ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

A Survey on Big Data Challenges and Solutions using Real Time Streaming Protocol

M.R.Sundarakumar, Dr.Vinay Hegde

Assistant Professor, Dept. of CSE. R V College of Engineering, Bangalore, India

Associate Professor, Dept. of CSE. R V College of Engineering, Bangalore, India

ABSTRACT: Big data analytics is an emerging technology which is used to handle Petabytes and Exabyte of unstructured data in social media, business organizations, health care etc., Storage and retrieval of large data in huge databases for transaction is challenging in short duration due to their locations and size in clusters. To enhance the data processing speed in big data analytics a fast data accessing techniques tool like Hadoop is essential. Advanced optimization method like Map Reduce and Hadoop Distributed File System is inevitable to handle big data. With the use of distributed file system in a high-speed internet connection, the velocity in periodical intervals is increased so as to reduce the time to transfer data from large databases. Even though the security of big data has to be controlled by authentication principles, distributed nodes in a network will make certain delay to access big data since huge databases are connecting to different networks. In this paper the way to increase the velocity of big data on a network using Hadoop distributed file system is proposed.

KEYWORDS: Big Data, Hadoop, Map Reduce, HDFS

I.INTRODUCTION

Big data analytics is an emerging technology that has the potential to extract information from large amount of structured and unstructured data from social media. It has 3 characterization process such as: (a) extremely large volume of data in petabytes and exabytes, (b) wide varieties of data types such as video, audio, text, image, etc, (c) velocity of data at which it can be processed like batch, periodical, real time etc. The main aim of Big data in NLP is to provide small data for people after analyzing big data using machines. Big Data as a Service (BDaaS) is used by an organization to understand and gain information from large database set with statistical analysis tool. Unstructured data is generated using BDaaS which in turn is intended to free up organizational advantage of predictive analysis. In a cloud computing environment, BDaaS is a combined effect of SaaS(Software as a Service) and IaaS(Infrastructure as a Service) with its scalability and self-service. A private cloud architecture environment is used to provide a high speed data analysis with BDaaS through a proprietary architecture

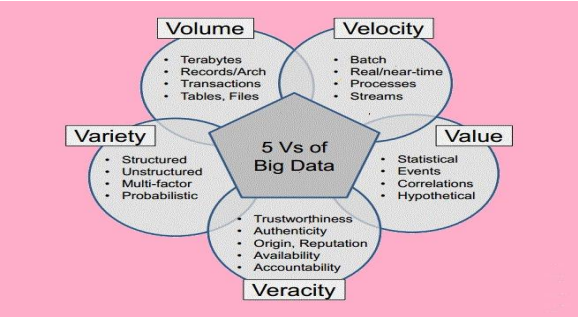

Fig:1 5V's of Big data

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

II. LITERATURE SURVEY

RDBMS (Relational Database Management System) is used to create larger databases and transform them in to different forms in earlier stage. But the increase in the size of the database ie., huge database sets, with Big data used for storing and analysis of the data merged with the Schema, increases time and cost considerably.

OLAP (Online Analytical Processing) tools are used to transform the high level databases in a dare warehousing concept which will lead the transaction processing. In data warehouse, huge amount of data processed by OLAP tool retrieve the data where the processing speed and particular time is unsatisfied.

Artificial intelligence is the human intelligence simulation processes by machines i.e. computer systems. The processes include learning, reasoning, and self-correction. Machine learning is the ability of the computers to learn without being explicitly programmed. It focused on the computer programs that can improve themselves to grow and change when exposed to new data. When this machine learning and artificial intelligence used for the extraction of unstructured data from a social media, it failed due to the formation of complex algorithm for repeated patterns.

Cloud computing environment is a phenomenon in which the distribution of the services is over the internet. For the process of data storage, data retrieval and distribution of the retrieved data to different nodes, Private Cloud is useful. In Private Cloud, the business data in the social media is distributed to the internal users. While using cloud computing, data distribution among different clusters of nodes without any query requires a tool called HADOOP. In HADOOP, uses 2 types of technologies: (a) Map-Reduced and (b) HADOOP Distributed File System (HDFS). Map-Reduced technology coordinates and combines the retrieved data from multiple resources. It is also responsible for scheduling, monitoring and executing the processed data. HDFS is used for parallel processing to overcome node failures and to distribute the Map-Reduced data among different clusters.

III CHARACTERISTICS OF BIG DATA

Big Data has some characteristics which are used to distribute the retrieved data from social network to different nodes. The characteristics are as follows:

- The ability to converge the both structured data and unstructured data from multiple sources such as social media to different clusters.
- The capability to provide realistic data extracted from data sources like mobile devices, web, sensors etc.,at expected time.

IV.MOTIVATION AND RELATED WORK

Real Time Streaming Protocol

While working with Stream SQL (Sequential Query Language), the retrieval of data is converged from the source at low speed. Hence the complex event processing occurs. So, here the future process is done with the Big Data Appliances. In this, the data is protected which is to be retrieved and distributed for different nodes. While processing, the time is reduced compared with other tools used for the distribution of retrieved data, life cycle management, cost effective. Although this, IoT (Internet of Things) can also be done which means business data and the human data are changed to machine data. By using this idea, it also removes the vulnerability of attacks with their platforms and infrastructure. The work is done through SPRAQL (secured SQP) and KERBEROS (used for plain text encryption).

IV.HADOOP: SOLUTION FOR BIG DATA PROCESSING

To perform the effective big data analytics HADOOP is a tool using Cloud computing concepts as private cloud characteristics. It can be done by Map Reduce and Hadoop Distributed File System (HDFS) concepts for processing the data from huge amount of databases in different clusters on the network. In Hadoop architecture everything deals with Master Slave method. Because Master Slave method is used to enhance the significance of Big Data by Name node with Datanodes.These nodes are accessed by the clusters on various network which helps to monitor all the big data analytics work like Extraction, Transformation and Location. If the connection is not established between the nodes then that node becomes the failure node. To overcome this problem job tracer and Task

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

tracer is used to update the current position of each and every nodes on the network while it is connected in private cloud environment.

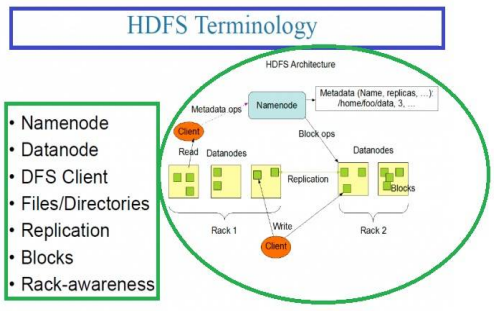

Fig:2 HDFS Architecture

VI.HADOOP DISTRIBUTED FILE SYSTEM

Hadoop Distributed File System is used to distribute the map reduced file among the different clusters on various network which are connected to the private cloud. The speed of the data which are processed with HDFS concepts is more compared with the existing SQL streaming technologies. When unstructured data has to be accessed with HDFS, the velocity of data will improve while retrieving from huge database.

VII. MASTER SLAVE ARCHITECTURE IN HADOOP

When Master Slave technology is used in HADOOP to find the failure nodes in the cluster with their accessing data types from the databases, it has to reduce the time taken to search the data failures during data transformation between two different networks.

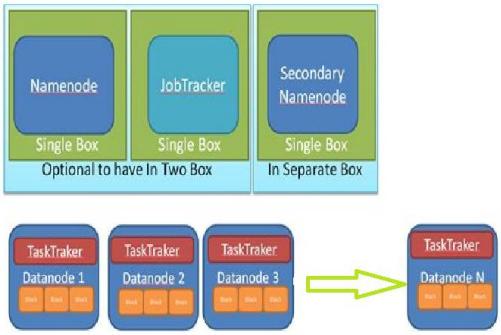

Fig:3 Master Slave Method

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

VIII.CONCLUSION

Big data has a better scope in big industries, institutions and social media's for faster data communication with periodic interval speed. The advantage of big data will be developed using various tools like Hadoop, Cloudera etc., which will help big data analytics faster and reliable than other methods. Most of the technologies of big data conserve latest terminologies in the world which will develop the real time usage of every day activities

REFERENCES

1. Puneet Singh Duggal," Big *Data Analysis: Challenges and Solutions"- International Conference on Cloud, Big Data and Trust 2013, Nov 13-15,*
2 Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, "A Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 1 ISSN 2250-3153
3. Sabia," Technologies to Handle Big Data:", International Conference on Communication, Computing & Systems (ICCCS–2014)
4 Sangeeta Bansal, Dr. Ajay Rana," Transitioning from Relational Databases to Big Data", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014 ISSN: 2277 128X
5 Jefry Dean and Sanjay Ghemwat, MapReduce:A Flexible Data Processing Tool, Communications of the ACM, Volume 53, Issuse.1,January 2010, pp 72-77.
6 Brad Brown, Michael Chui, and James Manyika, Are you ready for the era of „big data"?,McKinseyQuaterly,Mckinsey Global Institute, October 2011
7 DunrenChe, MejdlSafran, and ZhiyongPeng, From Big Data to Big Data Mining: Challenges, Issues, and Opportunities, DASFAA Workshops 2013, LNCS 7827, pp. 1–15, 2013
8 Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Big data using Apache Hadoop and Map Reduce" Volume 4, Issue 5, May 2014" 27
9 Kyong-Ha Lee Hyunsik Choi "Parallel Data Processing with MapReduce: A Survey" SIGMOD Record, December 2011 (Vol. 40, No. 4)
10 Apache: Apache Hadoop, http://hadoop.apache.org