical connected problems. HPCC system is a single platform, single architecture and a single programming language that can be implemented on data simulation. HPCC system implemented to analyze the extremely large amount of data for the purpose of solving complex problem of big data.

HPCC system is based on enterprise control language that can be declarative and on-procedural nature programming language.

The main components of HPCC are[5]:

- *HPCC Data Refinery*: Use parallel ETL engine mostly.
- *HPCC Data Delivery*: It uses structured query engine.
- Enterprise Control Language distributes the workload between the nodes in appropriate even load.

IV. FUTURE SCOPE

Today big data significantly influencing IT companies and through innovated development of new technologies only big data analytics can deal properly with refine results. Big data completely make a way for companies, regulatory authority, university and research centres by using number of tools to make the use of big data. Hadoop will be greatly in demand and the amount of data produced by association in next few years will be larger than today. In coming years cloud will play the significant role for every sectors and organisations to knob the big data impressively.

V. CONCLUSIONS

In this paper we conclude various technologies to handle the big data and discussed the challenges of big data (volume, variety, velocity, value, veracity). This paper discussed an architecture using Hadoop, MapReduce, Hive, Habse, and HPCC. Objective of our paper was to do a survey of various big data managing techniques those hold large volume of data from different sources and improves on the whole performance of systems.

REFERENCES

[1] Sangeeta Bansal, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X,2014.
[2] Vibhavari Chavan, IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7932-7939.
[3] Ranjana Bahri, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X.
[4] J.Archenaa, 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Procedia Computer Science 50 (2015) 408 – 413.
[5] Sabia, International Conference on Communication, Computing & Systems (ICCCS–2014).
[6] Sameera Siddiqui, International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-3, Issue-7).
[7] Rohit Pitre, A Survey Paper on Data Mining With Big Data, International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume 1 Issue 1 (April 2014).
[8] Grand Challenge: Applying Regulatory Science and Big Data to Improve Medical Device Innovation, Arthur G. Erdman*, Daniel F. Keefe, Senior Member, IEEE, and Randall Schiestl, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 60, NO. 3, March 2013.
[9] McKinsey Global Institute, Big Data: The next frontier for innovation, competition and productivity (June 2011).
[10] Apache Hive. Available at http://hive.apache.org.
[11] Apache HBase. Available at http://hbase.apache.org.
[12] Yuri Demchenko "The Big Data Architecture Framework(BDAF)" Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013.
[13] YouTube —YouTube Statistics Feb24, 2015‖https://www.youtube.com/yt/press/statistics.html.
[14] Facebook, Facebook Statistics, jan 27,2015, http://www.statisticbrain .com/facebook-statistics/.
[15] Twitter, —Twitter statistics,‖ 2015, http://www.statisticbrain .com/twitter-statistics/.
[16] A Review of Big Data Management, Benefits and Challenges 1 Oguntimilehin A., 2 Ademola E.O. 1,2Department of Computer Science, AfeBabalola University, Ado-Ekiti, Nigeria.

All rights reserved by www.ijsrd.com

[17] Linkedin Statistics‖, October 28th, 2014 http://www.statisticbrain.com/linkedin-company-profile-and-statistics/.
[18] Review Article:Big Data: Survey, Technologies, Opportunities, and Challenges By Nawsher Khan,1,2Ibrar Yaqoob,1 Ibrahim Abaker Targio Hashem,1 Zakira Inayat,1,3Waleed Kamaleldin Mahmoud Ali,1 Muhammad Alam,4,5 Muhammad Shiraz,1and Abdullah Gani1.
[19] Vishal S Patil, Pravin D. Soni, "Hadoop Skelton & Fault Tolerance in Hadoop Clusters", International Journal of Application or Innovation in Engineering & Management(IJAIEM)Volume 2, Issue 2, February 2013 ISSN 2319 – 4847.

All rights reserved by *www.ijsrd.com*