Survey on Big data

Manju Lakshmi[1] Smita C Thomas[2]
[1,2]Mount Zion College of Engineering, Kadammanitta, Kerala, India

Abstract— Big data is the term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on." Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of its ranging from a few dozen terabytes to many petabytes of data.

Key words: Big data, Hadoop

I. Introduction

At present, the industry does not have a bound together meaning of Big Data. It has been characterized in varying courses as takes after by different gatherings: As per McKinsey, "Enormous Data alludes to datasets whose size are past the capacity of regular database programming apparatuses to catch, store, oversee and break down". IDC characterizes Big Data advancements as another era of advances and models intended to concentrate esteem financially from substantial volumes of a wide assortment of information by empowering high speed catch, revelation and investigation. As indicated by O'Reilly, "Enormous Data will be information that surpasses the handling limit of ordinary database frameworks. The information is too huge, moves too quick, or does not fit the structures of existing database designs. To pick up quality from these information, there must be an option approach to process it." As indicated by Wikipedia, "Huge Data for the most part incorporates datasets with sizes past the capacity of generally utilized programming apparatuses to catch, clergyman, oversee, and handle the information inside a decent slipped by time". As indicated by Gartner, "Huge Data is high volume, high speed, and/or high assortment data resources that require new types of handling to empower improved basicleadership, knowledge revelation, and procedure enhancement". In the nutshell, efficacy of Big Data is that it is used to describe massive volumes of unstructured and structured data that are so large that it is very difficult to process this data using traditional databases and software technologies.

This much amount of data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This huge amount of the data is known as “Big data”[14]. Big data is a buzzword, or catch-phrase, utilizes to describe a massive volume of both structured and unstructured data that is so huge that it's complicated to process using traditional database and software techniques. In most enterprise scenarios the data is too large or it moves too fast or it exceeds current processing capacity. Big data has the potential to help organizations to improve operations and make faster, more intelligent decisions[5]. Big Data, now a days this term becomes common in IT industries. As there is a huge amount of data lies in the industry but there is nothing before big data comes into picture [3]. 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. Although big data doesn't refer to any specific quantity, so this term is often used when speaking about petabytes and exabytes of data[6]. Big data is an all-encompassing 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 larger datasets, organizations face difficulties in being able to create, manipulate, and manage big data. Big data is particularly a problem in business analytics because standard tools and procedures are not designed to search 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[15]. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on "Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;as of 2012, every day 2.5 exabytes (2.5×1018) of data were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization

II. Literature survey

Hadoop Map Reduce is a large scale, open source software framework dedicated to scalable, distributed, data-intensive computing. The framework breaks up large data into smaller parallelizable chunks and handles scheduling

- Maps each piece to an intermediate value

- Reduces intermediate values to a solution
- User-specified partition and combiner options Fault tolerant, reliable, and supports thousands of nodes and petabytes of data

• If you can rewrite algorithms into Map Reduces, and your problem can be broken up into small pieces solvable in parallel, then Hadoop's Map Reduce is the way to go for a distributed problem solving approach to large datasets
• Tried and tested in production
• Many implementation options. We can present the design and evaluation of a data aware cache framework that requires minimum change to the original Map Reduce programming model for provisioning incremental processing for Big Data applications using the Map Reduce model [4].

The author [2] stated the importance of some of the technologies that handle Big Data like Hadoop, HDFS and Map Reduce. The author suggested about various schedulers used in Hadoop and about the technical aspects of Hadoop. The author also focuses on the importance of YARN which overcomes the limitations of Map Reduce.

The author [3] have surveyed various technologies to handle the big data and there architectures. In this paper we have also discussed the challenges of Big data (volume, variety, velocity, value, veracity) and various advantages and a disadvantage of these technologies. This paper discussed an architecture using Hadoop HDFS distributed data storage, real-time NoSQL databases, and MapReduce distributed data processing over a cluster of commodity servers. The main goal of our paper was to make a survey of various big data handling techniques those handle a massive amount of data from different sources and improves overall performance of systems.

The author continue with the Big Data definition and enhance the definition given in [3] that includes the 5V Big Data properties: Volume, Variety, Velocity, Value, Veracity, and suggest other dimensions for Big Data analysis and taxonomy, in particular comparing and contrasting Big Data technologies in e-Science, industry, business, social media, healthcare. With a long tradition of working with constantly increasing volume of data, modern e-Science can offer industry the scientific analysis methods, while industry can bring advanced and fast developing Big Data technologies and tools to science and wider public.[1]

The author [6] stated the need to process enormous quantities of data has never been greater. Not only are terabyte - and petabyte scale datasets rapidly becoming commonplace, but there is consensus that great value lies buried in them, waiting to be unlocked by the right computational tools. In the commercial sphere, business intelligence, driven by the ability to gather data from a dizzying array of sources. Big Data analysis tools like Map Reduce over Hadoop and HDFS, promises to help organizations better understand their customers and the marketplace, hopefully leading to better business decisions and competitive advantages [3].

The author [5] stated there is a need to maximize returns on BI investments and to overcome difficulties. Problems and new trends mentioned in this article and finding solutions by combination of advanced tools, techniques and methods would help readers in BI projects and implementations. BI vendors are struggling and doing continuous effort to bring technical capabilities and to provide complete out of the box solution with set of tools and techniques. In 2014, due to rapid change in BI maturity, BI teams are facing tough time to have infrastructure with less skilled resources. Consolidation and convergence is going on, market is coming up with wide range of new technologies. Still the ground is immature and in a state of rapid evolution.

The author [8] given some important emerging framework model design for Big Data Analytics and a 3-tier architecture model for Big Data in Data Mining. In the proposed 3-tier architecture model is more scalable in working with different environment and also benefits to overcome with the main issue in Big Data Analytics for storing, Analyzing, and visualization. The framework model given for Hadoop HDFS distributed data storage, real-time Nosql databases, and MapReduce distributed data processing over a cluster of commodity servers.

III. Technologies and methods

Big data is a new concept for handling massive data therefore the architectural description of this technology is very new. There are the different technologies which use almost same approach i.e. to distribute the data among various local agents and reduce the load of the main server so that traffic can be avoided. There are endless articles, books and periodicals that describe Big Data from a technology perspective so we will instead focus our efforts here on setting out some basic principles and the minimum technology foundation to help relate Big Data to the broader IM domain.

A. Hadoop

Hadoop is a framework that can run applications on systems with thousands of nodes and terabytes. It distributes the file among the nodes and allows to system continue work in case of a node failure. This approach reduces the risk of catastrophic system failure.

Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Doug Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project. Hadoop is open-source software that enables reliable, scalable, distributed computing on clusters of inexpensive servers[1].

1) Components of Hadoop:

a) HBase:

It is open source, distributed and Non-relational database system implemented in Java. It runs above the layer of HDFS. It can serve the input and output for the Map Reduce in well mannered structure.

b) Oozie:

Oozie is a web-application that runs in ajava servlet. Oozie use the database to gather the information of Workflow which is a collection of actions. It manages the Hadoop jobs in a mannered way.

c) Sqoop:

Sqoop is a command-line interface application that provides platform which is used for converting data from relational databases and Hadoop or vice versa.

d) Avro:

It is a system that provides functionality ofdata serialization and service of data exchange. It is basically used in Apache Hadoop. These services can be used together as well as independently according the data records.

e) Chukwa:

Chukwa is a framework that is used fordata collection and analysis to process and analyze the massive amount of logs. It is built on the upper layer of the HDFS and Map Reduce framework.

f) Pig:

Pig is high-level platform where the MapReduce framework is created which is used with Hadoop platform. It is a high level data processing system where the data records are analyzed that occurs in high level language.

g) Zookeeper:

It is a centralization based service thatprovides distributed synchronization and provides group services along with maintenance of the configuration information and records.

h) Hive:

It is application developed for datawarehouse that provides the SQL interface as well as relational model. Hive infrastructure is built on the top layer of Hadoop that help in providing conclusion, and analysis for respective queries.

Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Doug Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project. Hadoop is open- source software that enables reliable, scalable, distributed computing on clusters of inexpensive.

Hadoop is:

- Reliable: The software is fault tolerant, it expects and handles hardware and software failures
- Scalable: Designed for massive scale of processors, memory, and local attached storage
- Distributed: Handles replication. Offers massively

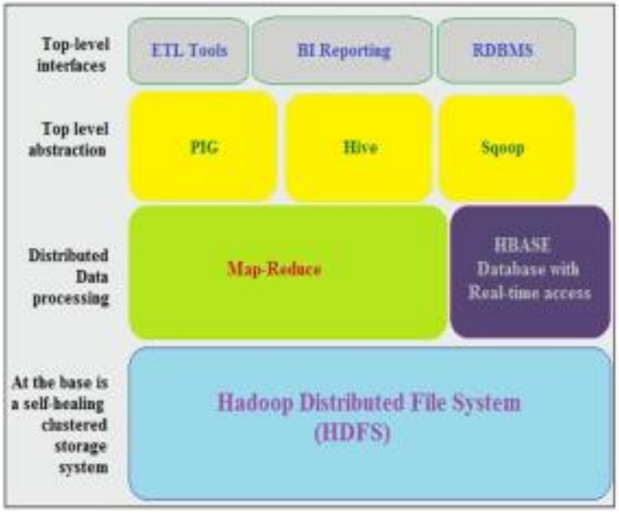
Top-level interfaces
ETL Tools
BI Reporting
RDBMS
Top level abstraction
PIG
Hive
Sqoop
Distributed Data processing
Map-Reduce
HBASE Database with Real-time access
At the base is a self-healing clustered storage system
Hadoop Distributed File System (HDFS)

Fig. 1: Architecture of Hadoop

B. HDFS

The Hadoop Distributed File System (HDFS) is the file system component of the Hadoop framework. HDFS is designed and optimized to store data over a large amount of low-cost hardware in a distributed fashion. Name Node*:* Name node is a type of the master node, which is having the information that means meta data about the all data node there is address(use to talk), free space, data they store, active data node , passive data node, task tracker, job tracker and many other configuration such as replication of data.

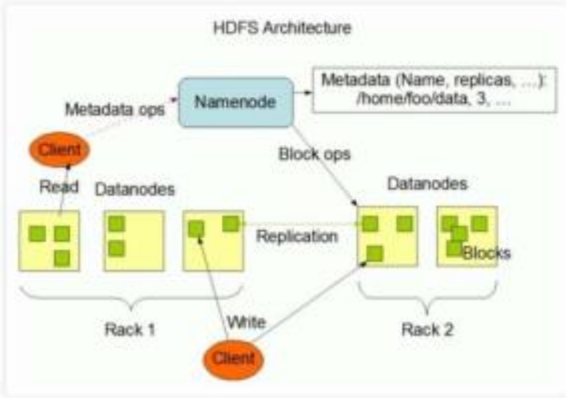

Fig. 2: HDFS Architecture

The NameNode[6] records all of the metadata, attributes, and locations of files and data blocks in to the DataNodes. The attributes it records are the things like file permissions, file modification and access times, and namespace, which is a hierarchy of files and directories. The NameNode maps the namespace tree to file blocks in DataNodes. When a client node wants to read a file in the HDFS it first contacts the Namenode to receive the location of the data blocks associated with that file.

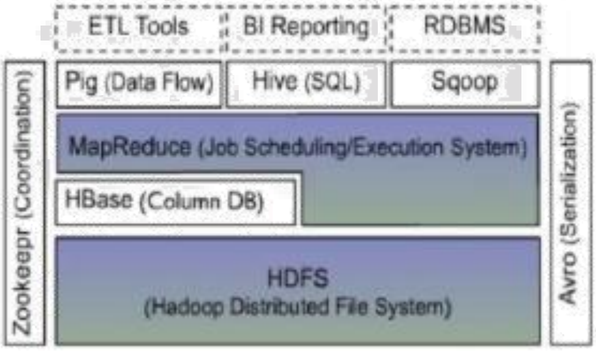

Fig. 3: Hadoop Ecosystem

C. Map Reduce:

Map-Reduce was introduced by Google in order to process and store large datasets on commodity hardware. Map Reduce is a model for processing large-scale data records in clusters. The Map Reduce programming model is based on two functions which are map() function and reduce() function. Users can simulate their own processing logics having well defined map() and reduce() functions. Map function performs the task as the master node takes the input, divide into smaller sub modules and distribute into slave nodes. A slave node further divides the sub modules again that lead to the hierarchical tree structure. The slave node processes the base problem and passes the result back 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) The parameters of map () and reduce () function is as follows: map (k1,v1) ! list (k2,v2) and reduce (k2,list(v2)) ! list (v2)

A Map Reduce framework is based on a master-slave architecture where one master node handles a number of slave nodes . Map Reduce works by first dividing the input data set into even-sized data blocks for equal 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.

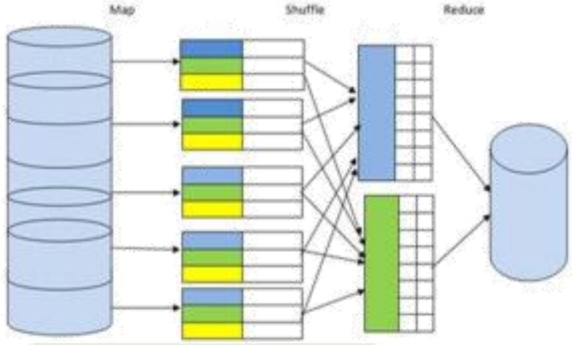

Fig. 4 Architecture of Map Reduce

Figure 4 shows the Map Reduce Architecture and Working. It always manages to allocate a local data block to a slave node. If the effort fails, the scheduler will assign a rack-local or random data block to the slave node instead of local data block. When map() function complete its task, the runtime system gather all intermediate pairs and launches a set of condense tasks to produce the final output. Large scale data processing is a difficult task, managing hundreds or thousands of processors and managing parallelization and distributed environments makes is more difficult. Map Reduce provides solution to the mentioned issues, as is supports distributed and parallel I/O scheduling, it is fault tolerant and supports scalability and it has inbuilt processes for status and monitoring of heterogeneous and large datasets as in Big Data. It is way of approaching and solving a given problem. Using Map Reduce framework the efficiency and the time to retrieve the data is quite manageable. To address the volume aspect, new techniques have been proposed to enable parallel processing using Map Reduce framework. Data aware caching (Dache) framework that made slight change to the original map reduce programming model and framework to enhance processing for big data applications using the map reduce model.

The advantage of map reduce is a large variety of problems are easily expressible as Map reduce computations and cluster of machines handle thousands of nodes and fault-tolerance. The disadvantage of map reduce is Real-time processing, not always very easy to implement, shuffling of data, batch processing

Map Reduce Components:

1) Name Node: manages HDFS metadata, doesn't deal with files directly.
2) Data Node: stores blocks of HDFS—default replication level for each block: 3.
3) Job Tracker: schedules, allocates and monitors job execution on slaves—Task Trackers.
4) Task Tracker: runs Map Reduce operations

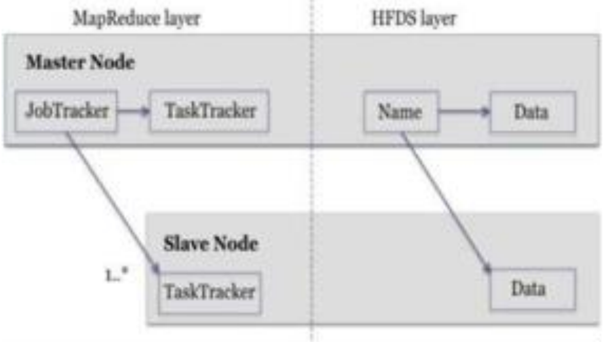

Fig. 5 Map reduce is based on the Maser-Slave architecture

D. Hive

Hive is a distributed agent platform, a decentralized system for building applications by networking local system resources. Apache Hive data warehousing component, an element of cloud-based Hadoop ecosystem which offers a query language called HiveQL that translates SQL-like queries into Map Reduce jobs automatically. Applications of apache hive are SQL, oracle, IBM DB2. Architecture is divided into Map-Reduce-oriented execution, Meta data information for data storage, and an execution part that receives a query from user or applications for execution.

The advantage of hive is more secure and implementations are good and well tuned. The disadvantage of hive is only for ad hoc queries and performance is less as compared to pig.

E. HPCC:

HPCC is an open source platform used for computing and that provides the service for handling of massive big data workflow. HPCC data model is defined by the user end according to the requirements. HPCC system is proposed and then further designed to manage the most complex and data-intensive analytical related problems. HPCC system is a single platform having a single architecture and a single programming language used for the data simulation. HPCC system was designed to analyze the gigantic amount of data for the purpose of solving complex problem of big data. HPCC system is based on enterprise control language which has the declarative and on-procedural nature programming language the main components of HPCC are:

1) *HPCC Data Refinery:* Use parallel ETL enginemostly.
2) *HPCC Data Delivery:* It is massively based on structured query engine used.

Enterprise Control Language distributes the workload between the nodes in appropriate even load.

IV. Conclusion

In this paper we have surveyed various technologies to handle the big data and there architectures. In this paper we have also discussed the challenges of Big data (volume, variety, velocity, value, veracity) and various advantages and a disadvantage of these technologies. This paper discussed an architecture using Hadoop HDFS distributed data storage, real-time NoSQL databases, and MapReduce distributed data processing over a cluster of commodity servers. The main goal of our paper was to make a survey of various big data handling techniques those handle a massive amount of data from different sources and improves overall performance of systems.

References

[1] Dhole Poonam B, Gunjal Baisa L, “Survey Paper on Traditional Hadoop and Pipelined Map Reduce” International Journal of Computational Engineering Research||Vol, 03||Issue, 12||
[2] Nilam Kadale, U. A. Mande, “Survey of Task Scheduling Method for Mapreduce Framework in Hadoop” International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA 2nd National Conference on Innovative Paradigms in Engineering & Technology (NCIPET 2013) – www.ijais.org
[3] Suman Arora, Dr.Madhu Goel, “Survey Paper on Scheduling in Hadoop” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014
[4] Wang, F. et al. Hadoop High Availability through Metadata Replication. ACM (2009).
[5] B.Thirumala Rao, Dr. L.S.S.Reddy, “Survey on Improved Scheduling in Hadoop MapReduce in Cloud Environments”, International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011
[6] Amogh Pramod Kulkarni, Mahesh Khandewal, “Survey on Hadoop and Introduction to YARN”, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)
[7] Vishal S Patil, Pravin D. Soni, “HADOOP SKELETON & FAULT TOLERANCE IN HADOOP CLUSTERS”, International Journal of Application or Innovation in Engineering & Management (IJAIEM)Volume 2, Issue 2, February 2013 ISSN 2319 - 4847
[8] Sanjay Rathe, “Big Data and Hadoop with components like Flume, Pig, Hive and Jaql” International Conference on Cloud, Bigs