Cloud Computing Techniques for Big Data and Hadoop Implementation

Nikhil Gupta (Author)
Research scholar
AIIT, Amity university
NOIDA-UP (INDIA)

Ms. komal Saxena(Guide)
Assistant Professor
AIIT, Amity university
NOIDA- UP (INDIA)

Abstract- Big data is identically modernistic and sizzling topic in today's scenario. Big data is a set of data, which is larger in size that a conventional database cannot or does not have the ability to capture, store, manage and analyze the data. The big data is implemented using Hadoop and Hadoop is on demand in cloud now a days.

Big data business ecosystem and its trend that provide basis for big data are explained. There is need of effective solution with issue of data volume, in order to enable the feasible ,cost effective and scalable storage and processing of enormous quantity of data, thus the big data and cloud go hand in hand and Hadoop is very hot and enormously growing technology for organizations . The steps required for setting up a distributed ,single node Hadoop cluster backed by HDFS running on ubuntu (steps only) are given.

Keywords: Big data, Hadoop, Map Reduce, HDFS, Big data business ecosystem, Scalable database management system.

I. INTRODUCTION

Everyday 2.5 quintillion bytes[3] of data is created or produced. Data is generated by smart phones, posts to social media sensors, blogs etc. Now days we are looking big data as a business perspective. Big data and cloud go hand in hand. The cloud computing has enabled the companies to get more value from their data even before, by enabling fast analytics at low cost then they were before. Thus companies can store even more data. Now this more data is increasing day by day and we need new technology and paradigms to collect, store and analyze data as and when needed. Now the problem with conventional database architecture is, it was not able to handle the data is petabytes size. So, ultimately there is requirement of new technology, hence Big data was introduced with new enhancements.

In section I the introduction is given about Big data and scalable database management system. The section II is about aspects of big data and its challenges. Limitations and issues are in section III. The section IV helps us to choose between the Hadoop or data warehouse. Section V is Hadoop for Big data. Big data management ,scalability and performance is outlined in section VI. Big data business ecosystem is described in section VII. Running Hadoop on Ubuntu Linux (Single-Node Cluster) is in section VIII. Finally the conclusion is in section XI.

BIG DATA definition- The set of data which is larger in size, that a traditional database cannot or does not have the ability to capture, store, manage and analyze.

1 petabyte=1000 TB (terabyte)

The 2 breakthroughs have helped to adopt the solution for handling big data

- Availability for cloud based solution.
- Distribution of data over many servers.

Scalable database management system[10]

It has been the vision of database community for 3 decades

- Distributed database- used for update intensive workloads.
- Parallel database- For analytical workload.

Parallel database system has grown very large commercial system but distributed system was not very successful.

So, changes in data access patterns led to the birth of new system referred as key value stores. This was the adoption of map reduce paradigm and its open source Hadoop implementation.

II. Aspects of big data and its challenges

1. *Volume*-Large volume of data is generated in year 2020 it is expected to store 35 zetabytes of data. On a daily basis the facebook generates 10 TB daily and twitter generates 7 TB of data.
   Challenges- thus virtualization of storages in data centers, and we can make use of no SQL database to store and query huge volumes of data.
2. *Velocity*- Speed of data i.e. more and more data is produced and must be collected in shorter time of frames.
   Example- Show all people currently affected by flood i.e. it is updated by GPS data in real time.
   Challenge: Real time data processing is needed.
3. *Variety*- To handle multiple sources and format as data can be a raw data, structured data, unstructured data etc.
   Challenge- This is against the traditional relational data model, the way we collect data, thus has created new data stores that are able to support flexible data model.
4. *Valu-:* The main concern for every organization is how to make data useful. Main point is to convert the raw data into valuable data.
5. *Data Quality*- How good data is? All the decisions will be made according to the quality of data. A good process ill make good decisions if based on good data quality.

III. LIMITATIONS OF BIG DATA

There are very limited persons who are having knowledge or highly skilled to take advantage of big data.

Issues

- *Data Policies:* As an example the storage computing, analytical software all these requires as in new for big data.
- *Technology and techniques:* Privacy, Security is required for data.
- *Access to Data:* When we have to access the data then we need to integrate the multiple data sources together.

IV. Which one to use? Hadoop or Data warehouse

Requirement	Dataware house	Hadoop
Low latency, OLAP, and interactive reports	✓	
SQL compliance is required (ANSI 2003)	✓	
Preprocessing or expedition of raw unstructured data		✓
Online archives alternative to tape		✓
High-quality purified and persistent data	✓	
100s to 1000s of concurrent users	✓	✓
Determine unexplored relationships in the data	✓	✓
Parallel complex process logic		✓
CPU intense analysis	✓	✓
System, users, and data governance	✓	
A number of limber programming languages running in parallel		✓
Unrestricted, ungoverned sand box explorations		✓
Analysis of provisional data		✓
Huge and vast security and regulatory compliance	✓	
Relative data loading and 1 second tactical queries	✓	✓

Table 1: which one to use Hadoop or data warehouse [2]

V. HADOOP FOR BIG DATA

Overview: In 2002, Dough Cutting develop an open source web crawler project, the Google published map reduced into 2006. Dough Cutting developed the open source, map reduced and HDFS.

What is Hadoop?

Hadoop is a framework that allows distributed processing of large data sets across the clusters of computers using a programming language. It is open source library & application programs written in JAVA language. Hadoop implements HDFS (Hadoop distributed File system).

Hadoop clusters running the same software can range the size from single server to as many of thousand servers.

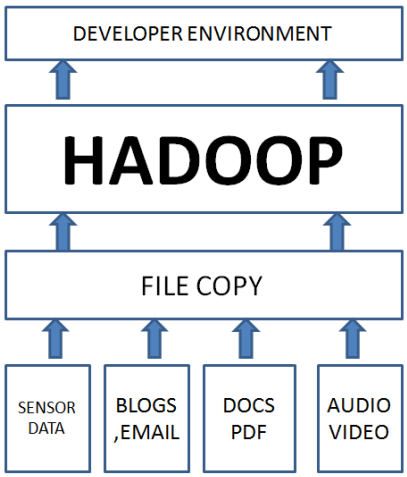

Fig 1: Hadoop supports different types of unstructured data with clustering.

Components of Hadoop

1.HDFS (Hadoop Distributed File System)

It is file system designed for storing very large files with streaming data access pattern running clusters on hardware. HDFS is highly fault tolerant. HDFS automatically distributed file across clusters and retrieves data by file name. HDFS does not change the file once it is written. Thus if any changes has to be made the entire file must be written.

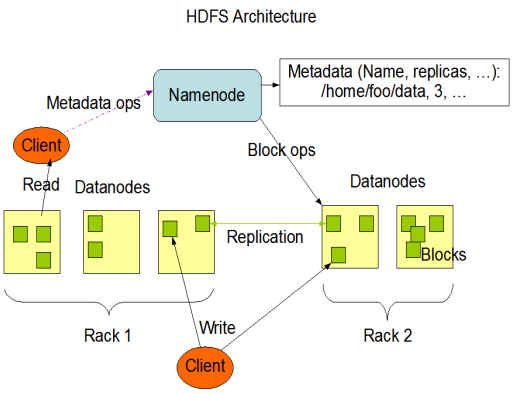

Fig 2: HDFS Architecture [16]

2. MAP Reduce

It is programming paradigm. It has 2 phases for solving query in HDFS:

- Map
- Reduce

Map is responsible for to read data from input location and based on input it generate a key value pair i.e. an intermediate output in local machine. Reduces the responsible for to process the intermediate the output receives from mapper and generate file output.

Map reduces for data processing enables Hadoop to process large dataset in parallel across all node in cluster.

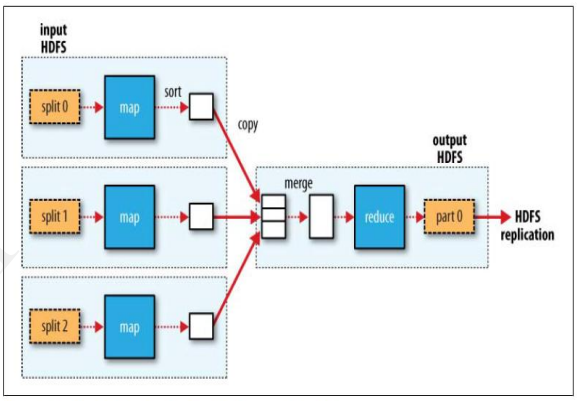

Fig 3: working of map reduce for a single node [17].

In many ways, Map Reduce can be seen as a complement to an RDBMS. Map Reduce is a good fit for problems that need to analyze the complete dataset, in a batch fashion, especially for ad hoc analysis. The relational database is good for some queries or updates, where the dataset has been tabulated to deliver low-latency retrieval and update times of a relatively small amount of data.

Considerations	Traditional RDBMS	MapReduce
Data size	Gigabytes	Petabytes
Access	Interactive and batch	Batch
Updates	Read and write many times	Write once and read many
Structure	Static schema	Dynamic schema
Integrity	High	Low
Scaling	Nonlinear	Linear

Table 2: difference between RDBMS and map reduce [17]

Advantages of Hadoop:

- No software license is required
- Low system acquisition cost
- Automatic, system maintenance.

Disadvantages of Hadoop:

Hadoop has a centralized metadata store i.e. namenode, which represents a single point of failure without availability .when the namenode is recovered it can take a long time to get Hadoop cluster running again.

VI. Big data management and scalability and performance with respect to Hadoop.

There is need of effective solution with issue of data volume, in order to enable the feasible, cost effective and scalable storage and processing of enormous quantity of data. There are quality constraints on both:

- Storage of big data: Up to which degree data has to be replicated and latency requirements .
- Processing: where with, which parallel requirement of computing resources are required.

VII. Big data business ecosystem:[6]

Trends that provide basis for big data ecosystem:

1. Data science and that associated skills are high in demand eg. data scientists.
2. Generalization of big data platform i.e. we need more responsive application. we need to write models that discover patterns in near real time.
3. Commoditization of big data platform:
   Two approaches
   (a) Make big data accessible to developers by making easy to create applications.
   (b) To find a use case for big data like face recognition, finger print reader voice recognition etc.
4. Increase cross enterprise collaboration:
   Requirement is for sharing, exchanging and managing data across platform.

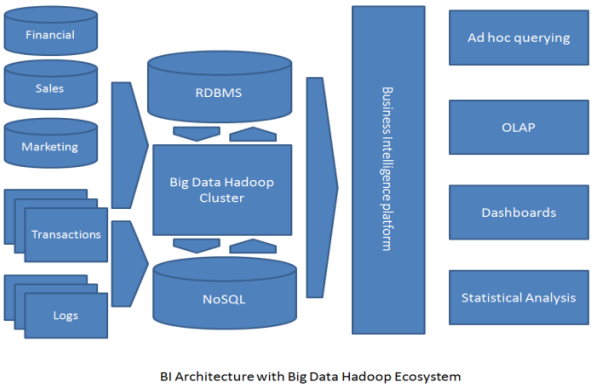

Fig 4: Big data Hadoop ecosystem

VIII. Steps required for setting up a distributed, single node Hadoop cluster, backed by HDFS.[16]

step 1- Install Ubuntu

step 2- Install Java in it

step 3- Add a dedicated Hadoop System user

step 4- Configure SSH

step 5- Now test SSH by connecting to your server

step 6- Disable IPV6

step 7- Hadoop Installations

step (7.1) -Configuring HDFS

step 8- Configure directory where Hadoop will stores files.

step 9- Format the HDFS file system via in name node

step 10- Start your own single node cluster

step 11- Run a map reduced job

step 12- Copy Local data to HDFS

step 13- Run map reduce job

IX. CONCLUSION

Big data is the next frontier for innovation competition and productivity. In horizon 2020, big data finds its place in Industrial leadership. There is need for structuring data in all sector of economy.
Hadoop and Cloud computing are in great demand in several organizations. In upcoming time, Hadoop will become one of the most required technology for Cloud Computing. This proof is given from total number of

Hadoop clusters offered by cloud vendors in many business.

Organizations are looking to expand Hadoop use cases to include business critical, secure applications that easily integrate with file based applications.

There is need for tools that do not require specialize skills and programmer. New Hadoop developments must be easier for users to operate and to get data in and out. Thus, this includes direct access with standard protocols using existing tool and techniques.

REFERENCES

1. Big Data Now by O'Reilly Media Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.
2. Hadoop and the Data Warehouse: When to Use Which Dr. Amr Awadallah, Founder and CTO, Cloudera, Inc., Map Reduce and the Data Scientist", Colin White, 2012.
3. Sam B. Siewert "Big data in the cloud", ,Assistant Professor,University of Alaska Anchorage,9th july 2013.
4. Dr. Satwant Kaur ,keynote at the CES show:" many trends and new technology developments for big data ",IEEE International Conference on Consumer Electronics (ICCE 2013).
5. Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.
6. NESSI white paper, Big data a new world of opportunities ,DEC 2012.
7. http://www.baselinemag.com/cloud-computing/managing-big-data-in-the-cloud
8. Divyakant Agrawal ,Sudipto Das, Amr El Abbadi, Department of Computer Science,University of California, Santa Barbara.
9. https://cloudsecurityalliance.org/research/big-data/#_news
10. The Apache Hadoop Project., http://Hadoop.apache.org/core/, 2009.
11. http://www.ibm.com/developerworks/library/bd-bigdatacloud/
12. D. Agrawal, S. Das, and A. E. Abbadi. Big data and cloud computing: New wine or just new bottles? PVLDB, 3(2):1647–1648, 2010.
13. http://www.edupristine.com/courses/big-data-Hadoop-program/?jscfct=1
14. http://www.edureka.in/blog/category/big-data-and-Hadoop/
15. www.forbes.com/big-data
16. www.michael-noll.com/tutorials/running-Hadoop-on-ubuntu-linux-single-node-cluster/
17. Hadoop: The Definitive Guide, Second Edition by Tom White, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.