A Survey on Hadoop Architecture & its Ecosystem to Process Big Data -Real World Hadoop Use Cases

Mr.Nagarjuna D N[1], Prof.Yogesh N[2]
[1]M.Tech (CNE) Scholar, Dept. of ISE, Acharya Institute of Technology, Bengaluru
[2]Assistant Professor, Dept. of ISE, Acharya Institute of Technology, Bengaluru

ABSTRACT

As in the present day context, growing with Big Data derive the importance of analyzing huge amount of data. The data can be frequent and speedy of increment and change in database and warehouses leads to Big Data. Due to dissimilar types of data, it needs a specialized merging system. So Apache foundation had come up with an open source tool called *Apache Hadoop* which deals with variety of data. In this paper, we have discussing the characteristic description about the Big Data and Apache Hadoop framework. It also discusses the challenges which are facing with Hadoop and their use cases. This paper gives brief idea for the beginners who are eager to know about the Big Data and Hadoop Framework and there use case applications.

Keywords *- Big Data, DataNode, HDFS, JobTracker, MapReduce, NameNode, TaskTracker.*

I. INTRODUCTION

Big Data is a Technology in which many vendors are thinking that it's an emerging technology. In a present day context, many business people were talking and moving towards a Big Data technology for the purpose of analysis and to make revenue out of that [1]. In order to provide a long term challenges, Big Data boost in new ways to help analytics, organizations, industries and even society itself.

Big Data means large and complex in nature of vast amount of data. They constitute structured, unstructured and semi-structured data. To process Big Data, software tools which are commonly used to constitute of forward-thinking about analytics condition such as predictive analytics, sensitive analytics and data mining etc. But the unstructured data which is used in Big Data analytics may not fit in to a data warehouse.

Furthermore, traditional data warehouses may be difficult to handles, process and analyze data demands posed by Big Data.
Conventional definition of Big Data can be characterized in main three different V's as shown in Fig.1:
1. Volume
2. Velocity
3. Variety
4. Variability/Veracity
5. Complexity

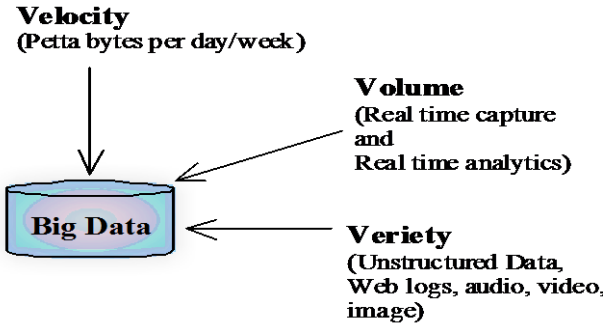

Figure 1: Big Data based on Doug Cutting 3V's Model

1. Volume

Big Data can be a larger in volume and can be measured in Tera/Peta bytes. If your old data is stored and new data is arrived at faster rate is very difficult to handle it, which is called a Big Data. Think that data can be kept on growing. Many factors bring to the sudden increase in data size. In the past years, this excessive data brings storage problem. In order to deal with such a large data, efficient software tool is used.

2. Velocity

Velocity or speed refers to how much fast the data is going to be generated, but also how well you can analyze and make use of it[2]. Velocity of data can be stream processing or batch processing. Many business processes that require real-time data analysis, this leads to a velocity challenge. Your advisor may need to pass suggestions business process changes to handle today's high-speed data. Velocity of data can be real time capture and real time data analytics.

3. Variety

Variety points to the data from different vendors applications to your databases. That might be documents, embedded sensor data, video uploads, phone conversations, social media, RFID, Satellite data and much more[3]. Data in the today era will be in all types of formats: Unstructured data, audio, video, images and numeric data in traditional databases. By today's estimation the 80% to 90% of data is unstructured and remaining is structured and semi-structured.

4. Variability

Data flows can be highly not consistent with periodic peaks. It's because of increase in variety and velocity of data, but also with unstructured data involved. Shortly we can say this as data in doubt.

5. Complexity

In present, the data generated from multiple sources. And it is still involved in linking, matching, cleaning and transforming data across systems[4]. However, it is essential to join and correlate their relationships, multiple data linkages and hierarchies.

II. BACKGROUND

Big Data incorporates different kinds of data. In present day context, data comes in all form traditional databases to hierarchical data stores created by end users, to text documents, email, meter-collected data, video, audio and financial transactions.

1. Structured Data

Any data which has standard defined format. Data is mainly text based. Distinctively conforms to ACID property. The term structured data refers to *having definite and highly organized structure*[5]. The most common method of structured data records is a database, where specific information is stored in the form of rows and columns. Structured data is also searchable by specifying the index. Structured data can be easily understood by computers and is also efficiently read by human.

Example:

Database, Electronic Spreadsheets, Enterprise systems, Data Warehouses, census and legal records, Library Catalogues, phone book etc.

2. Semi-structured Data

Semi-structured data is one on which is neither typed data, nor raw data in a schematic database system. If it is structured data, even though it is not represented a table or object based graph, or in relational database system. Data is mainly text based. Distinctively conforms to ACID property. Semi-structured data is a type of structured data but it is not represented in a standard defined format. Neither in the form of tables nor contain tags or with separate meaningful elements and apply hierarchy of records and data fields. Consequently, it is also as self-describing structure or schema less. In object-oriented databases, one often finds semi-structured data. Semi-structured data generally found in an object-oriented info.

Example:

Web Posts, Tweets Blogs, Scientific Data, Wiki pages, Forums, Instant Messages etc.

3. Unstructured Data

Any data which has no standard defined format. The term Unstructured refers to *lacking definite structure or organization.* The most common framework used in order to process information to extract useful data uses Unstructured Information Management Architecture. The Software that creates machine-process able structure exploits the auditory, linguistic, and visual structure inherent in all forms of human communication[6]. Unstructured information can contains and other small and large-scale patterns and ambiguities. While the main content being conveyed does not have a standard defined structure, it generally comes in objects package that itself have structure and are amalgamation of structured and unstructured, but together this is still referred to as unstructured data[7].

Example:

PowerPoint, PDF, Email messages, Word Documents, Audio files, Video, RSS feeds, Graphics and Multimedia etc.

III. ARCHITECTURAL DESIGN OF HADOOP CLUSTER

Hadoop cluster is an open source framework which consists of two main components: *HDFS* and *MapReduce.*

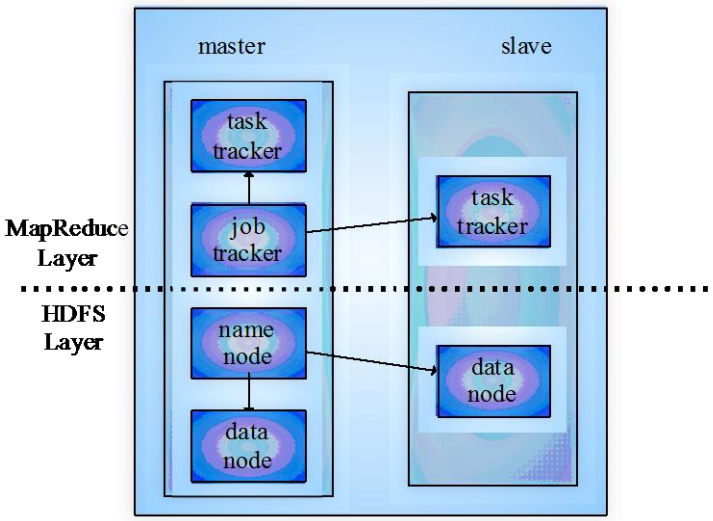

Figure 2: Master/slaves structure of Hadoop Cluster

HDFS architecture is mainly divided into three nodes are shown in Fig.3:

1. NameNode
2. DataNode
3. Secondary NameNode

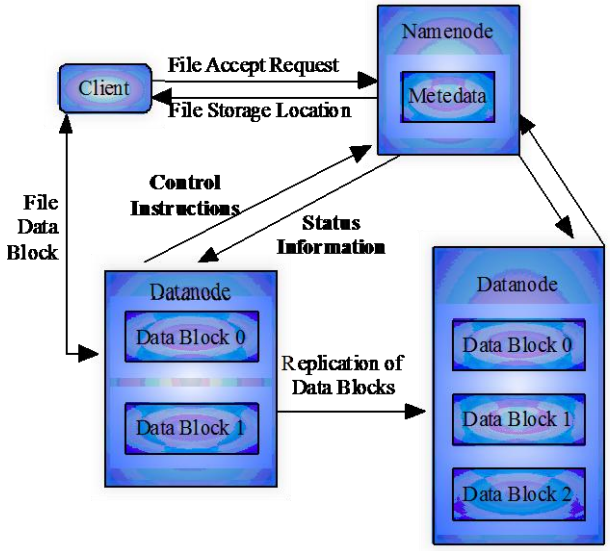

Figure 3: HDFS Architecture

1. NameNode

It's the central node which contains Meta information about Hadoop file system[9]. The important job of NameNode is that it registers all the metadata and also contains information about the position of files and blocks of data. NameNode is also generally called as *master node*, shown in Fig.2. It has contains all the replicated information about the blocks in the cluster.

2. DataNode:

It's the type of node in HDFS is DataNode. It's also known as *slave node*, shown in Fig.2. Hadoop cluster may contain one or more DataNodes, it depends on the number of systems are used in the cluster. DataNode is used to store in HDFS as blocks and for the running jobs it acts as a platform. First while starting the Data Node there must be a handshake happen with the NameNode. Each DataNode generates the block report and sends to the NameNode by an every hour. Heart beats also send to or from the NameNode for every 10 minutes to tell the NameNode that DataNode functioning correctly. If no heart beat received by NameNode at a particular timeout then NameNode decides the failure of the DataNode and it renders the replica of DataNode[9].

3. Secondary NameNode:

It's also called as HDFS *Clients/Edge node*[9]. It mainly acts as linker in between DataNodes and NameNode. In the Hadoop cluster always contains single client node. When any user want to read the application wants to read a file is contacts to the Name Node first and then gets the list of DataNodes which contains the data required by the clients. Then after client receiving the required data the NameNode contains a location of the file which is already red.

MapReduce architecture is mainly divided into two nodes are shown in Fig.4:

4. JobTracker
5. TaskTracker

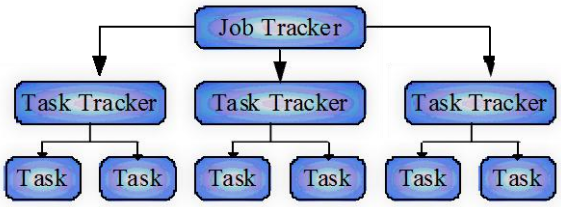

Figure 4: Task allocation in MapReduce

4. JobTracker

It's the service present in the Hadoop that helps in MapReduce tasks to specific nodes in the cluster. The nodes that contains data, it can be a rack. When the client applications went on submitting jobs to the JobTracker, the JobTracker communicates with the NameNode to check actual location of the data. The JobTracker allocates TaskTracker nodes to slots or near the data. The JobTracker assign work to the selected TaskTracker nodes. Each time the TaskTracker is monitored and if the heartbeat signals are not submitted often to the JobTracker, decides which is failed and work is assigned to other TaskTracker. The TaskTracker will notify the JobTracker whenever task fails. Then JobTracker decides further actions to be taken and resubmit the jobs. When the work is done, the JobTracker updates its status. Client applications can get results from the JobTracker.

5. TaskTracker

A TaskTracker is a node in the cluster which accepts tasks such as Map, Reduce and Shuffle operations from a JobTracker. TaskTracker are configured with a set of slots which indicates the number of task that can accept. When JobTracker looking to schedule jobs within the MapReduce operations, it first looks for empty slots on the same server or on same rack that contains the DataNode.

The TaskTracker spawns a separate JVM processes to do the actual work; this is to ensure that process failure does not take down the task tracker. The TaskTracker monitors these spawned processes, capturing the output and exit codes. When the process finishes, successfully or not, the tracker notifies the JobTracker. The TaskTracker also send out heartbeat messages to the JobTracker, usually every few minutes, to reassure the JobTracker that it is still alive. These messages also inform the JobTracker of the number of available slots, so the JobTracker can stay up to date with where in the cluster work can be delegated.

Fig. 5 shows the process of Hadoop dealing with large data sets. MapReduce model abstracts the parallel computing process on the large clusters into two functions, Map function and Reduce function.
Map function accepts a key-value pairs set as input, and outputs one or more intermediate state key-value pairs set.

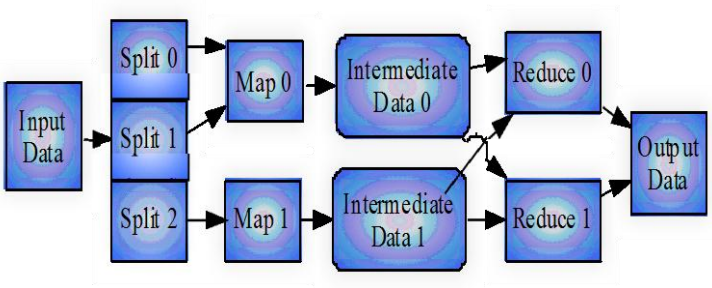

Figure 5: Data Processing in Hadoop

When a job is submitted to the MapReduce framework, MapReduce will divide it into several Map tasks and assign them to different nodes for running. Every Map task only deals with a part of the input data. After Map task processing, the results, those
value-keys that client requires (Table.1).

Table 1. MapReduce Example

Function	Input	Output	Directions
Map	(k1,v1)	(k2,v2)	The input keys (k1,v1) is mapped to the keys of an intermediate format (k2,v2) collection.
Reduce	(k2,v2)	(k3,v3)	Reduce a group of middle set values associated with k2 to a smaller set of values.

IV. COMMON HADOOPABLE USE CASES

The common Hadoopable Use cases are [10]

1. Risk Modeling

In many of the banks or other financial institutes Hadoop is a really useful way of digging deeper into their customers. They were takes in-data about their customer spending habits, their credit, repayments everything. It helps to analyze their customer needs and their comfort with the banks. It all together provides them to make more money.

2. Customer Churn analysis

Hadoop was mainly used here to analyze how a telecommunication organization helps their customers. Again, data from many different sources such as social networks and the calls themselves are recorded and then voice analyzed. They were used to know and why the telecommunication company were losing or gaining customers.

3. Recommendation engine

Hadoop was mainly used here to score the particular information details. Consider a Google; Hadoop is acts like the ranking algorithm. Extracting in a cluster of factors like; popularity, link depth, feeds from Facebook, buzz on Twitter etc and then giving scores to the links for display in score order later.

4. Ad Targeting

Hadoop was mainly used here Similar to the recommendation engine, but here deals with providing a better ad-space based on the dimensions of the advertiser paying.

5. Point of sale transaction analysis

Hadoop was mainly used here is simple and straightforward; analyzing the data that is provided by retail shops Point-of-Sale device. However, Hadoop analyze factors like weather and local news, which could help how and why shoppers/consumers spend money in the store.

6. Analysis network data to predict failure

Hadoop was mainly used here to deal with; electricity company which used to measure the electricity flying around their network. They could push in past failures and current fluctuations and then pass the whole lot into a modeling engine to predict where failures would occur.

7. Threat analysis/Fraud Detection

Hadoop was mainly used here to Modeling True Risk. Hadoop can be used to analyze spending habits, earnings and all sorts of other key metrics to work out a transaction are fraudulent. Yahoo! use Hadoop which helps to discover whether a certain piece of mail heading into Yahoo! Mail is really spam.

8. Trade surveillance

Hadoop was mainly used here similar to Threat Analysis and Fraud Detection, but this time pointed directly at the markets, analyzing collected historical and recent live data to see if there is Inside Trading or Money Laundering afoot!

9. Search quality

Hadoop was mainly used similar to the recommendation engine. This will analyze search efforts and then try to offer replacements, based on data gathered and pumped into Hadoop about the links and the things people search for. Hadoop gives similar matched links to the user while searching.

10. Data sandbox

Hadoop was mainly used here as a most useful Hadoop-able problem. A data sandbox is located at somewhere to dump data that formerly thought was too big. The data can be useless or disparate, but Hadoop helps to bring it into meaningful data. Instead of just throwing it away, throw it into Hadoop then see if there is data you can pick up from it. It's cheap to run Hadoop and anyone can attach a data source and push data in. It allows you to make otherwise arbitrary queries about stuff to see if it's any use!

V HADOOP ADVANTAGES AND DISADVANTAGES

1. Advantages

1.1 Store anything

Hadoop stores data in its clusters as it is coming from the data source. This doesn't force anything to transform and avoids information lost. Hadoop can be able to digest, analyze and transform any form of data. It allows analyst to make benefit from the stored data.

1.2 Control costs

Hadoop is open distributed software which runs on commodity hardware. The cost of commodity hardware is very less. The combination of these gives powerful cluster which helps in processing large amount of data.

1.3 Use with confidence

The Hadoop community, including both developers of the platform and its users, is global, active and diverse. Companies across many industries participate, including social networking, media, financial services, telecommunications, retail, health care and others.

The Hadoop community includes both developers and users. Companies across globe are including social networking, financial, telecommunications, retail, health care and others were also participating in this.

1.4 Proven at scale

Petabytes of data that we need to analyze today. Nevertheless, we can deploy Hadoop with confidence because big companies like Facebook, Yahoo! and others run very large Hadoop instances to manage huge amounts of data. The success of the major Web companies in the world demonstrates that Hadoop can grow as your business does. Hadoop lowers costs and extracts more value from data.

2. Dis-Advantages

2.1 Its complex

Not all data fits effortlessly into the rows and columns of a table. It comes from many sources in compound formats such as multimedia, images, text, real-time feeds, sensor streams and many more. Data formats normally change over time as new sources come on-line. Hadoop is able to store and analyze in its natural format.

2.2 There's lot of it

Many companies are forced to discard valuable data because the cost of storing it is simply too high. New data sources make this problem much worse such as people and machines are generating more data today than ever before. Hadoop's inventive architecture, using low-cost commodity servers for storage and processing, stores bulk amounts of data inexpensively.

2.3 It demands new analytics

Simple numerical summaries such as average, minimum, and sum – were sufficient for the business problems of the 1980s and 1990s. Large amounts of complex data, though, require new techniques. Identifying customer likings requires analysis of purchase history, but also a close analysis of browsing behavior and products viewed comments and reviews logged on a web site, and even complaints and issues raised with customer support staff.

Predicting behavior demands that customers be grouped by their preferences, so that behavior of one individual in the group can be used to predict the behavior of others. The algorithms involved include natural language processing, pattern recognition, machine learning and more. These techniques run very well on Hadoop.

VI CONCLUSION

In this paper we have discussed the details about the Big Data and Apache Hadoop Framework components such as HDFS and MapReduce. Here also discussed about the Hadoop use case applications and their advantage-disadvantages. This paper will gives the basic overview about the Big Data and Hadoop Framework and also there applications.

REFERENCES

[1] Blumberg, R., Atre, S.: The Problem with Unstructured Data. http://www.dmreview.com/issues/20030201/6287-1.html (19.02.2009) (2003)

[2] Abiteboul, S., Buneman, P., Suciu, D.: Data on the Web: from relations to semistructured data and XML. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA (1999)

[3] Manola, F., Miller, E.: Resource Description Framework (RDF): Concepts and Abstract Syntax. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ (19.02.2009) (2004)

[4] Brickley, D., Guha, R.: RDF Vocabulary Description Language 1.0: RDF Schema. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/ (19.02.2009) (2004)
[5] Aramburu, Juan Manuel Perez, Maria Jose, Rafael Berlanga and Torben Bach Pedersen, Integrating Data warehouse with Web Data: A Survey, IEEE transaction on knowledge and Data Engineering, Volume 20, July 2008.
[6] Allen, D: Seam in Action. Manning Publications Co. Greenwich, CT, USA (2008)
[7] Bauer, C: Java Persistence with Hibernate. Manning Publications Co. Greenwich, CT, USA (2006)
[8] Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis: http://ceur-ws.org/Vol-464/paper-14.pdf
[9] Vishal S Patil, Pravein D. Soni: Hadoop skeleton & fault tolerance in Hadoop clusters, IJAIEM, Volume 2, Issue 2, and February 2013.
[10] http://www.cloudera.com/content/cloudera/en/resources/library/recordedwebinar/10-common-hadoop-able-problems-recorded-webinar.html.