Volume 4, Issue 11, November 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# Big Data and Hadoop Technology Solutions with Cloudera Manager

**Dr. Urmila R. Pol**

Department of Computer Science, Shivaji University,
Kolhapur, India

***Abstract:*** *Big data represents a new era in data assessment and exploitation. Big data analysis is the strategy of examining large data sets containing a variety of data types to uncover hidden patterns and other useful business information. The primary goal of big data analysis is to help companies make more informed business decisions. many organizations looking to collect, process and analyze big data have turned to a newer class of technologies that includes Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases. Those technologies form the core of an open source software framework that supports the processing of big and different data sets across clustered systems.*

*In this paper I represent the process of formation of hadoop cluster of four nodes with VirtualBox. I wanted to get everyone familiar with the big data world with hadoop .*

***Keyword:*** *YARN, HDFS, RAM*

## I. INTRODUCTION

Every day, we generate 3.5 quintillion bytes of data — so much that 92% of the data in the world today has been created in the last three years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone signals to name a few. This data is **big data.** Big data spans three dimensions: **Volume, Velocity and Variety**. Enterprises are lacking with ever-growing data of all types, easily amassing terabytes—even petabytes—of information. Analyse 5 million trade events created each day to identify potential fraud. Analyse 500 million daily call detail records in real-time to predict customer churn faster. Big data is any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analysing these data types together. Monitor 100's of live video feeds from surveillance cameras to target points of interest.

Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Mainstream BI software and data visualization tools can also play a role in the analysis process. But the semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands

posed by sets of big data that need to be updated frequently or even continually , for example, real-time data for mobile applications.

## II. WHAT IS HADOOP

Apache™ **Hadoop®** is an open source software project that enables the distributed processing of large data sets across clusters of servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer.

Apache Hadoop has two pillars:

**YARN** - Yet Another Resource Negotiator (YARN) assigns CPU, memory, and storage to applications running on a Hadoop cluster. The first generation of Hadoop could only run MapReduce applications. YARN enables other application frameworks (like Spark) to run on Hadoop as well, which opens up a wealth of possibilities.

- **HDFS** - Hadoop Distributed File System (HDFS) is a 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.

  Hadoop is supplemented by an system of Apache projects, such as Pig, Hive and Zookeeper, that extend the value of Hadoop and improves its usability. Hadoop changes the economics and the dynamics of large scale computing. Its impact can be bubbled down to four salient characteristics.

- **Scalable**– New nodes can be added as needed, and added without needing to change data formats, how data is loaded, how jobs are written, or the applications on top.

- **Cost effective**– Hadoop brings massively parallel computing to commodity servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it affordable to model all your data.

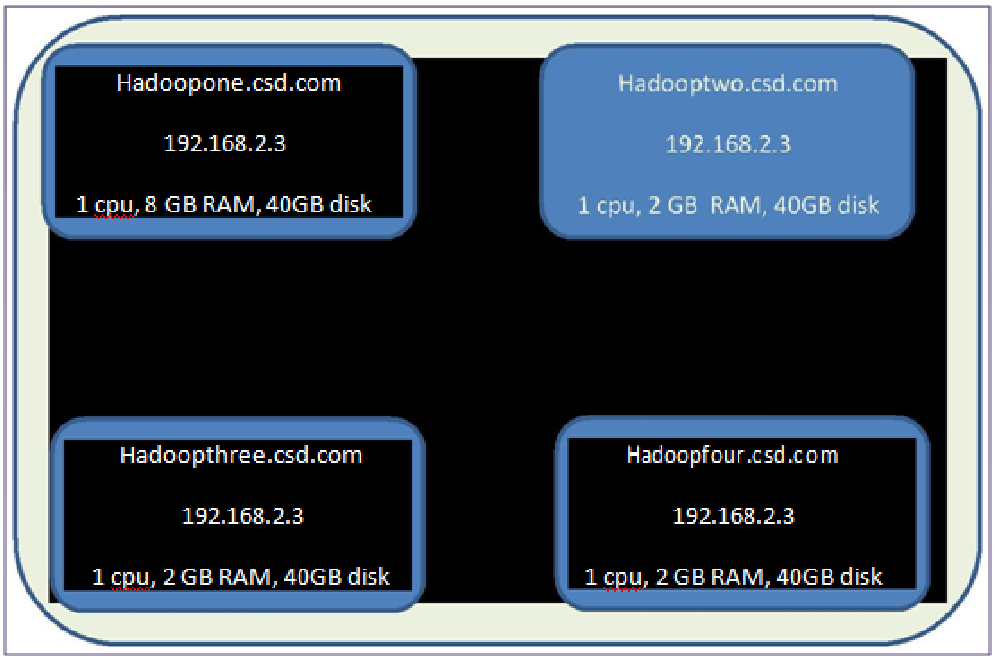
- **Flexible**– Hadoop is schema-less, and can absorb any type of data, structured or not, from any number of sources. Data from multiple sources can be joined and aggregated in arbitrary ways enabling deeper analyses than any one system can provide.

- **Fault tolerant**– When you lose a node, the system redirects work to another location of the data and continues processing without missing a fright beat.

### Simple Hadoop Cluster with VirtualBox

**Set up a CDH-based Hadoop cluster using VirtualBox and Cloudera Manager**

### Overview



**VirtualBox VM cluster running Hadoop nodes**

First create a virtual machine and configure it with the required parameters and settings to act as a cluster node (specially the network settings). This referenced virtual machine is then cloned as many times as there will be nodes in the Hadoop cluster. Only a limited set of changes are then needed to finalize the node to be operational (only the hostname and IP address need to be defined).

I created a 4 nodes cluster. The **first node**, which will run most of the cluster services, requires more memory (8GB) than the other 3 nodes (2GB). Overall we will allocate 14GB of memory, so ensure that the host machine has sufficient memory.

The prerequisites is that you should have the latest VirtualBox installed .You can download it from www.virtualbox.org . We will be using the CentOS 6.5 Linux distribution.You can download the

**CentOS x86_64bit DVD iso image** from www.centos.org/download.

### Base VM Image creation

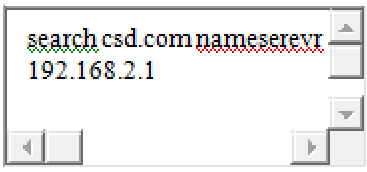
Create the reference virtual machine, with the following parameters:

Bridge network ,Enough disk space (more than 40GB) ,2 GB of RAM ,Setup the DVD to point to the CentOS iso image.when you install CentOS, you can specify the option _expert text‘, for a faster OS installation with minimum set of packages.

### Network Configuration

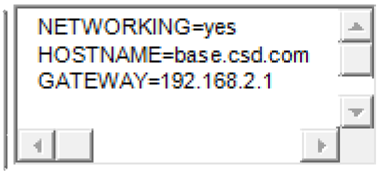
Perform changes in the following files to setup the network configuration that will allow all cluster nodes to interact.

*/etc/resolv.conf*



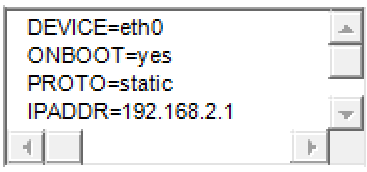
```
search csd.com nameserevr
192.168.2.1
```

*/etc/sysconfig/network*



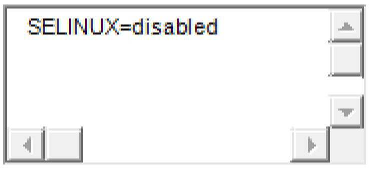
```
NETWORKING=yes
HOSTNAME=base.csd.com
GATEWAY=192.168.2.1
```

*/etc/sysconfig/network-scripts/ifcfg-eth0*



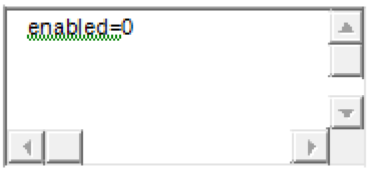
```
DEVICE=eth0
ONBOOT=yes
PROTO=static
IPADDR=192.168.2.1
```

*/etc/selinux/config*



```
SELINUX=disabled
```

*/etc/yum/pluginconf.d/fastestmirror.conf*



```
enabled=0
```

Initialize the network by restarting the network services:
$> chkconfig iptables off
$> /etc/init.d/network restart

**Installation of VM Additions**

You should now update all the packages and reboot the virtual machine:
$> yum update
$> reboot
In the VirtualBox menu, select *Devices*, and then *Insert Guest…*. This insert a DVD with the iso image of the guest additions in the DVD Player of the VM, mount the DVD with the following commands to access this DVD:
$> mkdir /media/VBGuest
$> mount -r /dev/cdrom /media/VBGuest

**Setup Cluster Hosts**

Define all the hosts in the /etc/hosts file in order to simplify the access, in case you do not have a DNS setup where this can be defined. Obviously add more hosts if you want to have more nodes in your cluster.
*/etc/hosts*
192.168.2.3 hadoopone.csd.com hadoopone
192.168.2.4 hadooptwo.csd.com hadooptwo
192.168.2.5 hadoopthree.csd.com hadoopthree
192.168.2.7 hadoopfour.csd.com hadoopfour

**Setup SSH**

To also simplify the access between hosts, install and setup SSH keys and defined them as already authorized
$> yum -y install perl openssh-clients $> ssh-keygen (type enter, enter, enter)
$> cd ~/.ssh

$> cp id_rsa.pub authorized_keys

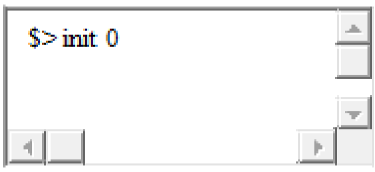
Modify the ssh configuration file. Uncomment the following line and change the value to no; this will prevent the question when connecting with SSH to the host.

*/etc/ssh/ssh_config*

StrictHostKeyChecking no

**Shutdown and Clone**

At this stage, shutdown the system with the following command:



```
$> init 0
```

We will now create the server nodes that will be members of the cluster in VirtualBox, clone the base server, using the _Linked Clone‘ option and name the nodes hadoopone, hadooptwo, hadoopthree and hadoopfour.

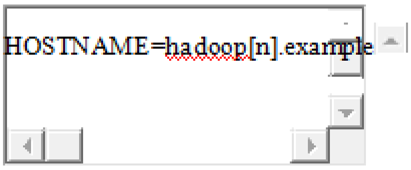
For the first node (hadoopone), change the memory settings to 8GB of memory. Most of the roles will be installed on this node, and therefore it is important that it have sufficient memory available.

**Clones Customization**

For every node, proceed with the following operations:

Modify the hostname of the server, change the following line in the file:

*/etc/sysconfig/network*



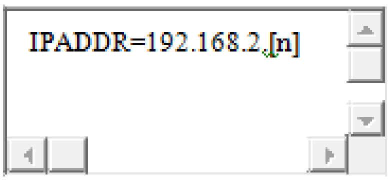
```
HOSTNAME=hadoop[n].example
```

1 HOSTNAME=hadoop[n].example.com

Where [n] = one..four (up to the number of nodes)

Modify the fixed IP address of the server, change the following line in the file:

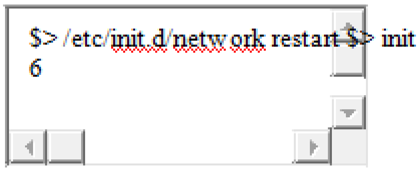
*/etc/sysconfig/network-scripts/ifcfg-eth0*



```
IPADDR=192.168.2.[n]
```

Where [n] = one..four (up to the number of nodes)

Let‘s restart the networking services and reboot the server, so that the above changes takes effect:

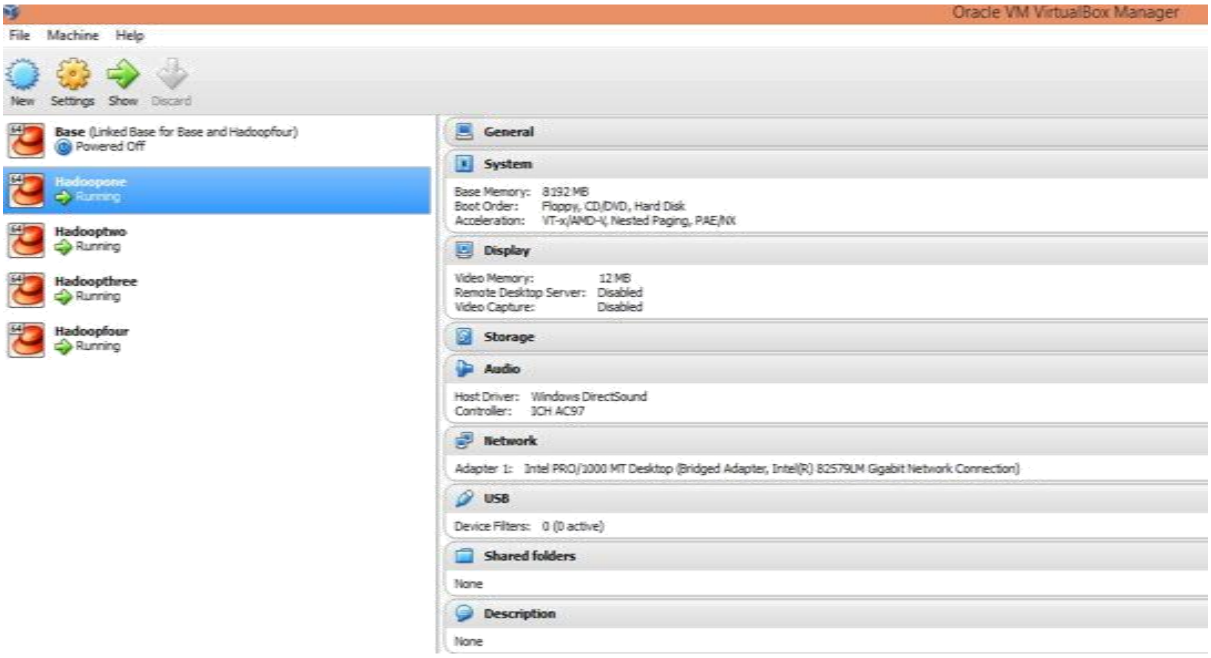


```
$> /etc/init.d/network restart $> init
6
```

1 $> /etc/init.d/network restart

2 $> init 6

at this stage we have four running virtual machines with CentOS correctly configured.



**Base** (Linked Base for Base and Hadoopfour)
Powered Off

**Hadoopone**
Running

**Hadooptwo**
Running

**Hadoopthree**
Running

**Hadoopfour**
Running

**General**

**System**

Base Memory: 8192 MB
Boot Order: Floppy, CD/DVD, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX

**Display**

Video Memory: 12 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

**Storage**

**Audio**

Host Driver: Windows DirectSound
Controller: ICH AC97

**Network**

Adapter 1: Intel PRO/1000 MT Desktop (Bridged Adapter, Intel(R) 82579LM Gigabit Network Connection)

**USB**

Device Filters: 0 (0 active)

**Shared folders**

None

**Description**

None

**Four Virtual Machines running on VirtualBox, ready to be setup in the Cloudera cluster.**

## Cloudera Manager

**Cloudera Manager** is the best way to install, configure, manage, and monitor your Apache Hadoop stack. Cloudera Manager runs a central server which hosts the UI Web Server and the application logic for managing CDH. Everything related to installing CDH, configuring services, and starting and stopping services is managed by the Cloudera Manager Server. The Cloudera Manager Agents are installed on every managed host. They are responsible for starting and stopping Linux processes, unpacking configurations, triggering various installation paths, and monitoring the host.

**Heartbeats** make up the primary communication channel in Cloudera Manager. The agents sends heartbeats (by default) every 15 seconds to the server to find out what the agent should be doing. The agent, when it‘s heartbeating, is saying: ―Here‘s what I‘m up to!‖ and the server, in its response, is saying, ―Here‘s what you should be doing.‖ Both the agent and the server end up doing some reconciliation: If a user has stopped a service via the UI, for example, the agent will stop the relevant processes; if a process failed to start, the server will mark the start command as having failed.

## Install Cloudera Manager on hadoopone

Download and run the Cloudera Manager Installer, which will simplify greatly the rest of the installation and setup process.

```
$> curl -O http://archive.cloudera.com/cm4/installer/latest/cloudera-manager-installer.bin

$> chmod +x cloudera-manager-installer.bin

$> ./cloudera-manager-installer.bin
```

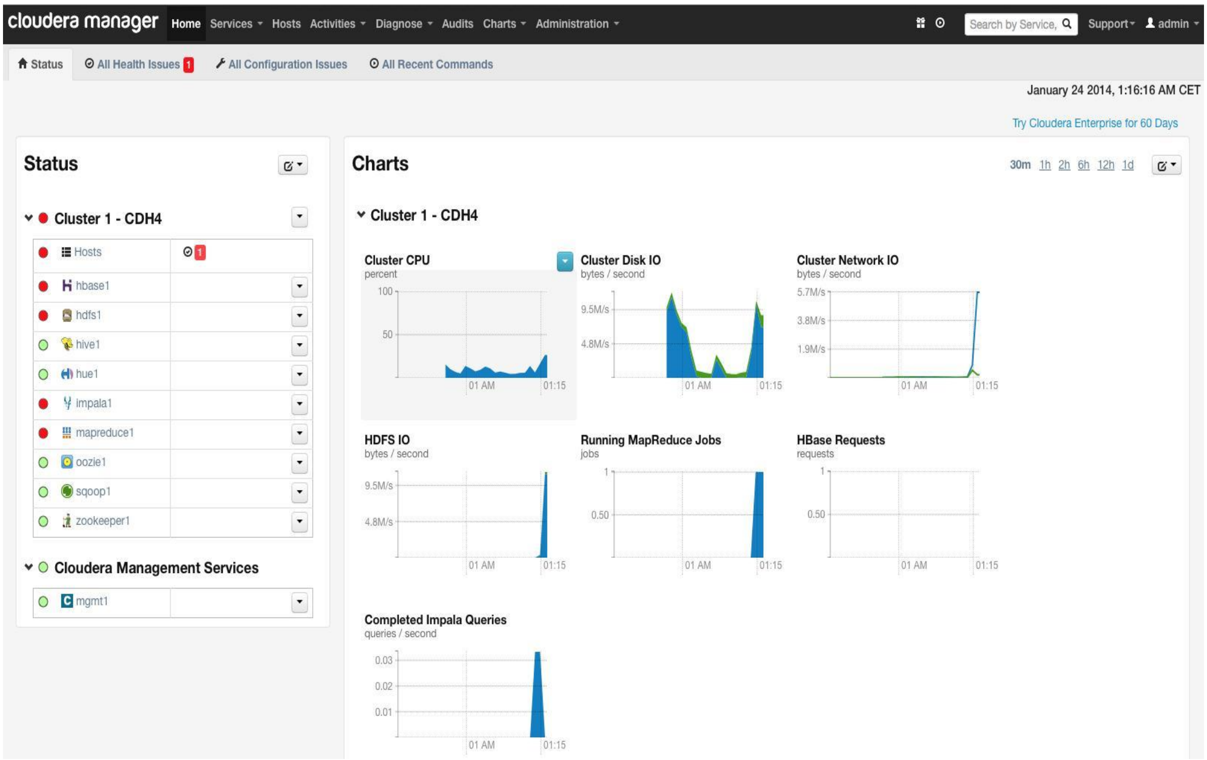
Use a web browser and connect to **http://hadoopone.csd.com:7180** (or **http://192.168.2.3:7180** if you have not added the hostnames into a DNS or hosts file).To continue the installation, you will have to select the Cloudera free license version. You will then have to define which nodes will be used in the cluster. Just enter all the nodes you have defined in the previous steps(e.g. *hadoopone.csd.com*) separated by a space. Click on the ―Search‖ button. You can then used the root password (or the SSH keys you have generated) to automate the connectivty to the different nodes. Install all packages and services onto the 1st node.Once this is done, you will select additional service components; just select everything by default. The installation will continue and will complete.

## Using the Hadoop Cluster

Now that we have an operational Hadoop cluster, there are two main interfaces that you will use to operate the cluster: Cloudera Manager and Hue.

## III. CLOUDERA MANAGER

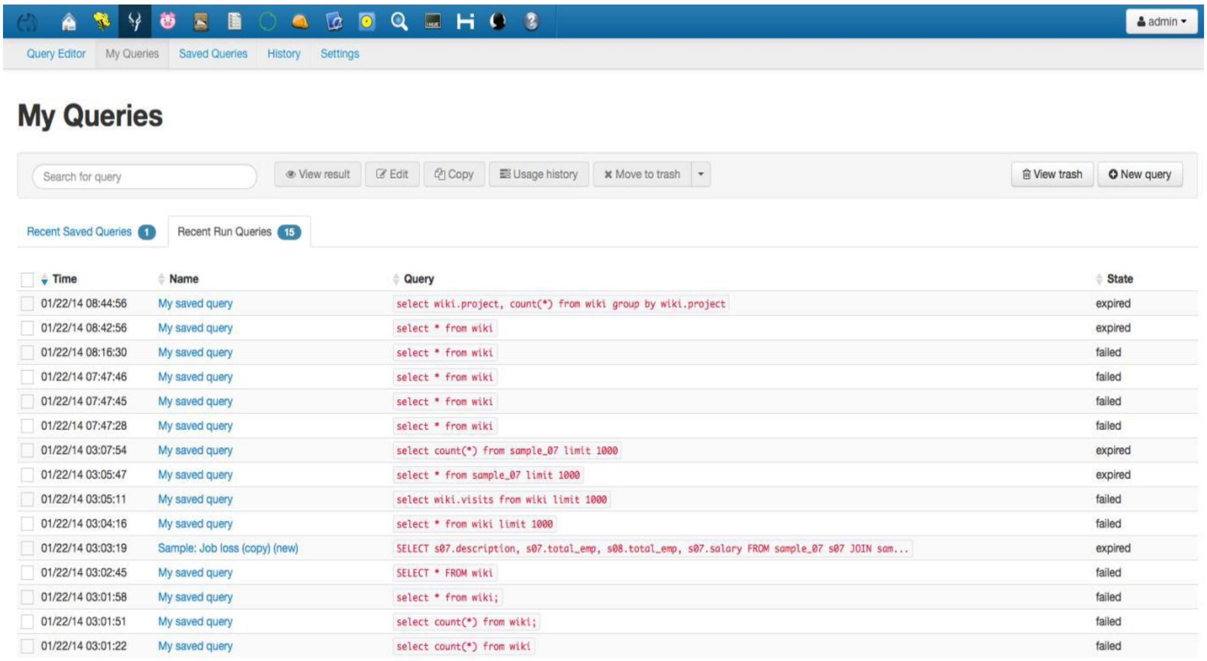
Use a web browser and connect to http://hadoopone.csd.com:7180 (or http://192.168.2.3:7180 if you have not added the hostnames into a DNS or hosts file).



**Cloudera Manager homepage, presenting cluster health dashboards**

**Hue**

Similarly to Cloudera Manager, you can access the Hue administration site by accessing: http://hadoopone.csd.com:8888, where you will be able to access the different services that you have installed on the cluster.



| Time | Name | Query | State |
|---|---|---|---|
| 01/22/14 08:44:56 | My saved query | select wiki.project, count(*) from wiki group by wiki.project | expired |
| 01/22/14 08:42:56 | My saved query | select * from wiki | expired |
| 01/22/14 08:16:30 | My saved query | select * from wiki | failed |
| 01/22/14 07:47:46 | My saved query | select * from wiki | failed |
| 01/22/14 07:47:45 | My saved query | select * from wiki | failed |
| 01/22/14 07:47:28 | My saved query | select * from wiki | failed |
| 01/22/14 03:07:54 | My saved query | select count(*) from sample_07 limit 1000 | expired |
| 01/22/14 03:05:47 | My saved query | select * from sample_07 limit 1000 | expired |
| 01/22/14 03:05:11 | My saved query | select wiki.visits from wiki limit 1000 | failed |
| 01/22/14 03:04:16 | My saved query | select * from wiki limit 1000 | failed |
| 01/22/14 03:03:19 | Sample: Job loss (copy) (new) | SELECT s07.description, s07.total_emp, s08.total_emp, s07.salary FROM sample_07 s07 JOIN sam... | expired |
| 01/22/14 03:02:45 | My saved query | SELECT * FROM wiki | failed |
| 01/22/14 03:01:58 | My saved query | select * from wiki; | failed |
| 01/22/14 03:01:51 | My saved query | select count(*) from wiki; | failed |
| 01/22/14 03:01:22 | My saved query | select count(*) from wiki | failed |

**Hue interface, and here more specifically, an Impala saved queries window.**

## IV. CONCLUSIONS

We have been create a small Hadoop cluster using Cloudera Manager Installer, which simplifies the installation to the simplest of operation. It is now possible to execute and use the various examples installed on the cluster, as well as understand the interactions between the nodes.

## REFERENCES

[1] Paramjot Singh, Vishal Pratap Singh, Gaurav Pachauri, "Critical Analysis of Cloud Computing Using OpenStack" IJCSMC, Vol. 3, Issue. 3, March 2014, pg.121 – 127

[2] OpenFlow as a Service; Fred Hsu, M. Salman Malik, Soudeh Ghorbani {fredhsu2,mmalik10,ghorban2}@illinois.edu

[3] Omar SEFRAOUI, Mohammed AISSAOUI, Mohsine ELEULDJ; ***"OpenStack: Toward an Open**-Source **Solution for Cloud Computing",*** International Journal of Computer Applications (0975 - 8887) Volume 55 - No. 03, October 2012

[4] OpenStack URL: http://www.openstack.org/

[5] CloudStack URL:http://www.cloudstack.apache.org/

[6] Open Nebula URL: http://opennebula.org/