ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

RICH ANALYTICS WITH HADOOP TECHNOLOGY

Dr.Shahnaz Fatima
Amity Institute of Information Technology
Amity University,Lucknow,India

DOI: 10.5281/zenodo.1252936

ABSTRACT

The period of 'big data' represents new challenges to businesses industry. Incoming data volumes are exploding in variety, speed, volume and complexity. It is not defined anywhere as such that how much volume of data will be considered as "big" data, but handling of such data requires lot of new tools and techniques to process it. To harness the power of Big data one needs an infrastructure and technologies which can deal with huge volume and variety of data as well as can draw inferences from it. There are various technologies for "Big Data Analysis" given by various vendors. The technologies are growing rapidly with the growing market of Big Data Analytics. The paper will provide you with a perspective on the technology platforms for big data and analytics. This paper presents an overview of Hadoop.

KEYWORDS: Hadoop,Apache,Analytics,Big Data

I. INTRODUCTION

BIG DATA ANALYTICS

Big data deals with the variety of a data. Different variety of data needs different analytic logic to produce results which work as an insight for the business[6].The data analysis is not at all a new concept in a business stream. It is essential for strategic planning in the business houses. The trend of analysis is to have an insight of the current market trends and customer demands. The big data technologies give you more and more accurate analytical results. This leads to the concrete data analysis as well. The robustness of big data can be best utilized by having an infrastructure which can manage multiple machines on the same time. The infrastructure should be capable of managing huge volume of structured and unstructured data, without compromising its security and privacy.

Big data technologies can be classified into two main categories-

a) Operational data
b) Analytical data

Operational data

Operational data includes sales, service, order management, manufacturing, purchasing and billing. The applications working at the operational level require significant amount of data related to the product sales, customer and other related data like price, discounts etc.

Analytical data

Analytical data is all about planning and decision making. It works as a support system of the business. It gives a deep insight about the Customer market needs, Suppliers performance, Market tendency, Product behavior etc.

II. BIG DATA TECHNOLOGIES

There are a number of disruptive and transformative big data technologies and solutions that are rapidly emanating and evolving in order to provide data-driven insight and innovation[1].The variety of technologies is given by the various vendors for big data. Apache has given the set of following technologies like **Apache hadoop, Apache Pig, Apache Hive, Apache Hbase, Apache Spark** and many more. The paper will give an overview of Hadoop technologies.

Apache HADOOP

Hadoop is an open source technology introduced by Apcahe for processing the data on a very large set of machine clusters. Apache Hadoop framework is written in Java and it is utilized by a community of users and contributors. Hadoop is designed to scale up with the clusters of machines capable of doing local storage and computation as well.

Hadoop Architecture

Hadoop Architecture is based on four major components –

1. Hadoop Common
2. HDFS
3. MapReduce
4. Hadoop Yarn

Hadoop Common

It has Scripts and Java Files that are used to start Hadoop. It mainly consists of Java utilities and Libraries required by Hadoop modules.

HDFS (Hadoop Distributed File System)

HDFS as names says is a distributed file system. Though it is quiet similar to the earlier distributed file system but still it has many significant variations as well. HDFS is capable of handling the clustered machine environment and provides a high throughput access to the application data. It is quiet suitable for the applications having really large data sets. It's efficiency lies in that it is written in robust language like Java hence provides scalability and portability.

HDFS Components

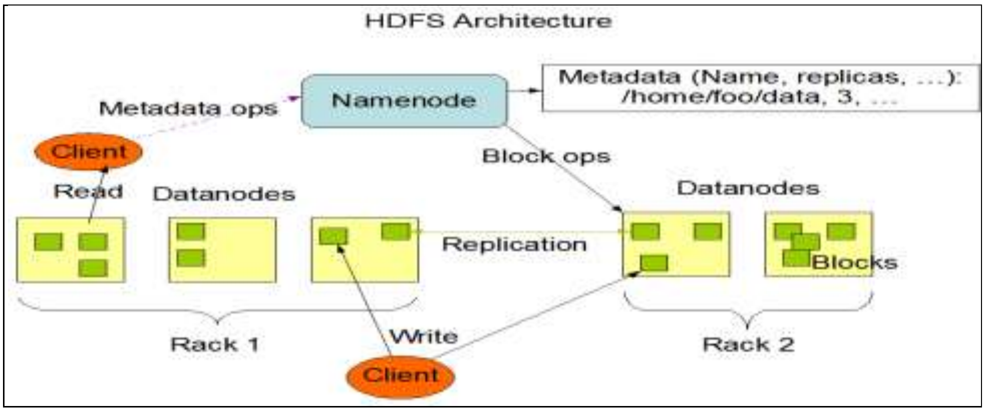

Figure 2: Hadoop Architecture[5]

Figure 2 clearly shows the Hadoop Framework containing the HDFS components

NameNode, DataNode and Blocks

There is a single NameNode and multiple DataNode. HDFS works on a principle of server and client. NameNode works as a server and manages the file system and make it available to all clients. DataNodes works as a client and can be many. Each of them serves the block of data over the network.HDFS namespace is divided into one or more blocks.

Files/Directories

HDFS supports a traditional file system which is hierarchical in nature. It gives you facility to create, modify, copy, delete and manage all files. The file system is maintained by NameNode. The number of permissible replica of a file is specified by the Application itself. It is called as the Replication factor of that file.

Replication

The HDFS is the distributed storage system which handles the storage of files in Hadoop. HDFS provides a reliable and fault tolerant architecture to store files[4]. The files are stored in the sequence of blocks of same size except the last block. The block size and replication factor is specified for every file. Replication factor can be given at file's creation time and can be changed later. NameNode which works as a master takes all decisions related to replication of blocks.

Every DataNode sends a block report and hearbeat to a NameNode, confirming it about its proper functioning.HDFS seems to be vulnerable as it is designed to store data on inexpensive data storage as well as unreliable too. However its designer took keen precaution for the safety of data and keep three copies of all the data blocks in order to bear any kind of data loss.

MapReduce

MapReduce is the heart of Hadoop. It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster[3].MapReduce can be called as a processor of the Hadoop framework. It helps in processing big amount of data parallel on various clusters. The MapReduce is basically a culmination of two tasks Map and Reduce. The **Map Task takes** each element of input data into key-value pairs also known as tuples.

The **Reduce Task** takes it input from the output of Map task and reduced all the tuples to the same key. The MapReduce framework works on master slave concept. There is a one master called JobTracker and multiple TaskTracker which works as a slave. JobTracker pushes the jobs to the various TaskTrackers. The roles of JobTracker being as master are to manage all resources, keeping the track of consumption/ availability of resources, scheduling of jobs for slaves and re-executing the fail tasks.The JobTracker tries to push the job to the node which contains the data or which is nearby. The JobTracker (master node) receives an incoming job through the JobClient (client). The received job is queued up into Pending Job List [2].

III. CONCLUSION

Big Data refers to volume, variety and velocity of data. Big data handles the data set which is so large and complex and it is impossible to handle such volume and complex data with ordinary software tools. The big data is also categorized into Operational and Analytical data. To harness such variety and volume of data various technologies are being introduced. The paper has discussed latest Apache hadoop technology that are in use to handle such a variety and volume of data.

IV. ACKNOWLEDGEMENTS

This **research** paper was supported by Dr.Parul Verma.I would like to thank Amity University

V. REFERENCES

1. M. Daud Awan, and M. Sikander Hayat Khiyal, Big Data in Cloud Computing: A Resource Management Perspective Saeed Ullah , Hindawi Scientific Programming Volume 2018,
2. Radheshyam Nanduri, Nitesh Maheshwari, Reddy Raja, Vasudeva Varma,Job Aware Scheduling Algorithm for MapReduce Framework , 3rd IEEE International Conference on Cloud Computing Technology and Science,2011
3. Ananthi Sheshasayee1 and J. V. N. Lakshmi ,Comparison of Machine Learning Algorithm on Map Reduction for Performance Improvement in Big Data, Indian Journal of Science and Technology, 2015
4. Franklin John#1, Suji Gopinath#2, Elizabeth Sherly#3, "An Efficient Dynamic Data Replication for HDFS using Erasure Coding ",International Journal of Computer Science and Information Technologies, Vol. 8 (2) , 2017
5. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html\
6. Dr.Shahnaz Fatima, Dr. Ranjana Rajnish and Dr. Parul Verma, CHALLENGES OF BIG DATA PRIVACY AND SECURITY: A REVIEW, International Journal of Latest Trends in Engineering and Technology,2017

CITE AN ARTICLE

Fatima, S., Dr. (2018). RICH ANALYTICS WITH HADOOP TECHNOLOGY. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, 7*(5), 561-563.