ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016

A BLENDED TECHNOLOGY INITIATIVE OF INCORPORATING BIG DATA AND CLOUD – HD INSIGHT CLUSTER CREATION AND SETUP USING MICROSOFT AZURE PLATFORM WITH BIG DATA HADOOP FRAMEWORK

Mr.D.R.Sudharsan– III CSE / Knowledge Institute of Technology, Salem
Mr.E.Arun – IIICSE / Knowledge Institute of Technology, Salem
Mr.R.J.Vigneswaran – IIICSE / Knowledge Institute of Technology, Salem
Mr.K.Sabari – IIICSE / Knowledge Institute of Technology, Salem
Mr.N.Vasanth- IIICSE / Knowledge Institute of Technology, Salem
Mr.S.S.Aravinth – ASP/ CSE / Knowledge Institute of Technology, Salem

ABSTRACT:

Big data is one of the emerging technology in this current scenario. In order to handle the data in big data a well-defined lab environment is mandatory. That lab environment should consists of a huge amount of commodity hard wares to form Name node, Secondary name node and set of data nodes. Even after this the next tedious part of managing the big data is configuration of hadoop cluster and map reduce strategy. Whenever it's needed to monitor the hadoop cluster the user have to access that particular lab environment. These all issues can be overcome by the solution called HD INSIGHTS (Hadoop on azure) a blended technological platform provided by the Microsoft .By using this the whole lab environment and hadoop cluster can be setup in cloud without need of any configuration in linux and the user can access the particular hadoop cluster via RDP (Remote Desktop Protocal) with highly secured and authenticated fashion from word wide.

I) INTRODUCTION:

1. Cloud Computing

Cloud computing is the term which defines that any of the service that can be provided via internet, that service may be a storage or a platform or a

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016

infrastructure or anything else. Therefore these services are broadly classified into three types namely

Infrastructure as a service

Platform as a service

Software as a service

2. Big Data

The term big data is used to represent the collection of complex and large sets of structured , unstructured and semi structured data Such that it's very difficult to capture , process , store , search and analyze these kinds of data by using the conventional data bases management tools and traditional database management systems.

3. Intention of Blended Technology

Blended technology is a influence of different domains in which consists of a combination of one or more technology to give the solution for a real world issue .lets think about IBM bluemix it is a blended technology with combination of web and cloud where it provides a sufficient platform to develop and deploy the user's web page. It is not needed to install any local web servers and text editors in a machine. The only need is an internet connectivity

4. Microsoft Azure

Azure is the cloud computing platform provided by the Microsoft Especially azure is focused on platform as a service. Because of that it provides sufficient platform for emerging technologies like big data, web of things, machine learning etc..,In domain of analytics it offers lot of services like

HD insight, Data factory etc.

5. Hadoop

Hadoop is the open source java framework managed by apache software foundation has proven to be very helpful in storing and managing vast amount of complex data like structured, unstructured and semi structured data cheaply and efficiently. Which allows you to store large amount of data of a cluster in a low cost commodity hardware. It is very important to understand that hadoop is not a database it is a distributed file system (HDFS-hadoop distributed file system) which distributes the data over number of data nodes and also provides you capability to process that distributed data using a simple programming model.

II) PHYSICAL HADOOP CLUSTER CREATION AND SETUP:

In order to create a physical hadoop cluster hard ware is one of the main requirement and it's required to get a clear idea on linux machine working and there is lot of configuring procedures as shown below.

Step-1

Download the JDK setup file for linux

Make a directory in following location /usr/lib/jvm by using command

mkdir /usr/lib/jvm

step2

Move the downloaded JDK file into the JVM directory created in /usr/lib/jvm

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016

Step-3

Install java by using the following command

Update-alternatives /usr/bin/java java /usr/lib/jvm/java

Update-alternatives /usr/bin/javac java /usr/lib/jvm/javac

Step-4

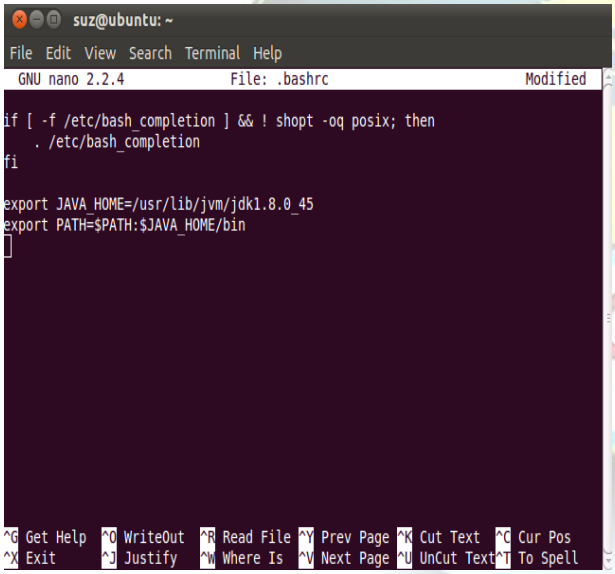
Open .bashrc file by using

nano .bashrc

Setup the java home path in .bashrc file

Export JAVA_HOME=/usr/lib/jvm/jdk-1.0.8

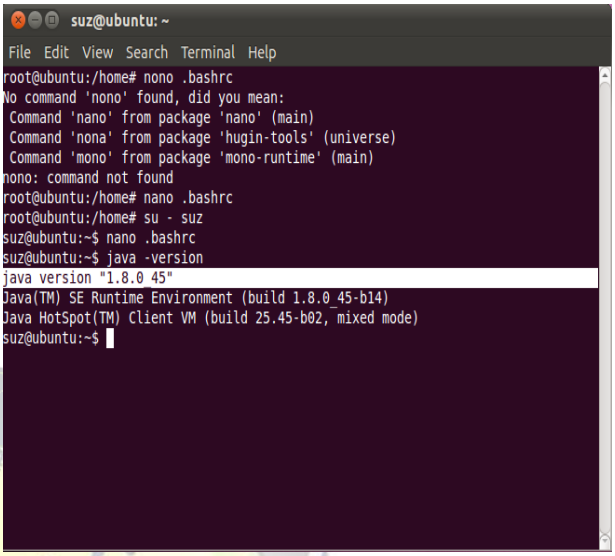
Export PATH=$PATH:$JAVA_HOME/bin



Step-5

Verify whether java had installed or not by the command

Java –version



Step-6

Install SSH client and server by using the command

sudo apt-get install openssh-client openssh-server

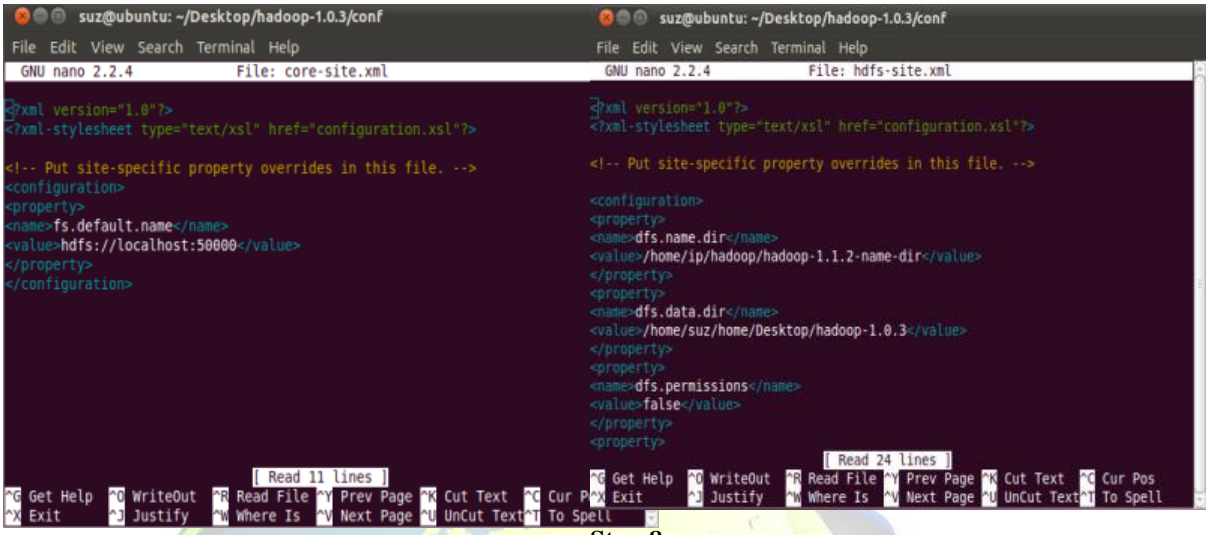
step-7

Download the hadoop setup and the place it to any of the directory you want

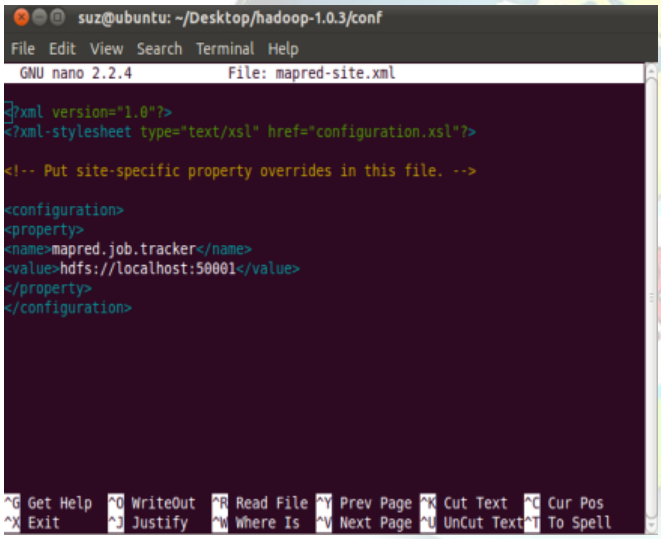
Make the following editing which shown in a picture given below in files

Conf/core-site.xml

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016



Conf/mapred-site.xml

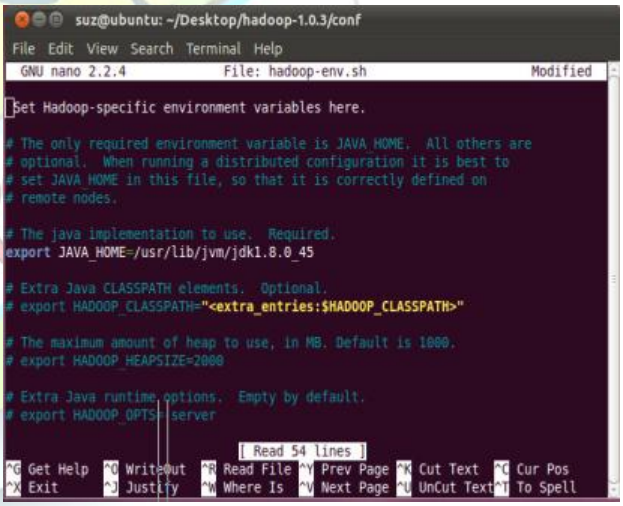


Conf.hdfs-site.xml

Step-8

Setup the hadoop environment path in a file

conf/hadoop-env.sh

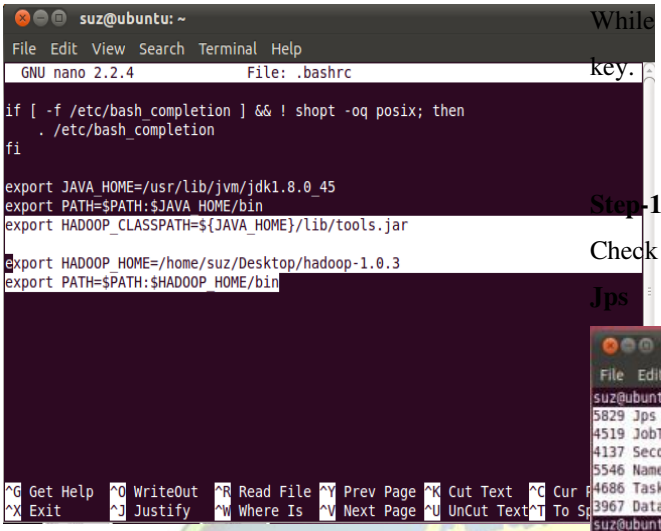


Step-9

Setup hadoop home directory path in .bashrc

As shown below

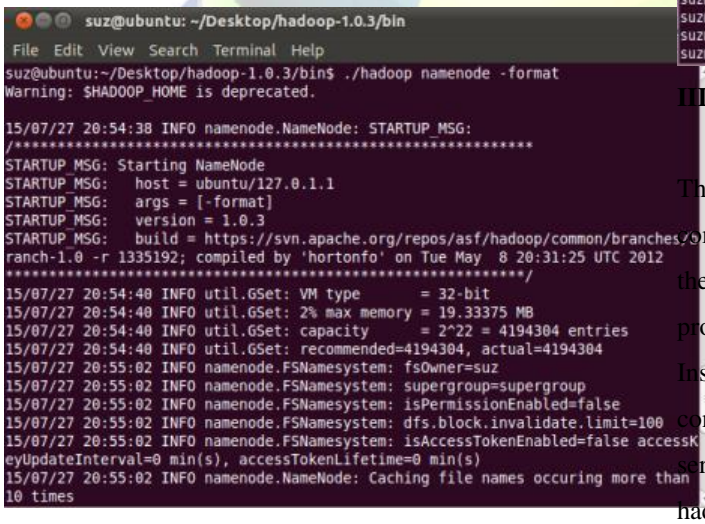
ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016



Step-10

Start formatting the namenode by the following command

Hadoop namenode –format (inside the bin directory)



Step-10

Start all the daemons by the command

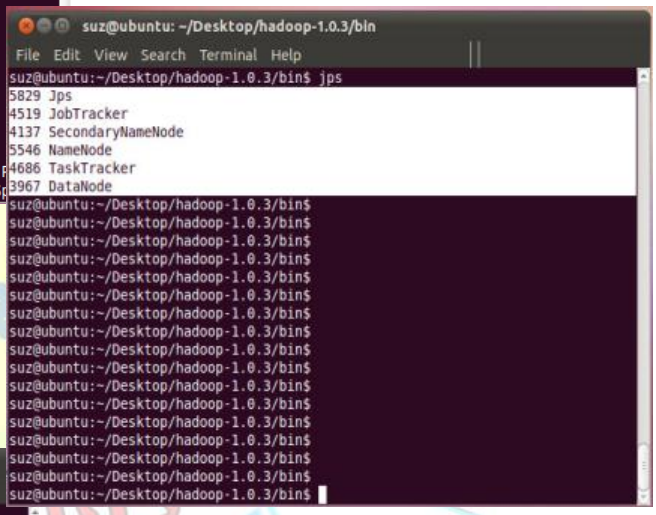
./start-all.sh (inside the bin directory)

While this process it asks you for the ssh password key.

Step-11

Check the java process status by the command

Jps



III) AZURE HD INSIGHT CLUSTER CREATION:

These kinds of hadoop installation by using linux commands will take much time to configure and also the process of importing and exporting data is also a problematic one

Instead of working with these complicated configuration for hadoop cluster, the HD INSIGHT a service provided Microsoft azure is used to create our hadoop cluster with in a few minutes.

Step-1

Login into the www.manage.azure.com web portal by using the Microsoft account

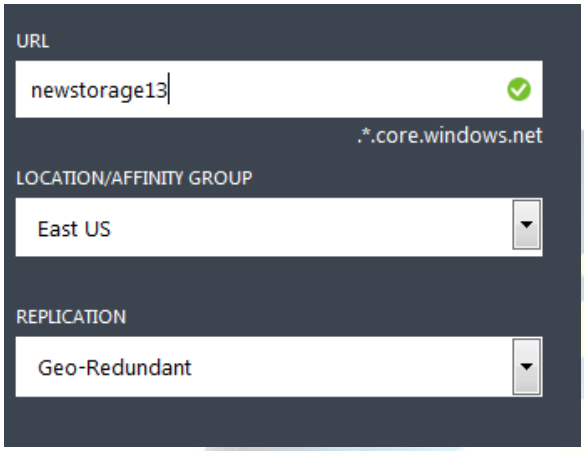
Step-2

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016

Then in the next page all the service provided by the Microsoft will be displayed like cloud, web apps, virtual machines etc..,

Among that select the STORAGE

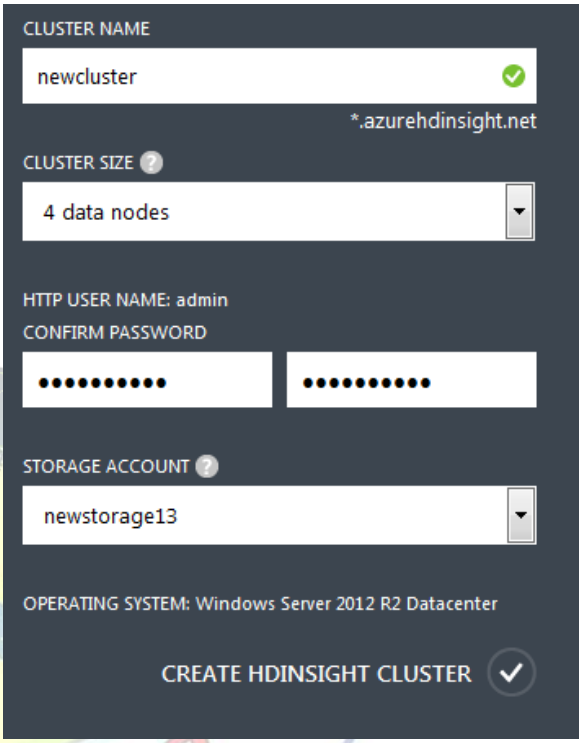
After that create your storage account.



Step-3

Select option hd insight and setup the hadoop cluster in the created storage account with number of data nodes and name node.

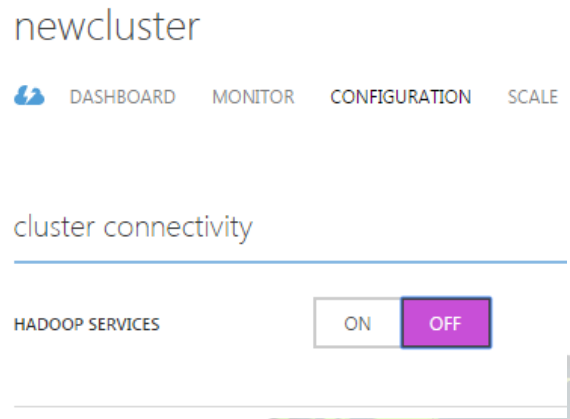
it is shows in the next page.



Step-4

In next page select configuration option at top of the page and switch on the hadoop services

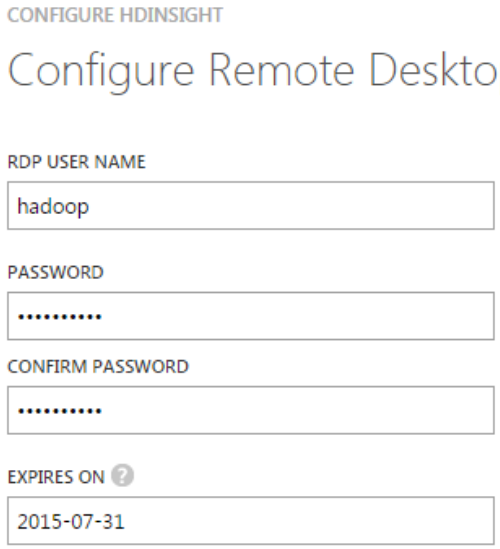
ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016



then in bottom select enable remote option the Remote Desktop Protocal will be downloaded.

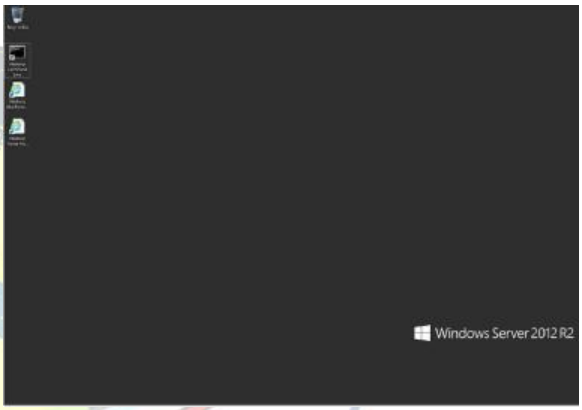
Step-5

The it will ask you to setup then user name and password for RDP.



Step-6

When the user opens the downloaded RDP it will ask you for the password created by you. After the verification of password then it opens up the virtual machine with the operating system of windows 2012. Here you are allotted with the memory space of 1000 GB only for your hadoop cluster in a cloud.



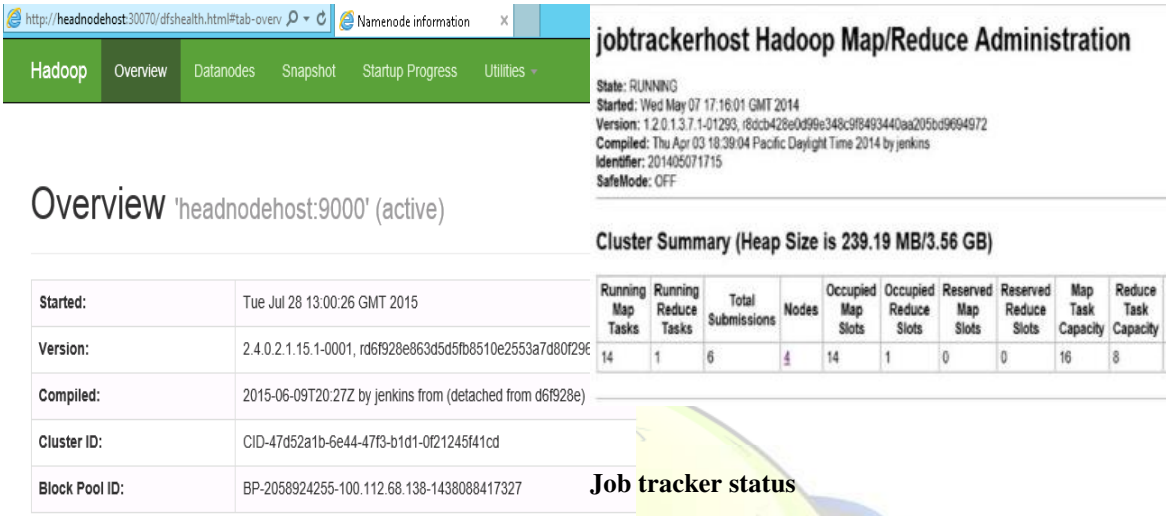
NAME NODE, MAP REDUCE, JOB TRACKER STATUS

Name node status.

The Hadoop data storage layer is the Hadoop Distributed File System (HDFS), and HDFS can be accessed from the management portal, the Hadoop command line, or from the Hadoop web service at port 30070 by the URL

http//headnodehost:30070/dfshealth.jsp

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016

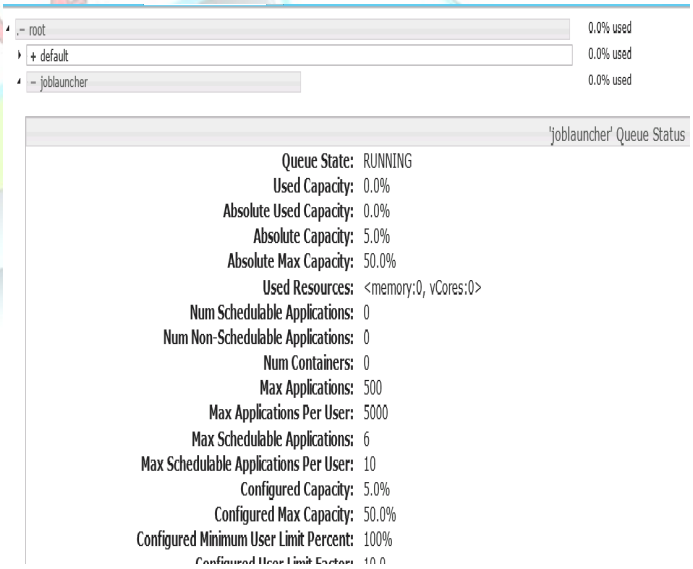


Job tracker status

it is a daemon which distributes the data towards the data node and also it stores the path that where a particular data is stored in data node.

Task Tracker

Task tracker is the daemon which perform the process in a part of data node and return the result to the job tracker. In this the scheduler information (queue status) can be known by the following link



Map reduce status

The Hadoop data processing layer has two components, JobTracker and TaskTracker. JobTracker usually runs on the same server as the name node. However, if needed, it can also run on a different machine. The role of JobTracker is to accept the jobs submitted to an HDInsight cluster from any source, break the job into tasks, and assign those tasks to the data nodes in the cluster. TaskTracker, which runs on each data node, sends status messages to JobTracker as it works on each job. To see the status of any job running on an HDInsight cluster, use the Hadoop web server on port 50030, following the URL template ***http://your_hadoop_server:50030/***, or the Hadoop MapReduce Status shortcut.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016

Hadoop command line & windows power shell

When the user opens the command prompt in windows 2012 server a virtual machine which is opened via RDP, it will display the prompt called hadoop command line. Windows power shell is the shell which support all hadoop operations by this HD insight cluster all the hadoop operations can be possible in a very easy manner

IV) BENEFITS OF PHYSICAL TO CLOUD

When compared to physical machine the HD insight cluster is cost wise efficient because it doesn't need commodity hardware that required memory is allocated from cloud in HD insight hadoop cluster and one of the main advantage is that it avoids the usage of hardware and tedious linux commands for hadoop configuration. In a physical hadoop cluster whenever the user needs to monitor the data then each and every time he need to access the same physical machine but user can access HD insight hadoop cluster from anywhere all over the world by using RDP the only need is internet access.

V) CONCLUSION:

This Paper gives a research on blended technology in a combination of big data and cloud called HD INSIGHTS (HADOOP OF AZURE) the service provided by Microsoft and gives the brief and elaborative knowledge on it, and it is known fact that almost all of the hadoop operations can be possible in HD INSIGHTS

References:

- https://azure.microsoft.com/en-in/documentation/articles/hdinsight-hadoop-tutorial-get-started-windows/
- https://azure.microsoft.com/en-in/services/hdinsight/
- https://channel9.msdn.com/Series/Getting-started-with-Windows-Azure-HDInsight-Service