ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 14, April 2016.

AN INTELLIGENCE HADOOP TOOL TO HANDLE STRUCTURED, UNSTRUCTURED AND SEMI STRUCTURED DATA

Mr.S.S.Aravinth, AP/CSE, Knowledge Institute of Technology, Salem

Mr.J.Gowrishankar,AP/CSE, Knowledge Institute of Technology, Salem

Mrs.R.Saranya, AP/CSE, Knowledge Institute of Technology, Salem

Mrs.T.Dhivya, AP/CSE, Knowledge Institute of Technology, Salem

Mail ID: ssacse@kiot.ac.in Mobile: 098944 – 48683.

ABSTRACT:

Data driven approach is a warm way to make efficient decision makings to all industries. Having huge amount of data affects the data processing. Due to the large augment of data, the industries are struggling to store, handle, and analyze data. This phenomenon is called big data in which the normal data base systems are not enough to do the above mentioned activities. Now a day IT industries face the real challenge to deal with these data. They need novel technology advancements to address huge amount of data processing. For that, the solution is Highly Available Distributed Object Oriented Platform (HADOOP). In Hadoop the enormous data will be stored and processed effectively and efficiently. Hadoop is the technology which has many frameworks such as data integration, management, orchestration, monitoring, data serialization, data intelligence, storage, integration and access. Industries are planning to resolve this entire surface of Hadoop frameworks. So In this paper the fore mentioned frameworks are studied and analyzed very well. This paper will help to all the Hadoop technology learners who are all seeking to do their research work in this domain. In addition to this the studies of the map reduce programming technique and Hadoop distributed file system are presented well.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 14, April 2016.

Keywords:

Business Analytics, Business Intelligence, Hadoop, Map Reduce, Distributed File System and Frameworks

I) Demerits of Existing Storages:

1. Traditional Databases:

In traditional DBMS problems such as data redundancy which means that the same data is present in more than one file. Same details are present under different categories in various files. This is further leads to data inconsistency. For example, if address is the detail present in various files, if any one address is to be updated then updating should be done in all the files wherever that address is present or else it is difficult to fetch details from files. Data isolation cannot be done and it also faces a serious problem in security and access control.

2. DBMS Storage:

Online accessing is carried out here which leads to access of same database by various users at the same time. Here maintaining privacy and data security is tough because of many numbers of users. Number of constraints will be more to ensure the quality of data base as it is accessed by many users. As the single data base is accessed by many parallel updating will leads to undesired output.

3. Data warehouse Storage:

Implementation under this technique is very long and it's not affordable too. There is a lack of flexibility.

4. Challenges in Dealing Enormous Data:

In industries data handling is a very big constraint since there is a huge clusters of data present in each firm. Data handling includes data storing and data processing. To handle this situation we are in need of a tool which will act as a solution for this constraint.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 14, April 2016.

So where is the Solution?

II) Hadoop – A Tool for Processing Big data:

Hadoop is a tool which should be used for the purpose of storage and access of clusters of data. Its frame works can be used to perform various tasks.

III) Introduction to Hadoop:

Nowadays we can find huge amount of data which is to be stored and accessed in various firms. To store and access huge amount of data which is also known as clusters of data a tool called 'HADOOP' is the solution.

HADOOP is a tool which is used for accessing and storing of structured data, unstructured data and semi-structured data.

IV) Hadoop Core Components:

There are two major things which we can do on clusters of data or any data which is storing and accessing. HDFS (Hadoop Distributed File System) is a tool which is used for storage purpose. Map reduce is a technique which is used for accessing clusters of data.

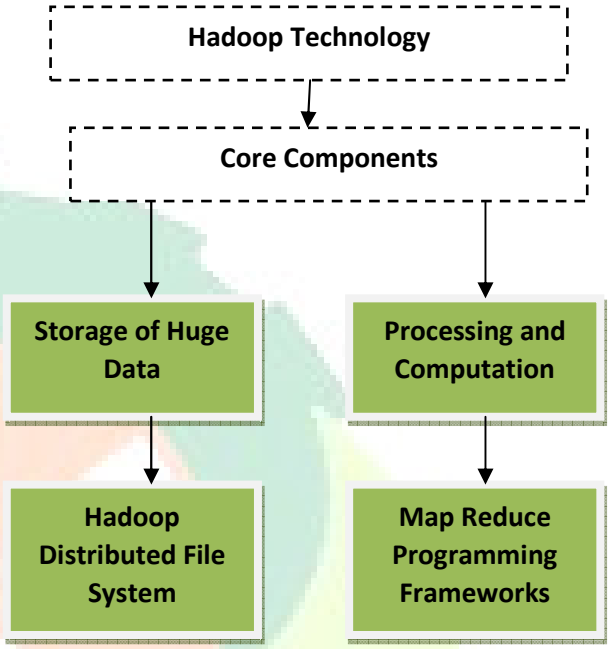

Figure 1. Hadoop Core Components

1. Cluster Storage with Commodity Hardware:

RDBMS is a technique which is followed in small network (i.e.) for domestic use. While considering an industry with clusters of data the technique called 'HADOOP' should be used for effective storage and accessing. Clusters of data can be handled with the frameworks of Hadoop.

V) Hadoop Technology Stack

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 14, April 2016.

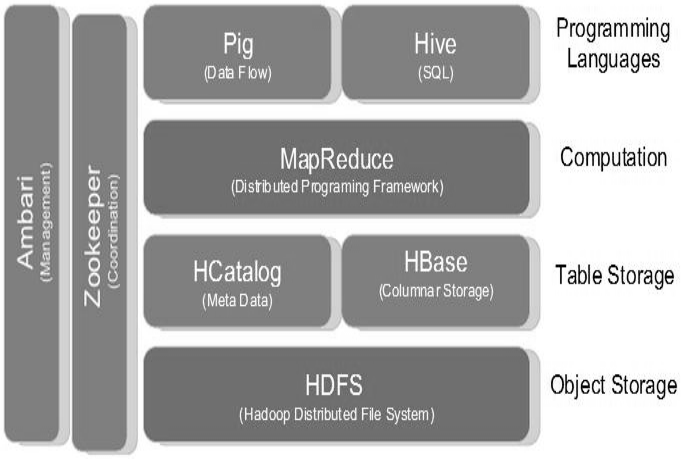

Figure 2. Overview of Hadoop Eco System:

Hadoop eco system comprises of frame works such as data integration, data access, data storage, data serialization, interaction and visualization, execution development, data intelligence, managing, monitoring, orchestration. There are few more part which is present within this frame works.

1. Data Integration:

To transfer data from relational database to Hadoop, an application called scoop is used. Scoop is a command line interface application. With the help of this exports can be done from Hadoop to relational database. Imports to hive or hbase is also possible with the help of scoop.

2. Data Management, Orchestration& Monitoring:

Ambari is a tool which aims to simplify Hadoop management over clusters of data. It is software which is used for provisioning, managing and monitoring of apache Hadoop clusters. Configuration and synchronization services are provided by Zookeeper. Yahoo, EBay are some companies which uses Zookeeper. Oozie is a system which is used to schedule the workflow of Hadoop jobs.

3. Data Access:

To create Map Reduce programs there is a high level platform called 'PIG' is used. The language which is used in this high level platform is 'PIG LATIN'. Initially researchers at yahoo research developed 'Pig' at 2006, Later it was moved to apache software foundation. User can write in java, python, JavaScript, ruby and they can be extended to pig Latin using (UDF) User

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 14, April 2016.

Defined Function. Hive is active from June, 2014 and it's licensed under Apache License 2.0. Hive is an infrastructure for data warehouse which provides things such as data summarization, query and analysis. Initially it was developed by Facebook and now it's used by many other companies. It provides HiveQL which is a SQL-like language to support Map Reduce.

4. Data Storage:

Apache Hbase was developed by apache software foundation and the stable release is on 21 July 2014. It was modelled after Google's big table which is an open source, non-relational, distributed database. A large amount of sparse data can be stored, provided fault-tolerant way approach. Compression, in memory operation, bloom filters are some of the features of Hbase. Cassandra is an open source distributed database management system which can handle clusters of data. There will be no failure is the key feature of Cassandra. It also features that the through put is very high. Cassandra Query Language (CQL) is the query language which is used here.

5. Interaction and Visualization:

Here a tool called Hcatalog is used which is nothing but a table. It also manages the storage of data. With the help of tools like pig and some other tools Hcatalog can be accessed. With the help of Hcatalog data availability can be notified and table abstractions provide us where the data is stored. A text search engine called lucene is used which provides accurate results. It is completely written in java language and it's ranked as best search engine.

6. Execution Development:

A tool called Hama is used which helps in bulk synchronous parallel computation for undergoing computation of massive scientific things and it's created by Edward J. Yoon.

7. Data Serialization:

Avro comes under data serialization frame work in Hadoop. Data type and protocols can be defined with the help of

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 14, April 2016.

JSON which is followed by Avro. Serialization of data in binary format takes place here. It provides communication between Hadoop nodes in a wire format and also between Hadoop services and client formats. With the help of programming languages like java, c#, c, c++, python, ruby we can access Avro as it provides API's written for this language.

8. Data Intelligence:

Drill is a technique which is used for analysis of data-sets which is present in a large scale. This frame work is an open source software and it is used to handle highly intensive database. In current scenario it's in incubation stage at apache. For implementation of distributed or scalable machine learning algorithms a platform called 'Mahout' is used. It covers areas like

collaborative filtering, clustering and classification.

Conclusion:

Processing of enormous data is very challenging tasks in many industries. Many of the industries use cases are being deployed by industry professionals. To take advantage of all the huge volume of data it's not possible to deal with old traditional database system a long drive with hadoop technology tool will focus about processing of various data types generated by various sources. To get the big insight in heterogeneous data platform the hadoop tool frameworks are being advised by most of industrialist, practitioners' and researchers. This paper will give a real picture of various hadoop platforms for developers and reserchers.

References:

1. Michael Minelli, Michelle Chambers, and Ambiga Dhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley, 2013.
2. P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence", Addison-Wesley Professional, 2012.
3. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012.
4. Eric Sammer, "Hadoop Operations", O'Reilley, 2012.
5. E. Capriolo, D. Wampler, and J. Rutherglen, "Programming Hive", O'Reilley, 2012.
6. Lars George, "HBase: The Definitive Guide", O'Reilley, 2011.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 14, April 2016.

7. Eben Hewitt, "Cassandra: The Definitive Guide", O'Reilley, 2010.
8. Alan Gates, "Programming Pig", O'Reilley, 2011.