Vandana Publications
IJEMR

Volume-7, Issue-4, July-August 2017
International Journal of Engineering and Management Research
Page Number: 103-104

Review of Big Data Mining Technologies

Samar Wazir[1], Shah Imran Alam[2]
[1,2]Department of Computer Science and Engineering, School of Engineering Sciences and Technology,
Jamia Hamdard, New Delhi, INDIA

ABSTRACT

The automation and modernization of industries leads to store huge amount of current and past data for analysis and processing which helps in decision making and business strategies for reducing risk. So as the size of data increases the processing power of hardware decreases and by this way the size of hardware got saturated. The different type of data which cannot be processed by already existing computer hardware due to its size, structure, variety and other properties is known as big data. So for handling big data we have two options, either we can switch to better hardware or we can change the strategy of processing. In this paper we focus on second option that is what can be the strategies for handling big data and which one is better. Hence various technologies of big data is comprehensively surveyed in this study.

Keywords-- Big Data, Map Reduce, HDFS, Cloud Computing, Structured and Unstructured Data

I. INTRODUCTION

The population of world increasing exponentially [3] and more than 2 billion of this population is using internet services. Furthermore, mobile devices become a necessary part of every one’s life and more than 4 billion of people are using mobile devices and generating data. In another case, remote sensors are producing huge amount of data every day. So in today`s world we are getting huge amount of data in structured or unstructured format every day and storing/ processing of such kind of data become a complex task. This data can be called as Big data[1]. In[1], Big Data is defined by HACE theorem, i.e., Big data starts at huge volume from Heterogeneous and Autonomous sources having decentralized control and distributed, also having Complex and Evolving relationship among data. As the Big Data problem increased various solutions were also proposed and a race started for processing data by different techniques. The data from different industries and different formats handled by different solutions[2]. So now the problem arise that which Big data solution is best suited for a particular data format[8]. Hadoop framework is one of the solution of Big data problem and provide by different vendors as open data platform. Some vendors are Cloud era, Horton Works, Amazon, Microsoft, Map R and IBM Info Sphere Insights.

The data created by different applications is very big in a single day and the study of recent years data shows that its speed is almost 90% hike in the size of data[9]. Therefore the demand of big data technologies[4] increasing rapidly and the main focused tools by big data vendors are Map Reduce[5], HDFS[6] and YARN. Some of the key points vendors keep in mind while providing big data solutions are:

- Easy support and simple for user
- Easy bug detection, fixes.
- Customization of apps for users.

In today`s market the top vendors for Big Data solutions are

- Cloudera
- Amazon web services elastic MapReduce hadoop distribution
- HortonWorks
- MApR
- Microsoft
- IBM InfoSphere Insights

In this paper we study how to select a big data platform for a particular organization.

II. CLOUDERA

Cloudera is one of the top big data vendor for providing hadoop framework as open big data paltform. It is one of the most trusted and reliable hadoop vendor from customer view.Cloudera having more than 300 paying customers including US army and Monsanto. It is occupied almost half of the Hadoop market. Interective GUI, automatic and fast recovery in case of node failure, valuable add-on tools e.g. Navigator, Manager and Impala are the key features of Cloudera Haddop framework.

III. AMAZON WEB SERVICES ELASTIC MAPREDUCE HADOOP DISTRIBUTION

Amazon did a tremandous job in this area by introducing Amazon Web Services and it is present since the realese of Hadoop. Amazon provide a very impressive and somple HDFS designh with easy and well-arranged data analytic stand as Amazon Elastic MapReduce(EMR). Amazon Web Services Elastic MapReduce is the biggest vendor across the globe. Amazon another improtant contribution is Dynamo DB.

IV. HORTONWORKS

It is surveyed about HortonWorks that it has more than 50 fresh customer with huge account and joint ventures with Microsoft, Teradata and many more[9]. Apache Ambari is another project provided by HortonWorks for administration of Hadoop clusters and running, controlling and supervising big data clusters. It is also the member of open data platform initiative by IBM and other vendors. It claims to provide 100% open source distribution.

V. MAPR HADOOP DISTRIBUTION

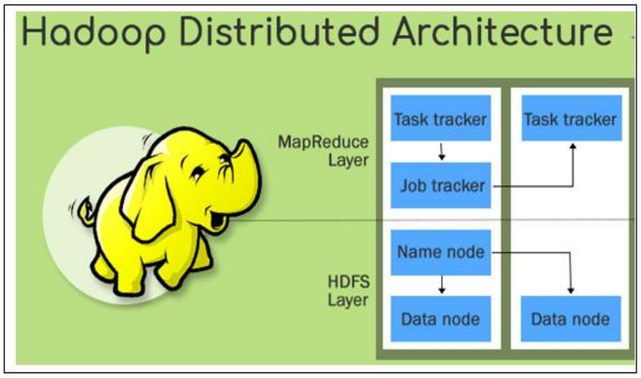
Technologies like MapR provide greater efficiency with less effort. MapR technologies provide built in API for HDFS and save lot of effort in writing code. It is also provide reliable and seccure delivery. It is very efficient for handing custers of large size[9].

VI. MICROSOFT HADOOP DISTRIBUTION

Microsoft also plays a vital role as Hadoop[7] vendor by providing different cloud based services. Microsoft offers window Azure`s HD insight as cloud manufactured goods. It is also provide Polybase for handling SQL queries efficiently.

VII. IBM INFOSPHERE INSIGHTS

Hadoop vendors diiferent solutions for handling diiferent kind of structured and unstructured data. In this environment where IBM is well known for its research projects, provide many of the Big Data solutions. The main focus of IBM is the data analytics part and data management part. It provide solution for machine learning problems as Apache System ML. the customers gain market repidly by their apps integration with IBM Big Insights.



VIII. CONCLUSION

Now the technology shifts to data analytics so it very important to develop new techniques for data analytics by studing current technologies and data designs. When an industry shifts from traditional technology to big data technology then it incuured lot of risk and if we donot shift then it will reflect in market competition failure. So it is necessory that before switching our business process we shold study the solution available very carefully.

REFERENCES

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu , "Data Mining with Big Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014
[2] Big Data Now: 2012 Edition, *O'Reilly Media, Inc.*
[3] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition.
[4] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, "Bigtable: A Distributed Storage System for Structured Data", *Google, Inc.*
[5] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, "The Google File System", Google, *SOSP'03,* October 19–22, 2003, Bolton Landing, New York, USA. 2003 ACM
[6] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplifed Data Processing on Large Clusters", *Google, Inc.*OSDI 2004
[7] Garry Turkington, "Hadoop Beginner's Guide".
[8] Yahoo! Developer Network Blog, https://developer.yahoo.com/blogs/ydn/hadoop-future-bright-48231.html
[9] https://intellipaat.com/blog/top-6-hadoop-vendors-providing-big-data-solutions-in-open-data-platform/