ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 2, February 2019

Study of Mobile, Distributed Agent Framework to Enhance Big Data Analysis

[1]Dr. Mallikarjun M Math, [2]Vinita Tapaskar
[1] Professor, Department of Computer Science and Engineering, KLE, GIT Belgaum
[2]Vinita Tapaskar , Research Scholar, VTU, RRC The Oxford College of Science, Bangalore

***Abstract* - There are large numbers of application getting developed and many users are getting benefited by growth in computational industry. The growth of data due to these application usages is giving rise to the concept of Big Data. Big Data is combination of structured and unstructured data which has to process efficiently to derive patterns which are useful to gain business intelligence. Initially HADOOP was bought by Apache open source software for distributed processing of large data sets. However HADOOP had drawbacks on reliability and performance. To remove HADOOP drawback and to provide efficient framework Map Reduce Agent Mobility framework is proposed which is based on Map Reduce Algorithm and Java Agent Development Framework (JADE). This paper is intended to study these frameworks and provide comparative study on HADOOP and JADE with MRAM framework**

Keywords: Mobile Agents, Big Data Analysis, JADE, MRAM, HADOOP

I. INTRODUCTION

The term "Big Data" is used to for the collection of complex and large data set which is difficult to capture, process, store, search and analyze by using conventional database management system. Big data comes with 3 important features which are called as 3Vs. They are Volume, Velocity, and Variety. Big data analytics is processing large and varied data sets to identify hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations in decision support system[1].
Big data will any one of these forms:

- Structured Data : Relational data.
- Semi Structured Data : XML data.
- Unstructured Data : Word, PDF, Text, Media Logs.

To process and manage the large volume of unstructured data initially the traditional database where used where the user with the help of application server the data will be accessed from traditional database server.

But due to volume of data traditional database failed. Google came up with the solution in the form of algorithm called Map Reduce which divides the task into parts and assigns the task to many computers connected through the network and collects the result set back and merge it.

There are two types of technologies which are in the market to handle big data [4]

- **Operational Big Data**

This includes systems like MongoDB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored.
NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement.
Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

- **Analytical Big Data**

This includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data.
MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 2, February 2019

from single servers to thousands of high and low end machines.

Apache came up with the open source solution called HADOOP. Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for huge amounts of data.

HADOOP also suffered drawbacks where the solution is provided in the form of Map Reduce Agent Mobility framework based on JADE.

II. HADOOP

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Architecture of HADOOP:

The idea of Hadoop is to provide a **cost-efficient High Performance Computing using the cloud infrastructure**

- Hadoop is designed to run on a large number of machines that don't share any memory or disks.
- That means you can buy a whole bunch of commodity servers, slap them in a rack, and run the Hadoop software on each one.
- When you want to load all of your organization's data into Hadoop, what the software does is bust that data into pieces that it then spreads across your different servers.
- There's no one place where you go to talk to all of your data; Hadoop keeps track of where the data resides. And because there are multiple copy stores, data stored on a server that goes offline or dies can be automatically replicated from a known good copy.
- Architecturally, the reason you're able to deal with lots of data is because Hadoop spreads it out.
- And the reason you're able to ask complicated computational questions is because you've got all of these processors, working in parallel, harnessed together.

Components of Hadoop:

The current Apache Hadoop ecosystem consists of the Hadoop kernel, MapReduce, Hadoop distributed file system (HDFS),a number of related projects such as Apache Hive, HBase and Zookeeper as shown in figure 1. MapReduce and Hadoop distributed file system (HDFS) are the main components of Hadoop[3][2].

- MapReduce:

The framework that understands and assigns work to the nodes in a cluster

- Hadoop distributed file system (HDFS):

HDFS is the file system that spans all the nodes in a Hadoop cluster for data storage. It links together the file systems on many local nodes to make them into one big file system. HDFS assumes nodes will fail, so it achieves reliability by replicating data across multiple nodes.

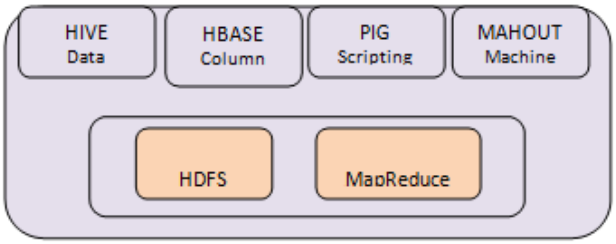

Figure 1: HADOOP Architecture

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 2, February 2019

Drawbacks of HADOOP

Major drawbacks of HADOOP system in big data analysis are:

- Hadoop needs high storage and memory to implement replication of data to provide fault tolerance.
- Hadoop can allocate task but do not provide any mechanism to schedule the task
- Single Master
- Task loading time of Hadoop system is more as allocation requires time

Other Drawbacks of Hadoop System are:

- Security Issues

Managing the complex application of HADOOP is challenging. The security model of it is disabled because of complexity issue which creates a issue for the security of the data. Hadoop even do not provide encryption at the storage and network level.

- Weak By Nature

Hadoop framework is written in java one of the largely used programming language. Java has been exploited by cyber criminals and as a result , as a result there can be lot of security breach which can happen.

- Not used for Small Data

All big data framework are suited for managing the small data needs. Hadoop is one of them. Hadoop has high capacity design, HDFS which affect the efficiency of reading of small files. Hadoop is not suited for organization with small amount of data.

- Potential Stability Problem

Hadoop is created by many developers contribution and constant improvement on it will be done and new versions are released. So it is recommended to make sure that organizations are used the latest stable version.

- Some Other Limitations

For most of the big data processing needs Hadoop is not only the solution. Apache Flume, MillWheel platform have ability improve processing efficiency and provide reliability of data collection, aggregation and integration which are missing by using only Hadoop framework.

Architecture of JADE

Java Agent Development Framework (JADE) is software in JAVA, which is used for the development of distributed multi-agent applications.

Additionally the platform has various debugging tools, mobility of code and content agents, the possibility of parallel execution of the behavior of agents[5][6].

The JADE platform is open source software. The internal architecture of JADE is shown in Figure 2

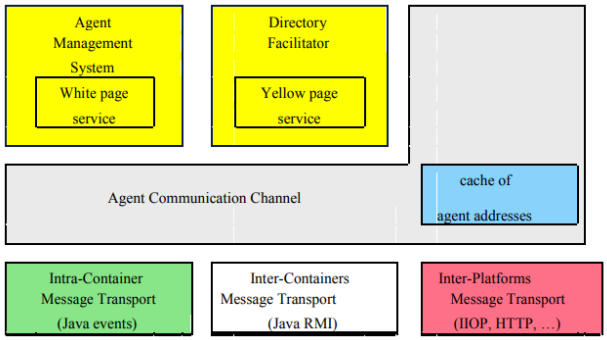

Figure 2: JADE Internal Architecture

A JADE platform comprises of agent containers that can be distributed over the network. Agents live in containers which are the Java process that provides the JADE run-time and all the services required for hosting and executing agents.

There is a special container, called the main container, which represents the bootstrap point of a platform: it is the first container to be launched and all other containers must join to a main container by registering with it. The programmer identifies containers by simply using a logical name by default the main container is named 'Main Container' while the others are named 'Container-1', 'Container-2', etc. Command-line options are available to override default names as shown in Figure 3.

The Responsibility of the main container is managing the container table, which is the registry of the object references and transport addresses of all container nodes composing the platform and managing the agent

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 2, February 2019

descriptor table, which is the registry of all agents present in the platform, including their current status and location

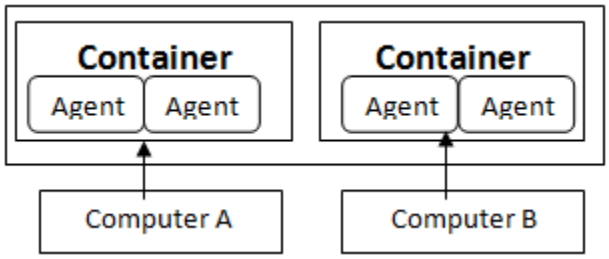

Figure 3: Container's and Agent

Mobile Agent:

- Hadoop provides fault tolerance through replication of data.

MRAM provides fault tolerance when machine is failed dynamically the system change happens the new agent with data and code moves to different machine and continue executing the task which is left out.
Each machine in the framework keep sending a copy of data and status to master machine after fixed interval period of time. In case of failure another copy of agent sent from master machine to new machine which carries data and code and completes the task.

- Following steps explain how the MRAM framework will function the same is also shown in figure 4.

1. Input text files which contain data to the MRAM Platform. Server will partition the data file with same size blocks and then application server will assign the data block to each computing machine.
2. Computing node applies map function of the block of data and produces intermediate data pair which is sent to application server the reduce operation
3. The reduce operation counts the summarization and saves the result for outputting purpose.

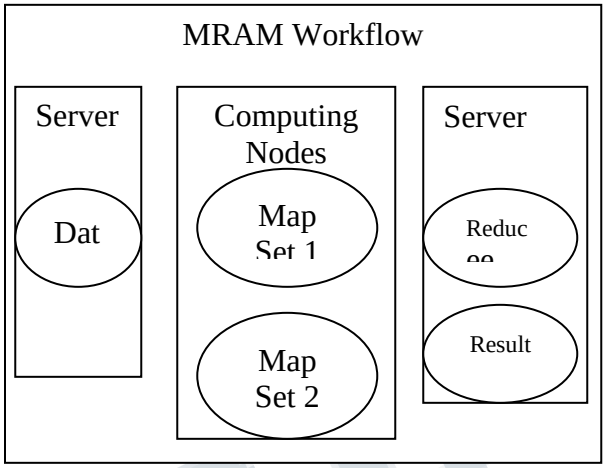

Figure 4: MRAM Workflow

- Mobile agent can react dynamically and autonomously to change in their environment.

In Hadoop Framework there is still dependency on single node that runs all the services needed to Map Reduce task distribution and tracking. The entire system is down when a single node is failed or down.
The solution of this problem is:

1. The master node is selected when a platform starts working. After that, the master node build linked list involves meta-data. These meta-data contains all information about tasks, dependences among them and information about all machines. Also, the master machine sends meta-data to all machines through network connection.
2. If any node receives a job, this node is elected as a new master.
3. When the master machine is shutdown or platform is down, all agents executing on master machine will be moved to another host that having a highest IP-address when meta-data is published. The agents continue executing tasks on new machine because it carrying its code, status and data.
4. A new machine becomes as a master node that is responsible for all acts of server expects to receive result from machines such as Task Tracker and informs all machines about a machine failed.

5. After all agents finished executing tasks, it is waiting to send the result to general server when it comes back online.[7]

Advantages of MRAM:

- MRAM support allocation and scheduling of tasks.
- MRAM provides fault tolerance and don't need high memory.
- Load time for MRAM is less than that of Hadoop.
- MRAM solves single master problem by using mobile agent.
- MRAM reduces execution time because of no need to huge processing for replication of data [7].

III. RESULTS AND DISCUSSION

Comparison of HADOOP and MRAM:

1. Architecture: Hadoop has Client Server Based - File system where as MRAM works on Distributed Multi agent framework
2. Startup Time: Less startup time for MRAM compare to Hadoop
3. Components :Map Reduce , HDFS are the modules used in HADOOP where as JADE uses agent management system and Directory Facilitator
4. Failure Management: In HADOOP if the main node failure system will shutdown but in case of MRAM failure of node or master node will not affect the system working
5. Application: HADOOP is used in Machine Learning application development and JADE is used in E-Commerce based application development
6. Mobility in the Network: HADOOP do not support mobility of components where as MRAM agent provide mobility in the network and can move around the network
7. Allocation and Scheduling Task: Hadoop does allocation of task but do not have feature of scheduling where as MRAM does allocation and scheduling activity.
8. Performance and Reliability : In comparison the MRAM has better performance and reliability than HADOOP in Big Data Analysis

REFERENCES

[1] Mobile Agent Based New Framework for Improving Big Data Analysis Journal: Cloud Computing and Big Data (CloudCom-Asia), 2013 International Conference on Dec. 2013

[2] Hadoop web site, http://hadoop. apache. org/, Sep2013.

[3] Kala Karun. A, Chitharanjan. K, "A Review on Hadoop-HDFS Infrastructure Extensions", In Proceedings of IEEE Conference on Information and Communication Technologies (ICT2013), pp. 132-137, 2013.

[4]https://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm

[5]file:///C:/Users/Rao/Desktop/10.1.1.819.4796.pdf

[6] International Journal of Research (IJR) Vol-1, Issue-9, October 2014 ISSN 2348-6848 Java Agent Development Framework Sujeet Kumar & Utkarsh Kumar

[7] New Framework For Improving Big Data Analysis Using Mobile Agent International Journal of Advanced Computer Science and Applications, Vol. 5, No. 3, 2014 Youssef M. ESSA, Gamal ATTIYA and Ayman EL-SAYED