International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 7, 2017 ISSN: 2348-6600
Reference ID: IJCS-188 PAGE NO: 1145-1149
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

A Survey on Technologies to Handle Big Data

[1]Mrs.K.Veena, M.C.A., M.Phil (Research Scholar)
[2]Dr.V.Sasirekha, M.C.A., M.Phil., Ph.D., (Supervisor)
[12]Assistant Professor, Department of Computer Science,
[12]J.K.K.Nataraja College of Arts and Science, Kumarapalayam, Namakkal Dt, Tamilnadu, India.
E-Mail: [1]veenabharathi44@gmail.com [2]sasirekhailangkumaran@gmail.com
Mobile: [1] (+91)9443520800, [2] (+91)9842522271

ABSTRACT- Big Data the term narrates itself as the data that is very big pointing towards large scale and multifaceted data processed using specific analytical methods. Big Data opens its eyes when traditional Relational Database tools and packages were unable to handle such huge data sets in terms of size, speed, source, storage and range of data. It focuses on various techniques to collect, process, analyses and to picture potentially large sets of data within a stipulated timeframe. The challenging task of Big Data analytics paves way to Distributed File System which is flexible, scalable and fault tolerant. Hence the architecture steps up from Centralized to Distributed architecture. The new technologies used to handle such massive data are Hadoop, MapReduce, NoSQL, Apache Hive, Pig, MongoDB etc. The researchers on Big Data seek to gather massive data in Social Media like Twitter, Facebook and News Services, Web Logs, images and videos, documents and PDFs, biometric data, government agency sources and human generated data. Because of complex in computations due to large data sets in areas like complex physics simulations, meteorology, genomics, biological and environmental research, the scientists moved on to Big Data techniques. For processing, Big Data needs HDFS(Hadoop Distributed File System) make use of MapReduce program, that relies on cluster computers which process massive amount of data in parallel and generates large amount of intermediate data. Big Data is used not only in IT Industries and Business but also it roots itself in the field of Healthcare, Social Media, Government, Transportation, and Education and Research. This paper focuses on analysing various technologies that are catered by several researchers to deal with heterogeneous data.

Keywords: Big Data, Hadoop, MapReduce, Social Media, Apache Hive, Twitter, Research

.

I Introduction

With a pace increasing in the growth of technological development and increasing in the

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 7, 2017 ISSN: 2348-6600
Reference ID: IJCS-188 PAGE NO: 1145-1149
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

options for the creating the data in all fields in heterogeneous ways, the size storage, flow, type, range of data is becoming unexpectedly increasing every second. This paves a broad way for new and different technologies to manage such enormous and different data. Big data seems to be small two words but it is the largest buzz word in the present IT domain. As Internet population is growing second to second, it is the necessity for the IT world to invent new and significant technologies and tools to handle the Big Data world which challenges the rest of the fields in front because of various types of data generated by various companies, IT, social media data made by public. Big Data is composed of three types of data namely Structured, Semi-Structured and Unstructured data. Structured data is organised in a predefined format in fixed fields. Human generated unstructured data like text documents, logs, survey results, e-mails, Social media data from YouTube, Facebook, Twitter, LinkedIn, Flickr, Mobile data, Web site contents. As traditional RDBMS is unable to handle such large varieties of data, a bundle of five parameters Volume, Variety, Velocity, Value and Veracity was evolved for the challenging Big Data Management. Volume addresses the size of the data. Data is growing rapidly from KB-MB-GB-TB-PB-EB-ZB-YB, to store such voluminous data scientists switch from traditional DBMS to advance storage systems. Data volumes are expected to grow 50 times by 2020.The massive volume of data made the scientists to backtrack and find a new paradigm to create and develop new tools for analysing it. Variety address to different types of data received for analysing from various heterogeneous data source structured, unstructured and semi structured data. Velocity addresses to the speed at which data flows. Velocity challenges most of the organisations as the data from social media, company data, bank transactions flows continuously and vigorously. Value addresses the way of exploiting the large amount of data to extract and draw valuable conclusions. Veracity address to the range of values for a data set. This veracity contains noisy and unwanted data for analysing. The unwanted dirty data is to be cleaned for further processing.

II. Big Data Technologies.

Big Data is the next frontier for advancement, competition and productivity by McKinsey & Co. The advent of Big Data not only increase the huge potential growth of individual companies revolutionizing in research and education but also effectively and gives a rapid increase in the productively and quantitative advancement and growth in entire arena of entire sectors and economic of the country.

Hadoop

Hadoop is the open source s/w written in Java, a framework for storing and procuring data in terabytes, with in cluster of nodes. Hadoop is a distributed master – slave architecture called HDFS

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 7, 2017 ISSN: 2348-6600
Reference ID: IJCS-188 PAGE NO: 1145-1149
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

(Hadoop Distributed file System).HDFS is a distributed file system designed to run on large clusters of commodity hardware based on Google File System (GFS). Master node known as Name Node and Slave nodes are known as Data Nodes. The master node divides the input into chunks of data and distribute to all slave nodes (Data Nodes) for processing. The data send signal / heartbeat to name node (Master). If the signal from any data node is stopped, the name node assigns the task to the other data node. The component which assigns task to data node is job tracker in master name node and which the nodes that that receives the task tracker in slave data node. Because of this, Hadoop is distributed, high throughput, replicates files, performs read and write large files, act on native file system, stable, highly tolerant to s/w and h/w
failures.

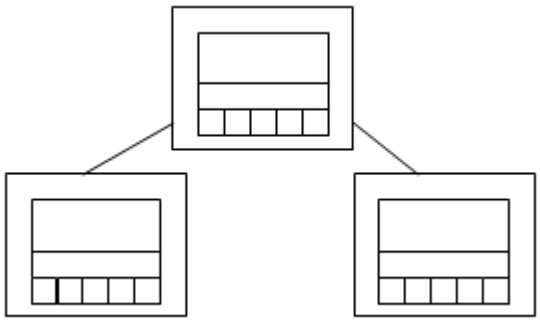

Fig.1 Hadoop Architecture

Hadoop comes in 2 Versions. Hadoop 1.0 & Hadoop 2.0

Limitations of Hadoop 1.0

Hadoop 2.0: The limitation of Hadoop 1.0 is resolved by Hadoop 2.0. Hadoop 2.0 comes with YARN (Yet Another Resource Negotiator) which is responsible for resource management, job scheduling and monitoring. With YARN, to run an application, client request to launch an application manager process from resource manager to find a node manager. The node manager launches a base to execute an application process.

The Hadoop core components are HDFS and Map Reduce and the ecosystem consist of Hive, Pig, Sqoop, Hbare, Oozie, Flume, Hahout, Avro, Chukwa, Zookeeper. Ambari is a web-based tool for Provisioning, Managing and Monitoring Hadoop Cluster. Pig is a high level data processing system used for data analysis that occurs in high level language. It was developed as a research project at Yahoo. Hive is a data warehousing tool, used to query structured data built on the top of Hadoop. Facebook created Hive component to manage their ever-growing volumes of log data. Sqoop is a command line interface application used to transfer data between Hadoop framework and relational databases. HBase is a no SQL database of Hadoop use column-oriented db used to store millions and billions of data provides random read / write operations. Chukwa is an analysis framework used to process and analyze, a highly flexible tool that makes large amount of log analysis, processing and monitoring. Zookeeper is a centralized service for maintaining configuration

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 7, 2017 ISSN: 2348-6600
Reference ID: IJCS-188 PAGE NO: 1145-1149
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

information, naming, providing distributed synchronization, and providing group services.

MapReduce.

MapReduce is a framework used for data processing of massive data sets in parallel on large clusters of commodity hardware. It supports distributed and parallel I/O scheduling, fault tolerant and support scalability . It has inbuilt processes for status and monitoring of heterogeneous and large datasets as in Big Data. The MapReduce model consists of two functions Map and Reduce.

Mapper: The Map task takes care of loading, parsing, transforming and filtering. The map task consists of four phases RecordReader, Mapper, Combiner and Pratitioner. RecordReader converts a byte-oriented view of the input into record-oriented view with key-value pairs and present it to Mapper task. The Map function generates zero or more intermediate key-value pairs. Combiner is an optional function or local reducer applies user specific aggregate function on the key-value pairs to only to that mapper. Partitioner splits the intermediate key-value pairs into shards and send then to the particular reducer i.e. key with same value goes to the same reducer.

Scalability: It is highly scalable because of its ability to store as well as distribute large sets across plenty of servers which can operate in parallel. Flexibility: It accesses various new sources data and operates on different types of data and generates values from all the data that can be accessed. Speed: As MapReduce is located in the same servers, it allows for faster processing of data. It takes minutes and hours to process terabytes and petabytes of data respectively. Security and Authentication: It works with HBase security that allows only approved users to operate on data stored in the system. Parallel Processing: It divides tasks and allows executing in parallel for multiple processors in less time. Availability and Resilient: If a node fails it copies the data from the available node and it has the ability of automatic recovery solution.

III. Apache Spark

Spark is a next generation paradigm for big data processing developed by researchers at the University of California at Berkeley. It is an alternative to Hadoop which resolves the disk I/O limitations and improve the performance of earlier systems. The major feature is its ability to perform in-memory computations, allows data to be cached in memory eliminates the disk overhead limitation for iterative tasks. Spark supports large scale data processing and it is tested up to 100x faster than Hadoop MapReduce when the data fit in the memory. It can run Hadoop Yarn manager and read data from HDFS.

IV. Conclusion

The term Big Data is coined to explain the age of fast growing of voluminous data which we

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 7, 2017 ISSN: 2348-6600
Reference ID: IJCS-188 PAGE NO: 1145-1149
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

are creating. In this paper we have described the concept, several details of technologies which assist Big Data, the working of HDFS, MapReduce, its advantages and disadvantages. This paper provides the readers with a comprehensive review of different platforms which can potentially aid them in making the right decisions in choosing the platforms based on their data/computational requirements. The future work involves analysis and investigation of algorithms and specific factors like amount of hard disk, memory and the speed required for optimally running the applications. Hope this survey will be beneficial for further progress and enhancement of Big Data Analytics in various research perspectives.

References:

[1] Kogent Learning Solutions, Bill Franks, Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman, Joris Meys, Andrie de Vries, Mark Gardener, Dr. Murray Logan, Michael J. Crawley, Deborah J. Rumsey, Johannes Ledolter, Stephane Tuffery, Dean Abbott, Wrox Certified Big Data Analyst(WCBDA), "Introducing Big Data Analytics and Predictive Modeling", Wiley Publishers.

[2] https://jeremyronk.wordpress.com/

[3]Yuri Demchenko "The Big Data Architecture Framework (BDAF)" Outcome of the Brainstroming Session at the University of Amsterdam 17 July 2013.