Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 9, September 2014, pg.599 – 606

RESEARCH ARTICLE

A Study of Data Management Technology for Handling Big Data

Amrit Pal[1], Dr. Sanjay Agrawal[2]

[1]Department of Computer Engineering and Application NITTTR Bhopal, India

[2]Prof, Department of Computer Engineering and Application NITTTR Bhopal, India

[1] er.amritkabir@gmail.com; [2] sagrawal@nitttrbpl.ac.in

Abstract— The amount of data is increasing daily. Data requires storage and effective processing for information retrieval. These both are challenge in case of the BigData due its velocity, variety and volume. It requires different management and efficient information retrieval schemes. There are different techniques available for the management of the Bigdata. The distribution of the storage and the processing power provides large gain in the storage and processing of the data. There are different tools available which use type's techniques for storage and processing of the Bigdata. In this paper we have analyzed different tools for handling the Bigdata. This paper provides a study of the setting of Bigdata processing and storage environment using different tools.

Keywords— Hadoop, Hortonworks, Cloudera, Map Reduce, Virtual image

I. INTRODUCTION

Data are tokens which can be interpreted or converted in some kind of information or values. These values can be quantitative or can be qualitative. Further these can be interpreted as qualitative and quantitative facts. Data can be converted into values or variables then interpret them into some information [1]. Data is what is stored for future use or for getting some information from it. This information can be relevant information and can be a irrelevant information. Data is there from starting of the information system, wherever the state of data is changed throughout this evolution. Data can be in form of structured, unstructured or can be in the form of the semi structured form. When the amount of data when it becomes in that much amount that it cannot be handled by the conventional database management system is called the Bigdata or it is difficult to capture cure and process the data then it is big data. Big Data term is related with the datasets which are in large size that they cannot be managed with the help of conventional database management system. The amount of data is increasing exponentially [2]. This data can be logs, sensors data, and scientific data or can be a data stream. According to the Gartner's definition big data can be represented with three V. These 3Vs are the three properties of the big data.

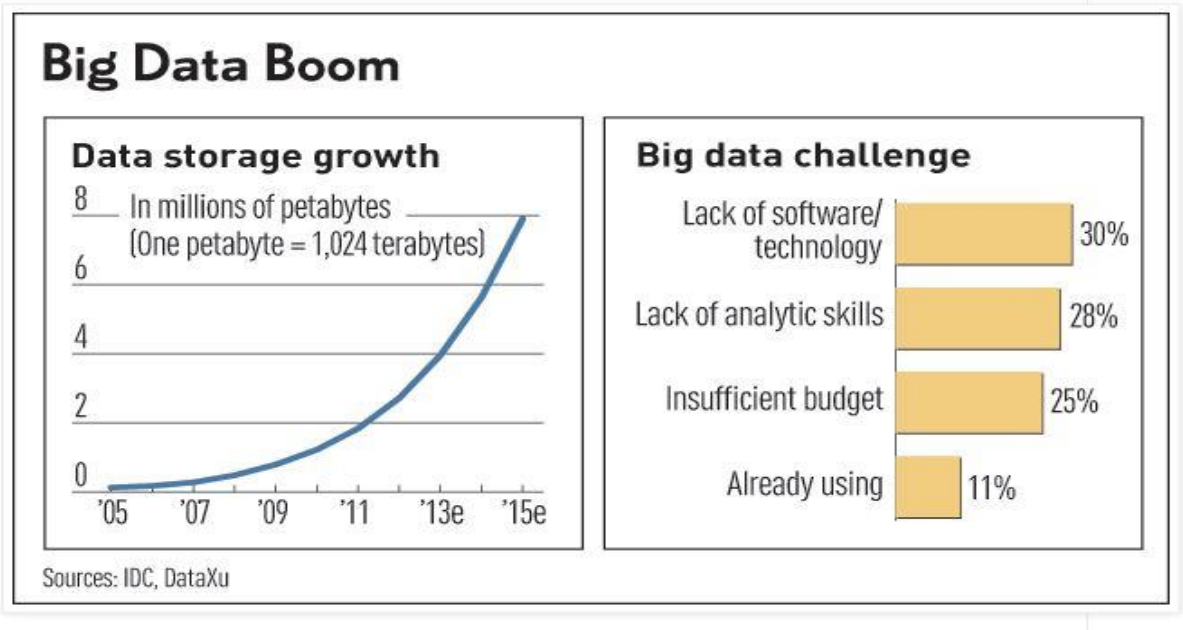

Figure 1. Big Data Growth and Challenges [3]

Figure 1 shows the growth of Big Data and also the percentage of challenges in different aspects of the Big Data [3].

II. RELATED WORK

There are number of BigData solution Available. Hadoop Distributed file system provides a distributed storage for the storage of the data in a distributed manner. Map reduce provides the distributed processing if the data [4]. Hortonworks uses this concept of for handling the Big Data [5]. Study shows Big Data can be used in different area like healthcare [6] education sector etc, it requires technological development. The cluster for handling Big data are setup by big vendors which provides data analytics services. In this paper we are concentrating toward setting the cloud processing environment using different tools and techniques.

III. BIG DATA HANDLING

There are different technologies available in the field of big data handling. In this paper setup different big data solution is done and the complexity in there setup is discussed. There are some issues in the setup of one technology with respect to other technology. All these issues are discussed and setup of the big data solution is done.

A. Hadoop

Hadoop is an Apache project. Hadoop have many more projects under its umbrella. Mainly there are two main aspects of the Bigdata.

- The storage of the data
- The retrieval of information and the processing of the data.

These two requirements are efficiently handled by the Hadoop. The information retrieval is handled by the map reduce and the storage of the data is handled by the Hadoop distributed file system. The Hadoop works efficiently in clustered environment. The development of the Hadoop cluster requires nodes connected by the communication channel like a simple LANs. There are two type of the installation that be done in case of the Hadoop, the single node and the multimode architecture of the Hadoop cluster. The main component of the Hadoop installation of is name node, data node, secondary name node, job tracker, task tracker. In

case of the single node all these resides on a single node and in case of the multimode case these are on different nodes. The specification should be done in different file for configuring the Hadoop cluster. The source code for the Hadoop is available at Hadoop website. The configuration of the Hadoop cluster requires setting in the files like core-site.xml, mapred-site.xml, hadoop-env.sh and also in the bash file of the system. Hadoop can be run on Linux operating system. The source code contains the setting for both the Hadoop distributed file system and for the map reduce. We have done the configuration on the Ubuntu system.

B. Hortonworks

Hortonworks provide there solution for dealing with the big data. They provide Hadoop for big data management. It is started in 2011 by twenty four engineers. The best thing about the Horton works is that they provide you with a fully functional Hadoop big data solution which is packed in a virtual machine disk. For using it we just need to add this to the virtual hypervisor tool which we are using and run this machine it will do all task by itself. Examples are Oracle virtual box. Oracle virtual box is freely available for download on the oracle website. It is easy to use and install. We show the screen shot below which provides steps for adding the disk to the virtual box. We have used Hortonworks_Sandbox_2.0_VirtualBox. It is an .ova extension file. It can be downloaded from [7] Hortonworks website easily. It is freely available. We have used a dell system with, 4GB RAM, Intel Xeon, 64 bit window 7 system with 1 TB disk space. When it is downloaded we need to add it in the virtual box. For this we need to create a virtual machine and in the settings at storage tab set the path of the .ova image file that is downloaded from the Hortonworks website. The figure 2 below shows the setup. This also shows the size of the disk can be expandable to the 50 GB and this expansion takes place as the data.

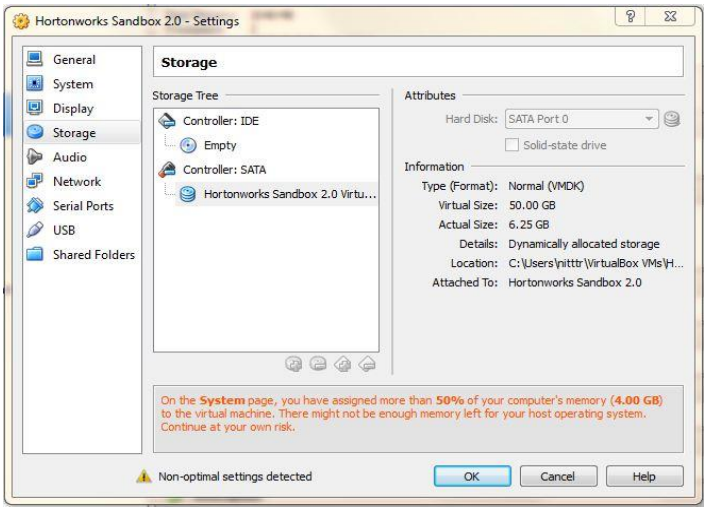

Figure 2 Virtual Box Setup

It uses Centos platform the starting screen of the virtual machine output is shown in the figure3.

There is an issue of secured login in the Hadoop. The apache Hadoop also provide access to different secured login techniques. These techniques should be separately implemented and configured on the Hadoop setup.

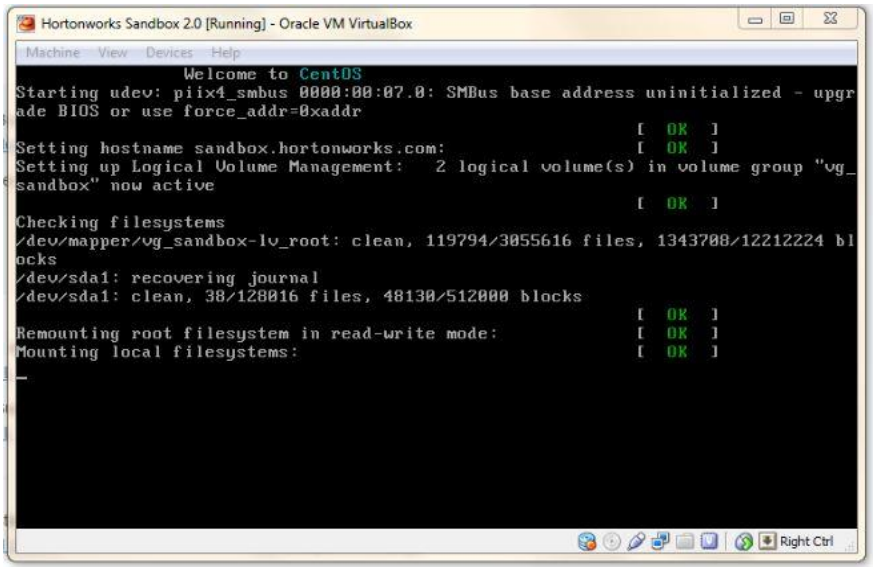

Figure 3 Centos Booting

Hortonworks also provides this kind of technique. Hortonworks uses SSH for secured login. The virtual machine automatically sets the secured login for the database. When we run this virtual machine the output as shown in the figure 4 comes up on the screen. The virtual machine is running and all the setup are automatically done in this virtual machine the main advantage of using the Hortonworks is that we don't need to go through all the steps for setting a hadoop node. It does all those automatically or they all are already done in the virtual machine. SSH request for this host can be done using ssh@127.0.0.1 –p 2222.

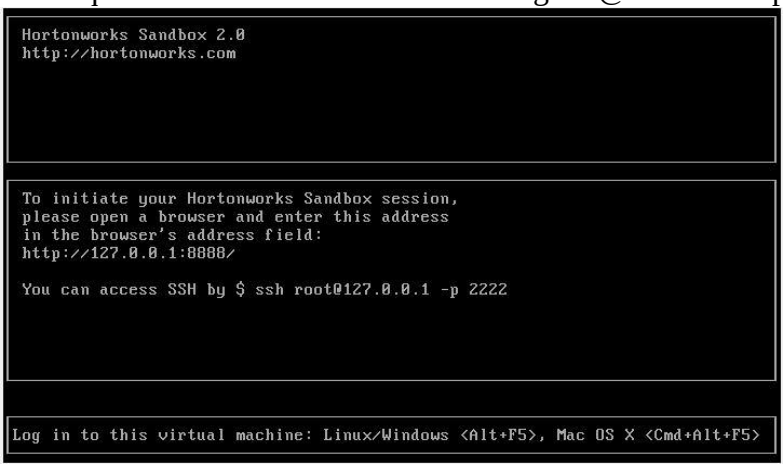

Figure 4 Runnig Output

We can login to the centos operating system on the virtual machine. The default username and password is root and password respectively. The Hortonworks interface is available for use on the main operating system browser. We can access it from the browser of the host machine with http://127.0.0.1:8888/. With this we can access the GUI provided by the Hortonworks. The package contains the interface for the Hive, Pig, HCat, File browsing etc. The documentation of the Hortonworks provides an example study of the temperature data. The weather dataset are freely available on the internet. The screen shot is shown in figure 5. Diagram shows the interface for the pig.

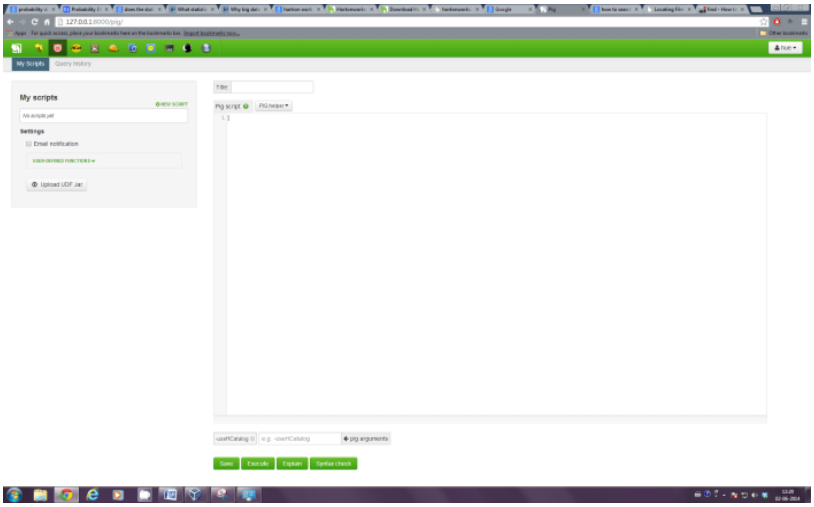

Figure 5 GUI Interface

C. *Cloudera:*

Cloudera also provide big data solution. It also uses the Hadoop for big data management. Cloudera also saves from the difficult installation of Hadoop on a node. It provides the virtual disk as well as a package repository which will install all packages for the Cloudera working. The Cloudera manager binary can be downloaded from [8] there website. It provides a cloudera-manager-installer.bin file. The Cloudera is available for the Linux platform like RHEL, Ubuntu, and Centos etc. The installer script or the .bin file is executed and it will install all the required packages for the Cloudera to work. According to the Cloudera documentation [9] the host on which the Cloudera is to be installed must be having at least 10 GB of RAM. The internet connection is required for the installation of the packages.

For installing the Cloudera we should have the root privilege to install the Cloudera. The commands used for installation are:

First change the file and make it executable by providing full 777 privilege to it by using

Sudo chmod 777 cloudera-manager-installer.bin

Then execute the following command

Sudo ./cloudera-manager-installer.bin

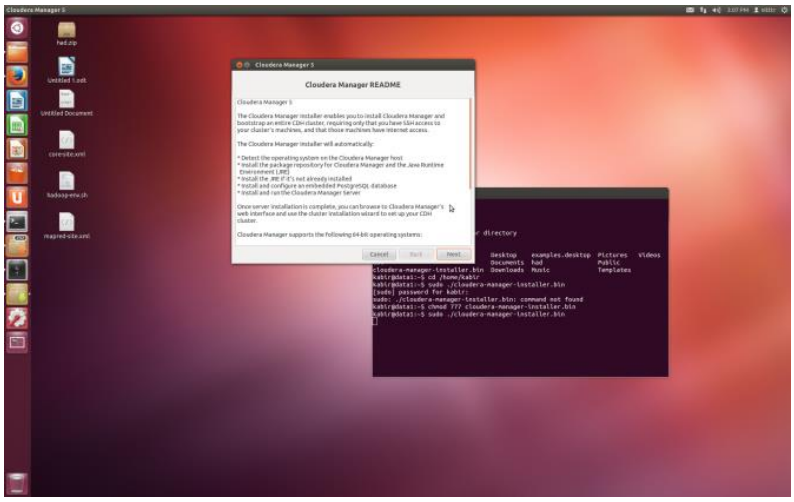

Figure 6 Cloudera Installations 1.

This is the way to install the installation process for the Cloudera big data solution. Figure 6 shows the starting of the installation process. Next it will ask for the licence. As it requires the internet connection for its installation it will start downloading the packages from the internet. Figure 7 shows the process of downloading the packages from the internet or from the repositories for its installation.

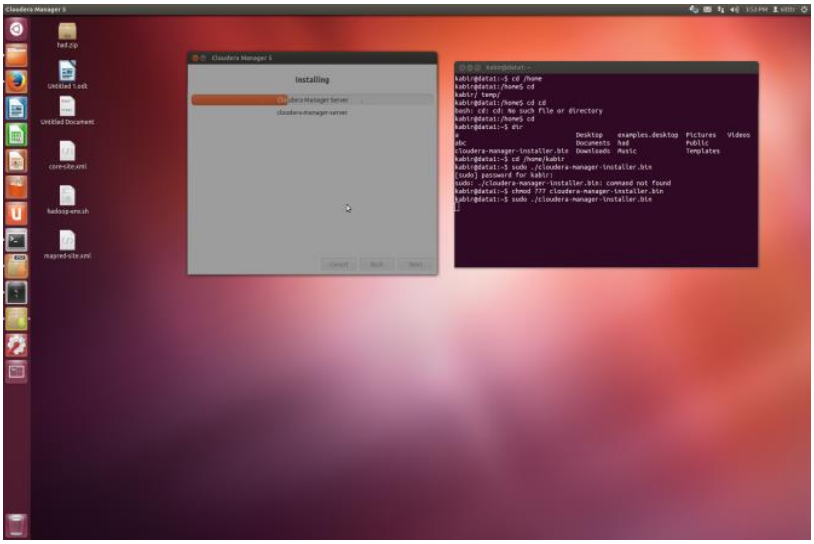

Figure 7 Cloudera Installation 2

As the installation of the packages and the repository is done, then the URL which can be used for access of the interface is provided after the installation. Diagram below shows that it is

http://localhost:7180 /

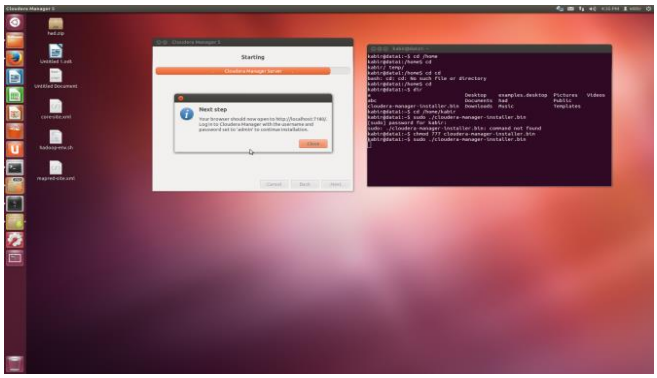

Figure 8 Licence Agreement

The licence agreement should be done after this step. There is availability of different types of licence which are available. The user should chose one of them then continue for the installation. The IP address for the DataNode should be specified. In our case it is 172.16.13.194.

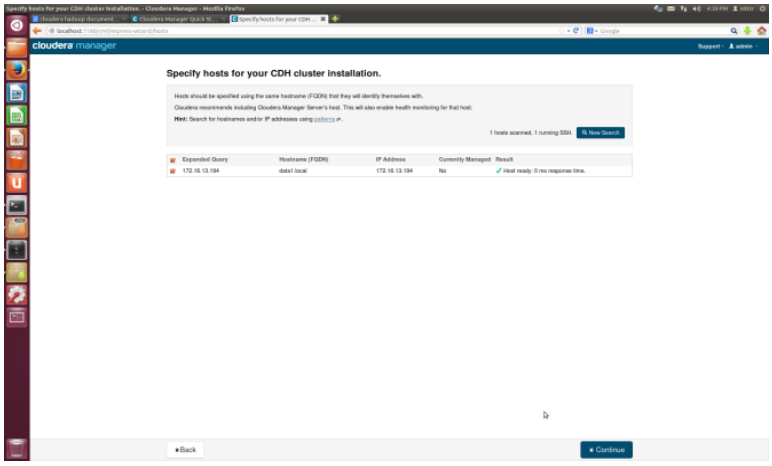

Figure 9Cloudera First GUI

The interface which is available after the installation of the Cloudera provides the information about the number of nodes, the task etc. All the information can be accessible through the interface. The default user name and password for the setup is admin. The login screen is shown below.

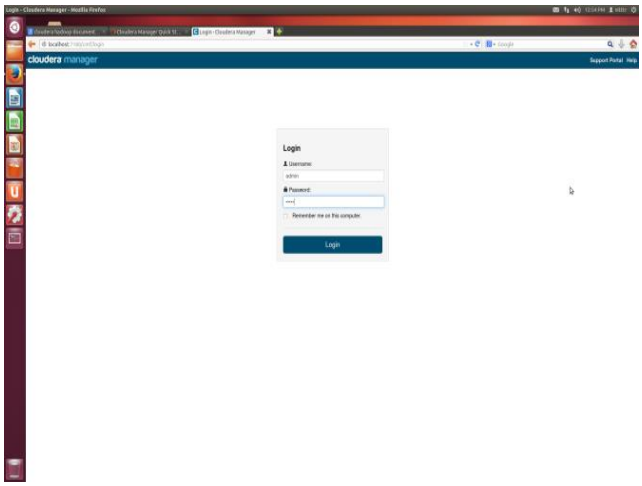

Figure 10 Login Screen

By doing these step the installation of the Cloudera setup can be done for big data.

IV. CONCLUSIONS

We have done setup of big data processing technology. Hadoop provides all project source code. It is easy to setup the big data processing environment using the virtual images. For a full working cluster it is recommended to use the Apache Hadoop project source code. While working with the Hadoop there is need to edit number files. The Hadoop distributed file system and the Map reduce provides efficient big Data handling. In future will develop map reduce algorithms for processing the data.

REFERENCES

[1] Data http://en.wikipedia.org/wiki/Data.
[2] Big Data http://en.wikipedia.org/wiki/Big_data.
[3] http://hdfpga.blogspot.in/2012/06/microsoft-oracle-ibm-supply-big-data.html.
[4] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Google, Inc.
[5] Hortonworks http://hortonworks.com/hdp/docs/.

[6] H.Gilbert Miller, Peter MorkFrom Data to Decisions: A Value Chain for Big Data 1520-9202/13/ © 2013 IEEE.
[7] http://hortonworks.com/hdp/downloads /
[8] http://www.cloudera.com/content/support/en/downloads/download-components/download-products.html?productID=4ZFrtT9ZQN.
[9] http://www.cloudera.com/content/cloudera-content/cloudera-docs/CM5/latest/Cloudera-Manager-Quick-Start/Cloudera-Manager-Quick-Start-Guide.html.

