INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

BIG DATA ANALYTICS ARCHITECTURE

ER. AMRINDER KAUR

Assistant Professor, Dept. Of Computer Sc. & Applications, Kurukshetra University, Haryana.

Accepted Date: 25/10/2015; Published Date: 01/11/2015

Abstract: - The term Big Data describe new architecture, new algorithms and new techniques to capture data from various sources like social sites, sensed data, web forms, emails, climate information etc. This captured data can be structured, semi structured and unstructured, resulting in incapability of conventional data management tools. To analyze the collected data and to find hidden patterns and useful knowledge various tools were developed, which is helpful to make analysis process easy and simple to some extent.

Keywords: Big Data, HDFS, Map-Reduce, Client-Server Architecture

Corresponding Author: ER. AMRINDER KAUR

Access Online On:

www.ijpret.com

How to Cite This Article:

Er. Amrinder Kaur, IJPRET, 2015; Volume 4 (3): 131-136

PAPER-QR CODE

1. INTRODUCTION

Today internet becomes ubiquitous in everyone's life. Every person is responsible for generating data ranging from few bytes to megabytes or even more. Data presented now a day's have size (volume), complexity (variability) and growth rate (velocity) which arises a need to make traditional processing systems up to date so that it will be able to process this data which is termed as Big Data . Various challenges of big data include analysis, curation, storage, visualization, and violations of privacy.[4][5]

2. AN ARCHITECTURE OF BIG DATA

The Internet and globalization is work as a fuel for economy, social networks and mobile computing are responsible to amplify it. Big Data is becoming worldwide concerned. Customer needs and dynamic market are responding by analyzing big data, and this analysis can provide immense competitive interest. This mined knowledge is very useful for a business and organization to earn profit or to define new policies so that optimization is obtained. To achieve these goals and for mining knowledge client-server architecture for Big Data is shown in Figure1.

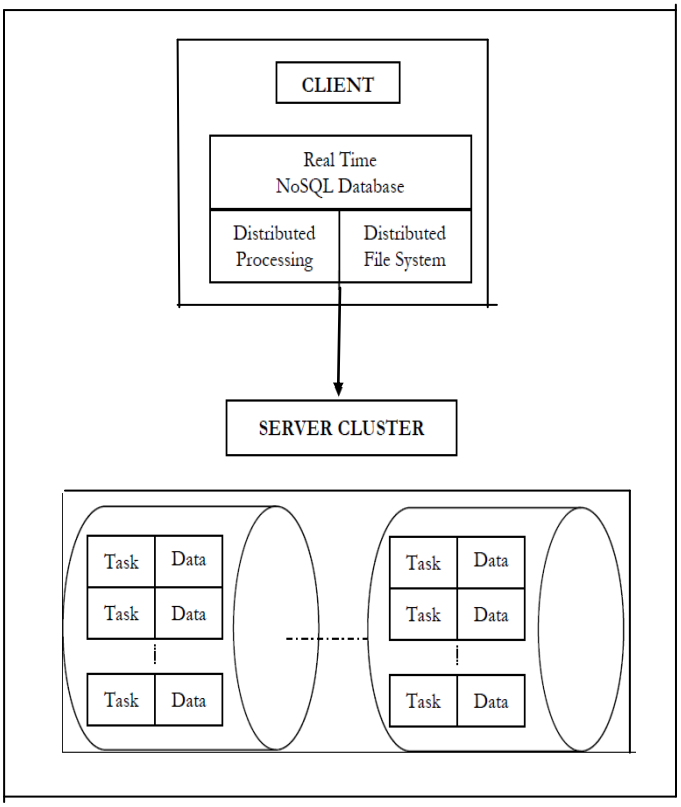

Figure 1: Big Data's conceptual architecture

2.1. ARCHITECTURE AT CLIENT LEVEL:

The architecture at client level consists of distributed file systems, distributed processing framework and NoSQL databases. NoSQL is interpreted as “Not Only SQL”. These databases are open source and non-relational databases which store data in pairs i.e. key-value for big data it is responsible for providing distributed and highly flexible data storage. NosQl data bases types are document base, graph stores, key value stores, and wide column stores. In document base type a key is associated with each document. Further this document can contain nested documents or key value pair. Graph stores are used to depict the information about social networks. In key value store, each data is stored as an attribute with its corresponding value. Whereas wide column store is best for queries over large database and it store data in column major order [2]. A popular example of NoSQL database is Apache Hbase. It is an open-source, It store data in column-oriented form that dispense arbitrary, real time read access and write access to data.

The next layer is distributed file system that distribute the data on clusters to provide scalability and able to handle bulk of data. Distributed processing frameworks disburse data processing or calculations over clusters and maintain load balancing. The impetus behind distributed processing was the availability of communication technology and powerful microprocessors. Apache Hadoop is the popular platform for distributed processing. For distributed processing it distribute the data across clusters of computers on the communication network. The computers used during computation have its own storage and computation ability. Hadoop is able to provide scalability from few servers to thousands of machines.

Map Reduce and HDFS known as Hadoop distributed file system are two essential component of Hadoop. To provide high throughput HDFS distribute data and files over server clusters. Initially mapreduce assigned the data processing jobs to every server of the cluster for computation and after that collects the outcome of computing from different systems. Map Reduce is also provides processing of large data sets concurrently or parallely. [1],[9]

2.2. Architecture at Server Level

Architecture at server level for Big Data made up of parallel computing platforms for handling the associated volume and speed. Clusters, parallel processing and high performance computing are three options available for parallel computing. Cluster is defined as collection of different computing resources that are working with each other as a single unit of computing resource to solve large and complex problems. In Massively parallel processing (MPP), processing is done on large number of systems parallel, which has its own memory and

processor to process the data locally. All communication is done via communication network. In high performance computing, computing power is aggregated for delivering higher performance. It is used to solve complex problems like scientific research, engineering etc and these advanced problems are solved by using supercomputers and system clusters.

Hadoop architecture consists of client machine and servers that are loosely coupled. These servers provide distributed data processing i. e. Map Reduce and distributed data storage i. e. hadoop distributed file system (HDFS)[6]. In hadoop node may belong to one of the three categories i. e. client node (machines), master nodes (server node), and slave nodes.

Job of the client node is to submit the map reduce job, load cluster with relevant data and when job is finished then retrieve its results. In hadoop HDFS nodes and map reduce nodes are recognized as master nodes. In hadoop distributed file system nodes consists of name node which is responsible for storing the directory structure of all files. When client applications submit jobs to map reduce node, it concerned its job tracker. After that job tracker linked with the name node to find the location of nearest data node. When data node is found it assign the task to its corresponding task tracker that inhabit in the previously find data node. Task tracker then executes the task and submits its result to job tracker.[1][8]

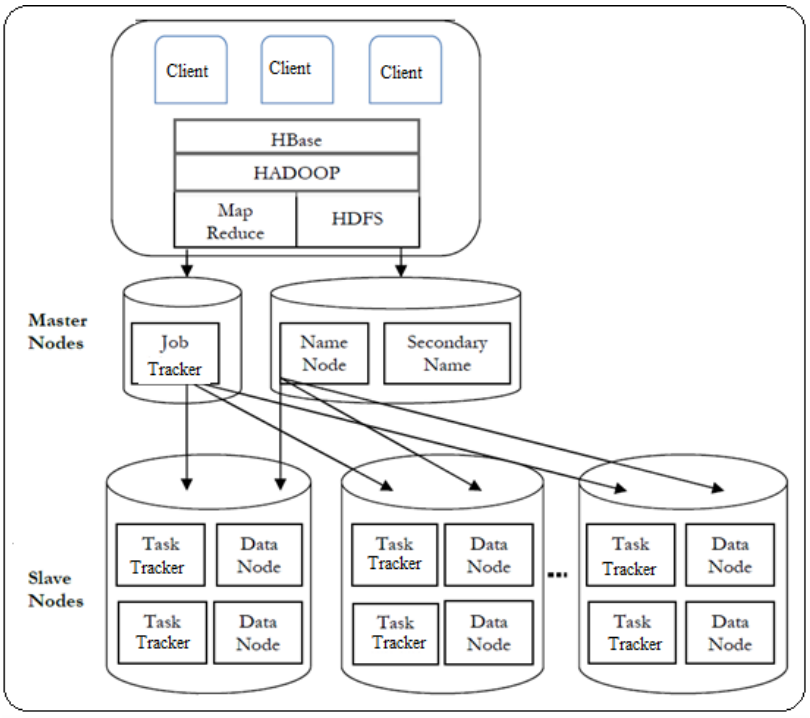

Figure 2: Cluster Architecture for Big Data

HDFS is well suited for storing large files but it is not able to provide fast lookups for every record. And to provide fast lookups on large files Hbase is rely on the top of HDFS. Therefore the responsibility of HBase is to yeild the real time random access. [7]

3. APPLICATIONS OF BIG DATA:

- Big Data is useful in understanding the behavior and preferences of the customer
- To Optimize business process by optimizing stock based on prediction generated from social sites and other sources
- Big data is used to personal quantification i. e. it is used to collect data about calorie consumption of an individual, his/her sleep pattern etc
- Big data is also useful in healthcare. For e. g. it is able to decode the DNA strings in a minute, which allow to find new cures and predict disease pattern
- Big data is useful in improving sports performance by using video analytics, that is able to track the performance of every player on the ground
- Big data is also useful in other fields like science and research, optimizing machine performance and for improving security and law enforcement.[3]

4. CONCLUSION

As big data is the recent trend in the market due to its volume, velocity and variety. For handling this data new techniques and Hadoop is one of them which use distributed processing for large computation and to analyze data for finding the hidden patterns. With the help of this paper research will be able to understand the architecture of the big data for analytics.

5. REFERENCES:

1. Joseph O. Chan, “An Architecture for Big Data Analytics”, Communications of the IIMA, Volume 13, Issue 2, 2013.

2. web_link1:https://www.mongodb.com/nosql-explained

3. web_link2: https://www.linkedin.com/pulse/20131113065157-64875646-the-awesome-ways-big-data-is-used-today-to-change-our-world, 25 October 2015.

4. Ferguson Mike. ,"White Paper: Architecting a Big Data Platform for Analytics", Intelligent Business Strategies, October, 2012.

5. Chanchal Yadav etal. , "Algorithm and approaches to handle large Data-A Survey", International Journal of Computer Science and Network, Vol 2, Issue 3, 2013.

6. Jens Dittrich etal," Efficient Big Data Processing in Hadoop MapReduce", 38th International Conference on Very Large Data Bases, Istanbul, Turkey. Proceedings of the VLDB Endowment, Vol. 5, No. 12, August2012.

7. web_link3:http://www.informit.com/articles/article.aspx?p=2253412

8. Jens Dittrich etal. , Efficient Big Data Processing in Hadoop MapReduce", The 38th International Conference on Very Large Data Bases, August 27th 31st 2012, Istanbul, Turkey. Proceedings of the VLDB Endowment, Vol. 5, No. 12.

9. Konstantin etal. , "The Hadoop Distributed File System", IEEE, 2010.