International Journal of Advances in Computer Science & Its Applications
Volume 6 : Issue 3 [ISSN 2250-3765]
Publication Date : 30 December, 2016

Processing of Big Data with Hadoop Environment in the Distributed System

Thurein Kyaw Lwin (PhD), Professor A.V.Bogdanov.

***Abstract*— In this paper, we would like to examine about cloud computing, Big data and Hadoop environment. The main focus of issues is big data with Hadoop are associated in distributed system. Applications of Big data are benefit to organizations of many large and small industries. In this article, we introduced with Hadoop analytics and cloud computing solutions, application, and testing real time in distributed system. We discussed about the various possible solutions of cloud computing and Hadoop for lage data sets. Cloud computing is a very vital role in database system, applications and the related infrastructure with technologies, controls and the big data tools. Bigdata applications and Cloud computing advantages are represented to the most important in distributed computing process.**

Keywords- Cloud Computing, Big Data, Analytics, Algorithms, Tactical Environment, Hadoop, Hadoop Distributed File System (HDFS) HDFS, Virtual server, JAVA, Map Reduce.

I. INTRODUCTION

In This article, we presented about the large data is being collected at the unprecedented rates to a wide range of high-resolution and high throughput sensors, it has also become a suitable algorithms and tools to analyze Big Data are largely missing. We believe that the Cloud computing involves extensive complexity rather than the providing solution to cloud processing, it would be ideal to make in the cloud and big data. Using Hadoop distributed file system is evolving as a software component for cloud computing with integrated parts such as MapReduce. In this paper, we were using with some approaches to providing of the data processing. Hadoop's framework can use to solve the problems and data managed conveniently by using different types of techniques with data processing technology. Hadoop Provides parallel processing to manage bigdata and it overcomes many of the above stated challenge of bigdata processing.

Thurein Kyaw Lwin (PhD)

St.Petersburg State University, Saint-Petersburg, Russia

Professor: Alexander Bogdanov

St.Petersburg State University, Saint-Petersburg, Russia

II. CLOUD COMPUTING

We can define the cloud computing is IT foundation for cloud services and it consists of technologies that enable cloud services. Cloud services and solutions are delivered and consumed in real-time over the Internet, while Cloud Computing is an emerging IT development and delivery model, enabling real-time delivery products, services and solutions over the Internet[3,4]]. The cloud computing is bringing together multiple computers and servers in a single environment designed to address certain types of tasks, such as scientific problems or complex calculations. This structure builds up a lot of data, distributed computing nodes and storage. Cloud Computing consists of a front end and back end. The front end contains the user's computer and software required to the cloud network. Back end contains various computers, servers and database systems are created the cloud.

III. BIGDATA

Bigdata described massive volumes of structured and unstructured data that are so large data sets, it is very difficult to process with traditional databases and software technologies. Collection storaged and processing of Bigdata in heterogeneous distributed computer networks are focused of different methodologies of data collection, storage and processing. In the cloud, it related areas to massive stored data and from the Internet and how the data is stored and utilized within distributed systems of enterprise storage[1,2]. Bigdata have to query loosely structured and very large distributed data.

IV. DISTRIBUTED DATA PROCESSING WITH HADOOP ENVIRONMENT

Hadoop is a free programming based on java technology and framework supports the processing of large data sets in a distributed computing. Hadoop cluster uses a Master (or) Slave structure. The large data sets can be processed of Hadoop across a cluster of servers and applications can be run on systems with thousands of nodes with so many terabytes. Hadoop Distributed file system helps in rapid data transfer rates and allows the system to continue, it is normal operation even in the case of some node failures. This approach can lower the risk of an entire system failure, even in the case of a significant number of node failures[4]. A computing solution of Hadoop is scalable to cost effective, flexible and fault tolerant. Framework of Hadoop is used by popular organization to support their applications involving huge amounts of data[1]. Hadoop has two main sub projects are Map Reduce and Hadoop Distributed File System (HDFS).

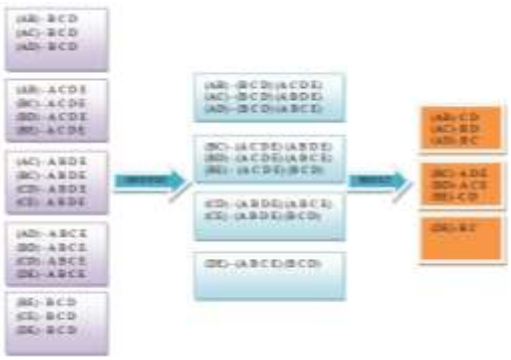

Fig.1. Reducing Process of our Example (collection mutual letter)

In this article we used the distributed system and there is support for Java technology and we added the required functionality before start. We worked with the operating system Linux for Hadoop, procedure allows to get ready for the binary package and eliminates that need to download and compile a similar code . At First, we created a new file and than we installed Hadoop on Linux Server. We set up password less SSH and some aspects of security, processing will be fully recovered by the work of Hadoop.

We tested Hadoop to running on our linux server in our Research Center and just need to run all of his demons tool. But at first we need to format the Hadoop File System (HDFS) is using with hadoop command. The NameNode node is requesting to format HDFS file system. This is the part of the installation, but it's very useful to create a file clean system.

Hadoop - 0.20 namenode –format

After confirmation file and this system is formatted, as well as some of the information will be return. In this configuration have five demons Hadoop: namenode, secondarynamenode, datanode, jobtracker and tasktracker. At the start of each demon, we can see a few lines of text (indicating where to store log-files). The daemon is running in the background and how a node is in the configuration after starting Hadoop.

Hadoop has several supporting tools to simplify the start-up. These tools are divided into two categories : the launch (for example, start-dfs) and stop (stop-dfs). The following small script shows how to run the Hadoop node.

```
# /usr/lib/hadoop-0.20/bin/start-dfs.sh
# /usr/lib/hadoop-0.20/bin/start-mapred.sh
```

To verify that all the daemons are running and we can use the jps, which can bring a list of five demons and their associated process identifiers. Demon namenode of major Hadoop server and manages file system namespace are stored on a cluster. There are also a daemon secondary namenode, which is not redundant with special namenode, but instead performs the control points and keeping the other supporting tasks. We can find one node type in the Hadoop cluster and the NameNode is one node type secondary namenode. The datanode manages storage connected to a single cluster, on each node, the data stored will always be running the daemon datanode.

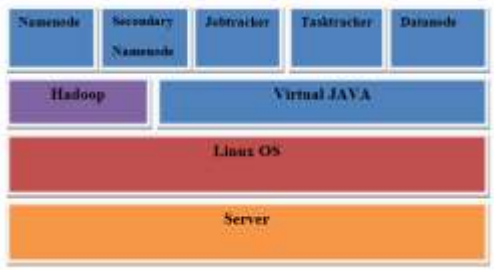

Fig.2. Hadoop Distributed configuration in Saint Petersburg State University Research Center

Finally, in each cluster, one instance jobtracker daemon is started, which is responsible for the distribution of tasks between DataNode nodes and one copy tasktracker daemon for each node of DataNode. Demons JobTracker and work in a "master-slave", a jobtracker datanode distributes tasks among the nodes and performs tasktracker received a job. JobTracker checks the results of the scheduled tasks and if the node datanode for some reason fails, the unfinished task is reassigned to another node. In our configuration design is we tested in our University Research Center (SPBSU), all nodes are on the same physical host (Figure 2). However, in this articles we shown how Hadoop provides parallel processing tasks. Although the architecture is quitely simple and the Hadoop provides a reliable, fail-safe way to implement data distribution, load balancing and parallel processing of large data volumes of our future databse system.

V. HADOOP MAP-REDUCE PERFORMANCE TESTING IN OUR LINUX SERVER

We installed Hadoop and learned about the basics of working with file system in a real time application in our server. We used one server for the standalone operation of Hadoop and the others for the fully-distributed operation. We separated the name node and the data node in the fully-distributed operation of Hadoop. This separation was necessary because the Job Tracker that executes Reduce task and the main controller run on the name node, while the Task Tracker that actually executes the Map task runs on the data node. In this article, we tested, how the MapReduce process with a small amount of data. The names of the operations map and reduce are taken from the names of the respective operations in functional programming and provide basic functionality for data processing and its in order to reduce their volume. The map operation means the process of breaking down the input data into smaller subsets for subsequent processing. Reduce operation means the process of linking from the processors into a single set of output data. The processing as a framework allows to define itself and

MapReduce counts the number of words in a set of documents.

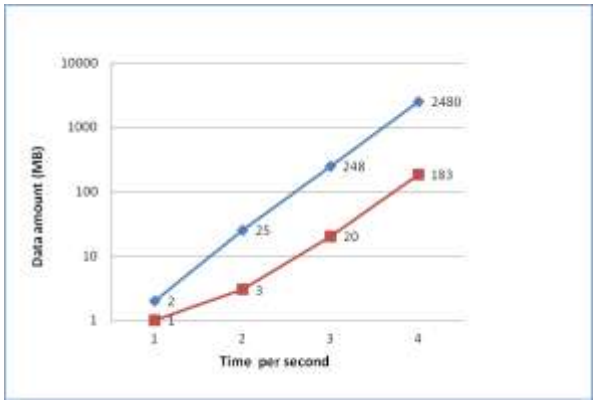

Fig.3. Result of MapReduce Performance in Our Linux Server

The Fully-distributed operation is good for processing large data. Processing small data with the fully-distributed operation is undesirable because the time it takes to collect distributed data in the same nodes during a Reduce operation outweighs the advantages of distributing data to each node using Map. In this test, the fully-distributed operation of 10 million rows of data outperformed other test conditions. The test result varies depending on the amount of data and the system specification for the future of Big data and data distributed system.

VI. CONCLUSION

Summarizes the database requirements for cloud databases and compares the suitability of different database architectures to cloud computing. Whether we need to assembling, managing or developing on a cloud computing platform and need a cloud-compatible database. The newest wave of big data is generating new opportunities and new challenges for businesses across every industry. The challenge of data integration incorporating data from social media and other unstructured data into a traditional environment is one of the most urgent issues for IT managers[5,6]. Apache Hadoop provides a cost-effective and massively scalable platform for ingesting big data and preparing it for analysis. In This article we shown the initial setup simple on Hadoop cluster. When we need to create a big data processing with Hadoop, we need to scale the Hadoop cluster and hardware resources need to enough[1,7]. This is popularity of Hadoop by the fact that we can easily run it in the cloud computing infrastructure on servers with Hadoop

VMs. In this article, it is easy to see how Hadoop distributed computing makes it easier to handle large data sets for Bigdata Processing.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. A.B.Vogdanov (Saint Petersburg State University) and Research Center of Saint Petersburg State University, Russia. for the continuous support of this article and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this article. I could not have imagined having a better advisor and mentor for my DsC study.

REFERENCES

[1] Venkata Narasimha Inukollu , Sailaja Arsi and Srinivasa Rao Ravuri, Security Issues Associated with Big Data In Cloud Computing// International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014.

[2] Bogdanova, A. V., Thurein Kyaw Lwin, Storage database in cloud processing.// КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ И МОДЕЛИРОВАНИЕ// Computer Research and Modeling,// Computer 7, no. 3 (2015): 493-498.

[3] A.V.Bogdanov, Thurei Kyaw Lwin, Ye Myint Naing. Database Used for Consolidation of Cloud Computing [текст] (База данных используется для консолидации облачении вычислении) // CSIT-2011 International Conference. (Armenia, September 26-30,2011). p-237-239.

[4] Thurein Kyaw Lwin, Prof. A.V.Bogdanov., Consolidation Technology in the Cloud Data Processing.// ISBN 978-93-84422-37-0., 2015 International Conference on advances in Software, Control and Mechanical Engineering (ICSCME'2015)., p-72-76.

[5] Hao, Chen, and Ying Qiao. "Research of Cloud Computing based on the Hadoop platform." Chengdu, China: 2011, pp. 181 – 184, 21-23 Oct 2011.

[6] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.

[7] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications.". Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct 2012.

Dr.Thurein Kyaw Lwin (PhD)

Saint Petersburg State University,

Russia.

Researcher of SPBSU Scientific Research Center. Russia.