Journal of Advances and Scholarly Researches in Allied Education
Vol. 13, Issue No. 1, April-2017, ISSN 2230-7540

Big Data and Hadoop Ecosystem: A Brief Analysis

Ruchi Sawhney[1]* Prof. (Dr.) K. P. Yadav[2]

[1] Research Scholar, Department of Computer Science, Himalayan University, Itanagar, Arunachal Pradesh

[2] IIMT College of Engineering, Greater Noida

Abstract – Since Last few years Data in internet is growing tremendously which is coined as Big Data. This Big Data includes data collected from different sources and in different forms. In this paper different Data Forms, Characteristics and Pipeline Analysis of Big Data is explained. Our objective is to explain 6V's of Big Data and Big Data Model. For Handling this big data there are different software's are available in market. Among of all Hadoop is one of the popular Software. It is open source software used to store big data of large dataset across a commodity of clustered servers. Concept of HDFS and Map Reduce is being illustrated. Apart of that Hadoop Ecosystem with its supporting Apache software's have been mentioned with their characteristics.

Keywords : Big Data, Structured data, Semi Structured, Unstructured data, 6 V's Volume, Velocity, Variety, Value, Veracity, Valence, HDFS, Map Reduce, Hadoop Ecosystem.

----------X----------

INTRODUCTION

Electronic World around us is having large amount of data. We are in the world of internet where daily large amount of data get generated. During Past many years the internet plays a vital role for data generation. Several changes are happening in the field of cloud computing, Big Data, Mobility and Internet Of Things. This data is getting generated from feedbacks, Facebook, Twitters, You tube, Google, ATM's, Drop box, Picasa etc. The data which get collected from different-different sources we store them and utilize it for different type of Analysis. Which will help us for many different type of surveys. The Data is in different variety like Audio, Video, Pictures, Text etc. It's of massive volumes that's why called Big Data.

BIG DATA

The volume of data is too big/large or complex that can't be handled by traditional ways of processing application and as we know the growth of data is increasing drastically. So, it will become more difficult for us to handle it. It's a new data challenge that requires leveraging existing system differently. Few decades back the conventions of data units are in Kilobytes, Megabytes and now a days they are in Gigabytes, terabytes, Petabytes, Exabytes and Zettabytes and as data grows these units soon becomes smaller.

Data is stored in various servers which are not even known by anyone. That data can be utilized for several surveys. So, we can make out something from that also. Careful analysis of this type data can transform this data into information and information into insight.

A Pool of large sized dataset to store, capture, transfer, search, analyze and visualize related information or data within an acceptable elapsed time.[1]

Data = Information

Information= Insight

Definition

Big Data means a collection of a Structural, Semi structural and Unstructural data. It's qualitative in nature.

Big Data is the concept and can be defined in various definition.

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Challenges include capture, storage, analysis data curation, search, sharing, transfer, visualization, querying, updating and information privacy.[2]

Big Data is the term which defines the hi-tech, high speed, high Volume, complex and multivariate data to capture, store, distribute, manage and analyze the information (TechAmerica Foundation, 2014)

Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enables enchanced decision making, insight discovery and process optimization (Gartner,2014 : Gursakal, 2014).

Data Forms:

1. **Structured Data**: Structured Data that means has predefined format and resides in fixed length for big data. The data which usually stored in Database.

2. **Semi- Structured Data:** Semi- Structured data is not in structured way to store in database but has tags or other markers to separate semantic elements and enforce hierarchies of records and fields with in the data.

3. **Unstructured Data:** On web/Internet this type of data we get which in unstructured that is can't be accommodate in any type of database. Extremely large dataset which is difficult to analyze with traditional tools. It can be data, number, facts etc.

Characteristics of Big Data:

As 3v's (Volume, Variety, Velocity) get extended to 4v (value added) then 5v (v for veracity) and now 6v (another v for valence).

1. **Volume**: This is in context with huge capacity. It can include any kind of data. Either data generated from communication or including the data is generated from all the connected.

2. **Velocity**: Means speed of data processing. It could be incoming data or stream outgoing data.

3. **Variety:** As we know big data is consisting of different types of data. So there is variety in data. Data forms like structured (database system), Semi structured (tags, web logs) or unstructured (audio, image, video)

4. **Value:** When we talk about 4v then additional v stands for value. This value is valuable for big data and fetch data from the information stored. This helps in analyzing data and the only reason that it allows to generate useful information for business.

5. **Veracity:** Big data veracity means the abnormality, partially or distributed data. Veracity checks for quality of data.

6. **Valence :**

Valance = = connectedness

It's a concept of chemistry which is responsible for connectedness or bonding with other atoms. It have the highest energy level and in it is in outer most shell [3].

Similarly the more connected data means more valence which will again help to connect connected data. A high valence data set is denser. This makes many regular analytics critiques very inefficient.[3] Many problem comes due to dynamic behavior of the data as data keeping changing with time and volume. If there is more valence among data then the complexity of the analysis get increased.

Big data – Pipeline Analysis:

As shown in figure 1 below we are explaining accordingly.

1. **Data Acquisition and Recording:** Origin of data is from different forms. We get raw data so need to do filtering and compression but keeping one thing in mind that important information should not get filtered. Next is to generate right metadata and data fidelity.

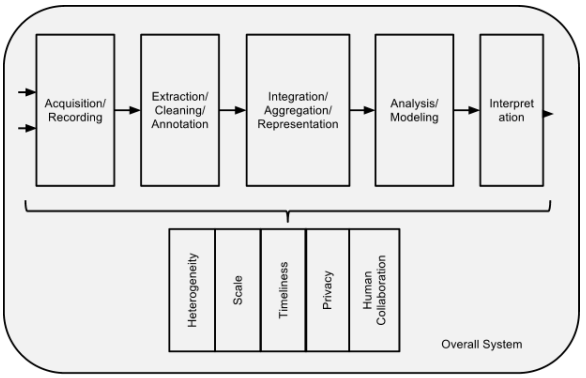


Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

1. **Information Extraction and cleaning:** First we need to transform unstructured data to analytics friendly format so that they can extract right information. Doing aforementioned process again and again with correctness being the top priority. Data cleaning techniques consists of well controlled constraints to valid data and well verified error models to ensure the quality of the data.

2. **Data Integration & Aggregation:** We get heterogeneous data from multiple sources. A set of data transformation & integration tool helps the data analyst to resolve heterogeneities in data structure & semantics this leads to integrated data which is uniform and can fit for standardization schemes and analysis needs.

3. **Query Processing, Data Modeling and Analysis:** Big Data is different from traditional data hence its query methods also. When we talk about big data, we often associate it. When we talk about Big Data, we often associate it with being dynamic, inter-relation, untrustworthy & noisy as opposed to the process of mining, which will require clean, trustworthy, integrated, efficient, accessible data which can be accessed via mining interfaces using declarative queries, scalable mining algorithms and computing environment for big data.[5]

4. **Interpretation:** It means after completing all the process we get output in the form of reports or graphs etc. This should be properly understandable by the user. Interpretation can't happen in vacuum. Usually, this involves examining all the hypothesis made and retracing the analysis.

Big Data Applications in different Areas

1. Manufacturing
2. Procurement
3. Product Development
4. Distribution
5. Price Management
6. Sales
7. Store operation
8. Human Resources
9. Merchandising

And many more….

Big Data Ecosystem:

Dealing with big data we know that it is not a simple job. So, all the association between the components and the interconnected relationship form Big Data Ecosystem which incorporate in itself all the data, supporting infrastructure, model during entire big data life cycle

Techniques: There are many techniques present which we can use in projects like Association rule learning, Data Mining, Cluster analysis, Crowd sourcing, Machine learning, Text Analytics.[5]

Due to the generation of variety of data in messy forms a new business started to utilize this electronic media. Big data is an umbrella term that encompasses the management and processing of such data.

To handle the problem of storing and processing and complex and large data, many software frameworks have been created to work on the big data problem.

The organizations that deals with high volume of data must deals with following mentioned areas.

1. Data Capture / acquisition from various sources
2. Data Massaging or curating
3. Organization and storage
4. Search, analysis and querying
5. Sharing or consumption
6. Security and Privacy

BIG DATA TECHNOLOGY

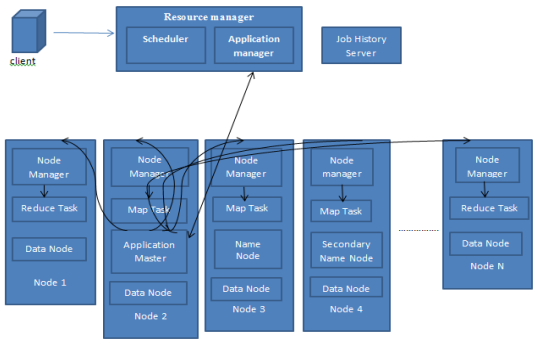
Hadoop Ecosystem

Hadoop is a software to enables the distributed processing of large dataset across a commodity of clustered servers. It manages from single server to thousands of commodity hardware machines.

Primary Components

1. **Hadoop distributed file system (HDFS)**
2. **Map Reduce Framework**
3. **Yarn**

HDFS : It is a file system that can be used to store data in a replicated/ Duplicated and distributed manner across the various nodes, which are part of Hadoop cluster. In HDFS it allows to append data only and no modification is allowed.



Map Reduce: A program that helps to provide a distributed data processing framework for large dataset.

A programming task that takes a set of data (key-value pair) and converts it into another set of data, is called Map Task. The results of map tasks are combined into one or many Reduce Tasks. Overall, this approach towards computing tasks is called Map Reduce Approach [8]

Yarn: stands for Yet Another Resource Negotiator. It is a new component included in Hadoop 2.x Architecture also known as "MR V2".

Core components of Hadoop

1. Resource Manager (RM): It manages resources of a cluster. RM consisting of Scheduler provides resource allocation and Application Manager which is responsible for client interaction like accepting job, identifying and assigning them to application master.

2. Application Master (AM): It is responsible to complete a life cycle of Map Reduce Job. Even to interact with RM for the resources required by the job.

3. Node Manager (NM): This works node side. Responsible to manage all the containers present in node will keep check on resources required for process like CPU usage and Memory and even prepare the report for the health to RM.

4. NameNode (NN): It is the Master Node which Stores the Metadata related To HDFS. It performs coordination activities among Data nodes, Stores the mapping of blocks on the data nodes. There can be only one active node to regulates access.

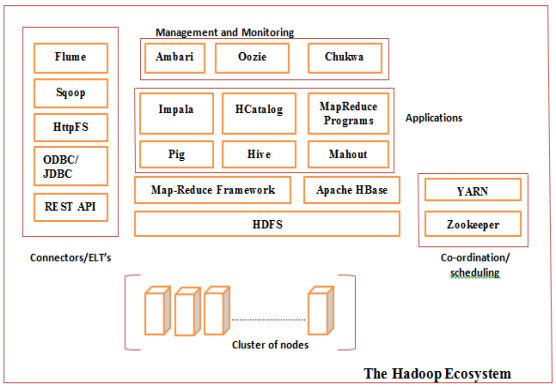
5. As a backup Hadoop introduces secondary NameNode, which constantly synchronizes with First NameNode and can be used as a backup when NameNode is not available.

6. DataNode: They are the slaves where the actual data is stored. They are deployed on all the nodes in a Hadoop cluster. The data get split into data blocks and then get stored in various data nodes in a cluster. The block size is 64 MB by default and can vary also.

Each Hadoop file block is mapped to two files in the DataNode. First one is the file block another one is the checksum of the same.

Hadoop Startup each DataNode get connected to NameNode to inform its availability to server. During runtime DataNode sends the heart beat signals in every 3 seconds. If the heartbeat is not received by NameNode in 10 minutes then it is assumed to be not available. NameNode also replicates the data blocks to other DataNode also.

Now Let we discuss about other supporting Software's of Hadoop. Below mentioned diagram shows the Hadoop Ecosystem and it's supporting Software. Here in this paper only the name of software with its basic functionality is specified. Rest in detail will be the future scope.



Apache Hbase

As shown in diagram Hbase runs on top of HDFS. Hbase is an open-source distributed, random access & non-relational database. It allow read, write and update data. It doesn't support SQL hence called NoSql database (Accept Hbase some more are NoSql database like couchDB, MangoDB and Cassandra).The data gets stored as Key Value Pair.

Apache Pig

Apache pig is on top of Map Reduce. It is used to analysis large dataset that runs on HDFS. Writing Map Reduce queries is time consuming to write. Pig is a easier way to write Map Reduce Queries. Pig is similar to python Language and allow to write short code, more efficient code. We can use scripting language known as PigLatin. Which later can be translated to Map Reduce before execution. Pig was developed at Yahoo! Research to enable developers to create Map Reduce jobs for Hadoop. Since then many big organizations are using it.

Apache Hive

It provides Data warehouse .Hive also runs on top of Apache Hadoop and use HDFS for storing its data. Hive is an alternative to Pig. Python is not understandable by most of the developers. Hence Facebook decided to create Hive. Hive is an SQL like query language called HiveQL. It accordingly converts Sql-like queries into Map Reduce job which get executed on Hadoop.

Apache Zookeeper

It's mandatory part of Hadoop Ecosystem which coordinates between or among Nodes. It maintains configuration information & the group services to the distributed system. Due to its in memory management of information it offers distributed coordination at a high speed.

Apache Mahout

It is an open source machine learning software. This is consisting of many Algorithms that enables machine learning to be performed on Hadoop. With Mahout you can perform clustering classification or filtering on you data .It is highly effective on large Dataset. Mahout Algorithms are highly optimized to run the MapReduce framework over Hadoop.

Apache Hcatalog

It provides Metadata management services on top of Hadoop. All the software which runs on Hadoop can use Hcatalog to store Schemas in HDFS. It helps to create, edit and expose (using REST APIs) table definition to any third software. It ensures that without worrying about the location and format of data it get stored.

Apache Ambari

It monitors the Apache Hadoop cluster, Hiding the complexities of the Hadoop framework. It offers many features like installation wizard, system alerts and metrics, provisioning and management of Hadoop cluster, and jobs performance. It allows administrators to allow integration with any other software.

Apache Oozie

It is a Java Web application which is used for workflow scheduler for Hadoop jobs. It can be used with MapReduce as well as pig scripts to run the jobs. It even clubs multiple jobs sequentially into one logical unit of work.

Apache chukwa it is a monitoring application which is used for Distributed Large System. This built on top of HDFS and Map Reduce.

Apache Sqoop

Apache Sqoop loads large Dataset into Hadoop cluster. It establishes connectivity between non Hadoop data source and HDFS. Use mappers to load and unload data across HDFC and a data source.

Apache Flume

It is a distributed data collection service that extracts data from the heterogeneous sources, aggregates the data and stores it into the HDFS. It is used as an ETL(Extract-Transform-Load) [8]

Comparison between Map Reduce, Pig, Hive

Through the analysis of various research papers we compare different fetching techniques used in Hadoop on the basis of different attributes.

Attributes	Map Reduce	Pig	Hive
Language	Algorithms of Map and Reduce Functions	Scripting Language named as Pig Latin	SQL like Language named as Hive QL.
Size	More lines of code.	Fewer	Fewer than pig and Map Reduce because of using SQL like language.
Development Time	More time consuming	Rapid development	Rapid development
Data Forms	Can deal with Structured, Semi Structured and Unstructured Data	Can handle Structured, Semi Structured and Unstructured Data	Deals mostly with structured and semi structured data
Joins	Hard to attain join functionality	Joins can be easily written	Easy for joins

CONCLUSION

This paper propounds the concept of big data, different data forms, characteristics, its applications areas and some techniques to handle big data such as Hadoop. Hadoop is open source software used to manage big data. In this paper we also proposed Primary components of Hadoop (HDFS, Map Reduce Yarn techniques) and its core components with its working, Hadoop ecosystem with its supporting software and comparison between different data fetching techniques. As future work direct this research towards various fetching techniques. A deep investigation of techniques to optimize fetching technique will be tried.

ACKNOWLEDGEMENT

My sincere thanks to my honorable guide Dr. K.P. Yadav who has contributed towards the preparation of this paper.

REFERENCES

1. International journal of emerging trends in engineering research (IJETER- held on 2015, by Purti Jain1 , Dr. Shruti kohli.

2. https://en.wikipedia.org/wiki/big_data

3. https://www.coursera.org/learn/big-dataintroduction/lecture/qlrsf/characteristics-of-big-data-valence

4. by h.v. jagadish, johannes gehrke, alexandros labrinidis, yannis papakonstantinou, jignesh m. patel, raghu ramakrishnan, and cyrus shahabi "exploring the inherent technical challenges in realizing the potential of big data." communications of the acm | july 2014

5. Tanvi ahlawat and dr. radha krishna rambola " literature review on big data" – ijeter , 2016

6. Sabitha M. S., Dr. Vijayalakshmi, R. M. Rathikaa S.R.E. (2015). "Big data – literature survey" (IJRASET) Dec 2015 Hrishikesh Vijay Karambelkar "scaling big data with hadoop and solr "

7. A Comprehensive Survey on Big Data Issues and Alternative Approaches to Hadoop MapReduce Gauri S. Rapate1 , Nandita Yambem2.

8. A Review Paper on Big Data and Hadoop Harshawardhan S. Bhosale1 , Prof. Devendra P. Gadekar

9. Review of Hadoop Performance Optimization Dongyu Feng, Ligu Zhu, Lei Zhang.

10. A Survey on Big Data privacy using Hadoop Architecture Priyank Jain1 , Manasi Gyanchandani , Nilay Khare , Dhirendra Pratap Singh, Lokini Rajesh

11. http://hadoop.apache.org/hdfs

12. Apache Hive. Available at http://hive.apache.org

13. Sagiroglu, S.Sinanc, D.,‖Big Data: A Review‖,2013, 20-24

14. Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule (2014). Survey Paper on Big Data‖ International Journal of Computer Science and Information Technologies.

15. 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||

16. https://www.researchgate.net/publication/282281171_big_data_challenges

Corresponding Author

Ruchi Sawhney*

Research Scholar, Department of Computer Science, Himalayan University, Itanagar, Arunachal Pradesh

ruchidakshsawhney@gmail.com