The Novel Approaches on Big data and the Hadoop Environment

R. Ganeshkumar[1], Dr. A. Marimuthu[2]
M.Phil - Research Scholar [1]
ganeshkmphil@gmail.com
Associate Professor & Head [2]
mmuthu2005@gmail.com
PG & Research Department of Computer Science
Government Arts College (Autonomous)
Bharathiyar University

Abstract -- **The term big data is very large data sets that may be analyzed computationally to reveal patterns, trends and associations especially relating to human activities and communications. These large data sets processed its size (volume), complexity (variety), speed of data (velocity), potential value (value), noises (veracity) make them difficult to capture, manage, analyzed and increasing its processing speed. To analyze this huge amount of data Hadoop framework can be used. Hadoop is an open-source software for distributed storage and distributed processing of very large data sets on computer clusters built from product hardware. The heart of Apache hadoop consist of a storage part is called HDFS (Hadoop Distributed File System) and processing part called Map Reduce. The technologies used by big data application to handle the enormous data are hadoop framework, Map Reduce, HDFS, HPCC and Apache Hive. These technologies handle large amount of data in GB (Giga Byte),TB (Tera Byte),PB (Peta Byte),EB(Exa Byte),ZB (Zeta Byte),YB (Yotta Byte),BB (Bronto Byte) and GB (Geop Byte).**

Key Words - **Big data, Hadoop framework, HDFS, HPCC, Map Reduce, Hadoop eco system**

I. INTRODUCTION

Big data is a term for data sets that are so large or difficult that outdated data processing applications are inadequate to deal with them. Big data is a developing term that describes any huge amount of structured, semi structured, unstructured data that has the possible to be mined for information. The need of Big data generated from the large companies like Google, yahoo, face book and YouTube etc[1] for the purpose of analysis from massive amount of data also Google contains the huge amount of information.

The term big data is describes the large volume of data set whose size, complexity and rate of growth make them difficult to captured, managed, processed by relation as databases.

There are various technologies to handle big data such as Amazon, IBM, and Microsoft etc.

A. TYPES OF DATA

TABLE I
THE TYPES OF DATA IN BIG DATA

S. No	Type	Example
1	Structured data	Relational data
2	Unstructured data	Word, PDF, Text, Media Logs
3	Semi Structured data	XML data

B. CHARACTERISTICS OF BIG DATA

Big data can be described by the following characteristics [2]

1. Volume:
The quantity of generated and stored data.It represents the size of the data how the data is large.Ex: Facebook generating 500+ terabytes of data per day.

2. Variety:
The nature and type of the data. This helps people who analyze it to effectively use the resulting insight. The files come from various formats and any type, such as text, audio, video & log files and sensor data.

3. Velocity:
Velocity represents the speed at which the data is generated and processed to meet the demands and challenges that lies in the path of growth as a development. Ex: Analyzing 2 million records each day to identify the reason for losses.

4. Variability:
The potential value of big data. Inconsistency of the data set can hamper processes is handle and manage it.

5. Veracity:
Veracity is represents accuracy of data. It also refers to noise, biases, when dealing with high volume, velocity and variety of data.

C. APPLICATIONS OF BIG DATA

The following figure gives the applications of big data.

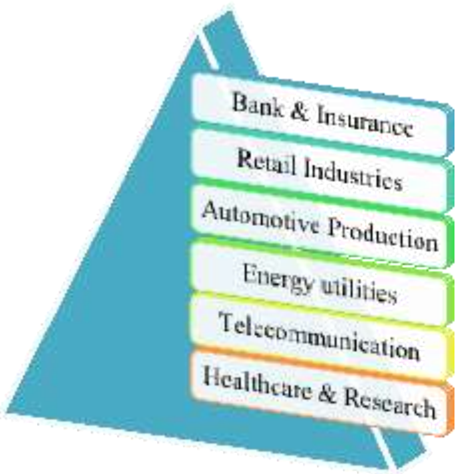

Fig 1 – The Structure of Big Data Applications

1. Bank & Insurance – The use of customer data always raises privacy issues [6]. Fraud detection has also been improved. The massive data from digital channels and social media, real-time checking of claims throughout the claims cycle has been used to provide visions.

2. Retail Industries – Enhanced staffing through data from shopping patterns, local events, and so on. Exactly the frauds are reduced. Timely data study of inventory.

3. Automotive Production – The automotive industry continues to face a growing number of challenges and forces. Cost pressure, competition, globalization, market shifts, and instability are all increasing. Big data and analytics today propose before unthinkable opportunities for tackling these and many other challenges automakers face.

4. Energy Utilities – Smart meter readers allow data to be collected almost every 15 minutes as different to once a day with the old meter readers. This rough data is being used to analyze consumption of utilities better which allows for better customer feedback and improved control of utilities use. In utility companies the use of big data also allows for better asset and workforce organization which is useful for recognizing errors and correcting them as soon as possible before complete disappointment is experienced.

5. Telecommunication – In telecommunication field the big data plays a good role. Service providers are trying to complete in the cutthroat world of telecom facilities. Where mare and more subscribers rely on over the top players as providers of value added facilities are focused on increasing revenue, reducing opex, chum and enhancing the customer experience as key business objectives.

6. Healthcare & Research – Some hospitals are using data collected from a cell phone app, from millions of patients, to allow doctors to use evidence-based medicine as opposed to handling numerous medical/lab tests to all patients who go to the hospital. A battery of tests can be resourceful but they can also be costly and usually in real. Free public health data and Google Maps have been used to make visual data that allows for faster identification and resourceful analysis of healthcare information, used in following the spread of long-lasting disease.

II. HADOOP: THE WAY OF SOLVING ANALYSIS OF BIG DATA [3]

Everyday a large amount of unstructured data is getting dumped into our machines. The major challenges are not to store huge data sets in our systems, but to recover and analyze the big data in the organizations.

Hadoop has the ability to analyze the data present in changed machines at changed locations very quickly and in a very cost effective way. It uses map reduce concept which enables it to divide the request into small parts and process them in parallel.

Hadoop is open-source software that allows to store and process big data in a distributed setting across dusters of computers using simple programming models. It is considered to scale up from single servers to thousands of machines, each offering local computation and storage.

The heartbeat of hadoop is Hadoop Distributed File System and Map reduce. Hadoop Distributed File System using storage part in hadoop. Map reduce using processing part in hadoop.

A. APACHE HADOOP ECO SYSTEM

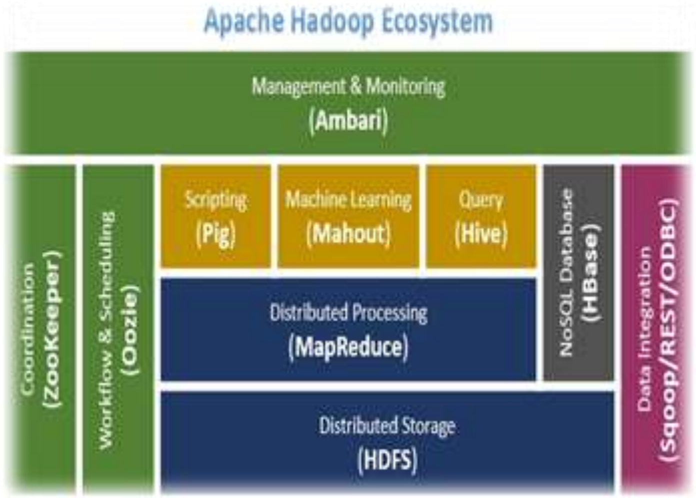

Fig 2 - The Architecture Of Hadoop Eco System

B. HADOOP COMPONENTS

Hadoop Eco System Comprises Four Core Components [4]

1. Hadoop Common:

Apache Foundation has pre-defined set of advantages and collections that can be used by other modules within the Hadoop eco system.

2. HDFS: Hadoop Distributed File system:

The default big data storage layer for Apache Hadoop is HDFS.HDFS is a file system designed for storing very huge files with streaming data access pattern, running clusters commodity hardware. HDFS manages storage on the cluster by flouting incoming files into pieces called blocks and storing each blocks are redundantly across the pool of the server.

3. Map Reduce:

Map Reduce is the java-based system. Map Reduce is a programming framework for distributed computing to break the large complex data into small units and process using divide and conquer method [5].

Map Reduce includes two stages:

1. Map (): In this map consist the inputs are divide into smaller sub parts and given to the worker nodes that is masternode.A worker node received that inputs and again leads to the multi- level free structure.

2. Reduce (): The master node saves the answers from all the sub problems and combines them together to from the outputs [5].

4. YARN:

YARN is great enabler for dynamic resource utilization on hadoop framework as users can run various hadoop applications without having to difficulty about increasing workloads.

C. SUB COMPONENTS

1. AMBARI:

Ambari provides step-by-step wizard for installinghadoop eco system services. A hadoop component, Ambari is a RestfulApl which provides easy to use web user interface for hadoop management.

2. MAHOUT:

Mahout is an important hadoop component for machine learning. This provides execution of various machine learning algorithms. Mahout is divided into four important groups: collective filtering, classification,grouping and mining of parallel frequent patterns.

D. DATA ACCESS COMPONENTS

1. PIG:

Apache Pig isanalyzing huge data sets efficiently and easily.Pig can process half dozen lines of code with tera bytes of data.

2. HIVE:

Hive is a data warehousing framework. It is on top of hadoop. It allows writing SQL queries to analyze and process the bigdata stored in HDFS.Hive makes querying faster through indexing.

E. DATA INTEGRATION COMPONENTS

1. SQOOP:

Scoop component is used for importing data from external sources into related hadoop components. It can also be used for exporting data from hadoop other external structured data stores.Sqoop parallelized data transfer, mitigates unnecessary loads, allows data imports, efficient data analysis and copies data quickly.

2. FLUME:

Flume component is used to aggregateand gather large amounts of data. Apache flume is used for gathering data from its origin and sending it back to the resting location (HDFS).Flume consist of three primary structures that are channels, sources and sinks.

F. DATA STORAGE COMPONENT

1. HBase:

HBase is a column-oriented database that uses HDFS for storage of data.HBase supports batch computations using map reduce and also random reads. With HBase No-SQL database enterprise can create large tables with millions of row and columns on hardware machine.

G. MONITORING, MANAGEMENT AND ORCHESTRATION COMPONENTS

1. OOZIE:

Oozie is a workflow scheduler where the workflows are communicated as Directed Acyclic Graphs. The workflows in Oozie are executed based on data and time dependencies.

2. ZOOKEEPER:

Zookeeper is the king of coordination and provides simple, fast, reliable and ordered functioning services for a hadoop cluster. Zookeeper is responsible for organization service, distributed configuration facility and for providing a naming archive for distributed systems.

III. HPCC (HIGH PERFORMANCE COMPUTING CLUSTERS):

HPCC is an open source computing platform and provide the services for management of big data workflow. HPCC is user designed data model. HPCC is managed most complex and data-intensive analytical problems. HPCC system was built to analyze the big volume data for the purpose of solving difficult problem. The followingcomponents are heart of HPCC.

1. HPCC data refinery: Massively parallel ETL engine.

2. HPCC data delivery: Massively structured query engine.

3. Enterprise control language distributes the workload between the nodes.

IV. CONCLUSIONS

In this survey entered the big data processes and hadoop environment. Big data solving lot of problems in future using hadoop and spark. Big data applications the data will be stored and analyzed for future enhancement. The paper also focuses on big data processing problems. The technical languages are using to process the big data. In the future of big data process will be solved many of the open-source software's.

V. REFERENCES

[1] Varsha B.Bobade, "Survey paper on Big data and Hadoop", IRJET, Jan,2016

[2] Hilbert, Martin, "Big data for development: A review of promises and challenges Development policy review", martinhilbert.net, Retrieved 2015

[3] HadoopTutorial: http://developer.yahoo.com/hadoop/tutorial/module1.html

[4] Hadoop Tutorial: https://www.dezyre.com/article/hadoop-components-and-architecture-big-data-and-hadoop-training/

[5] shilpaManjitKaur," BIG Data and Methodology- A review" ,International Journal of Advanced Research in

Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.

[6] Sravanthi, Subba Reddy," Applications of Big data in Various Fields", IJCSIT, 2015.

[7] Jean-Pierre Dijcks, "Oracle: Big Data for the Enterprise", 2013.

[8] DeWitt &Stonebraker, "MapReduce: A major step backwards", 2008.

[9] Dean, J. and Ghemawat, S., "MapReduce: a flexible data processing tool", ACM 2010.

[10] J. Dean and S. Ghemawat,"MapReduce: Simplified Data Processing on Large Clusters", p.10, (2004).

[11] Andrew Pavlo, "A Comparison of Approaches to Large-Scale Data Analysis", SIGMOD, 2009.

[12] Apache Hadoop: http://Hadoop.apache.org

[13] Hadoop Distributed File System, http://hadoop.apache.org/hdfs