ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 39-45 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp20.pdf

Running Hadoop 2.7.3 single node cluster on Ubuntu 16.04 LST

Ziad Munther Al-Muhaisen
Amman Arab University
Faculty of Computer Science and Informatics
Amman,Jordan
zmuhaisen@hotmail.com

Dr. Mohammad Othman Nassar
Amman Arab University
Faculty of Computer Science and Informatics
Amman,Jordan
moanassar@aau.edu.jo

Abstract— ***Apache Hadoop has the potential to revolutionize the processing of very large data sets. Due to its importance, Hadoop is receiving an increasing attention among research literature in recent years, where many implementations and algorithms have been proposed to overcome constrains and challenges in Hadoop framework in order to improve its data availability and overcome heavy load failures. Hadoop consists of multiple nodes which offer transparency and redundancy by allowing the distribution of storage and processing over multiple different locations.***

This paper gives a brief outline to the storage part of apache Hadoop, known as "Hadoop Distributed File System (HDFS)", and describes installation procedures, the environment requirements needed to run the Hadoop approach on a single node cluster, and how to overcome the problems during the installation.

Keywords- Hadoop, HDFS, Ubuntu 16.04 LST

I. INTRODUCTION

Apache Hadoop provides a distributed file system for processing very large datasets using the distribution of data and parallel computation across thousands of hosts, servers store and analyze large sets of by distributing storage and computation across servers, Hadoop clusters at Yahoo! span 25 000 servers, and store 25 petabytes of application data, with the largest cluster being 3500 servers [1].

Hadoop ecosystem includes other projects for a particular needs, for example Hive gives an SQL-like interface to query data stored in a Hadoop cluster, Pig is for analyzing large data sets, Zookeeper is used for federating services and Oozie is a scheduling system[2].

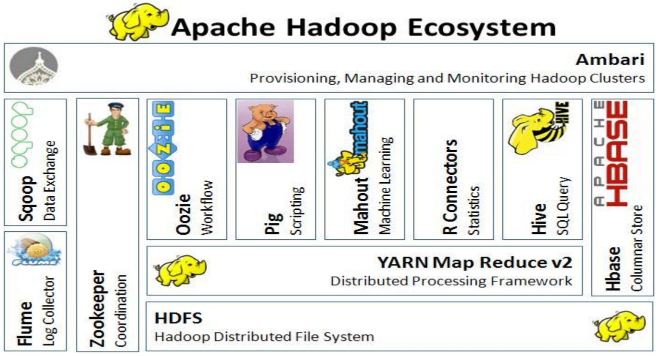


Figure 1[3]

APACHE HADOOP 2

Is the second iteration of the Hadoop framework. Hadoop project framework includes these modules, Hadoop Common utilities, Hadoop Distributed File System, Hadoop YARN, and Hadoop MapReduce [4].

The major types of machines in a Hadoop deployment are Client machines, Masters and Slave nodes [5] [10].

The Master nodes which is responsible for storing data (HDFS), and running parallel processing (Map Reduce). The Name Node coordinates the data storage, while the Job Tracker manages and coordinates the parallel processing using Map Reduce. Slave Nodes store the data and run the computations [5] [6]. In a Hadoop single node cluster all of the nodes will be running on the same machine [6].

II. EQUIPMENTS AND METHODS

The PC: Intel Core I5 machine , 4 GB RAM , 250 GB hard disk.

The OS: Hadoop is supported by GNU/Linux as a development and production platform, Windows is also a supported platform, Hadoop version 2.x includes native support for Windows [7].

As mentioned Hadoop development is supported by GNU/Linux operating system therefore an open source

Ubuntu 16.04 LTS will be used to run the Hadoop single node environment.

STEP 1:

Install Ubuntu 16.04 LTS from https://www.ubuntu.com/download/server

Advice: download the ISO image from the site and burn the iso image on a cd and use the CD to install Ubuntu 16.04 OS on the PC (fresh install).

The installation of Ubuntu is a straightforward operation but for people working with Windows environment it's somehow different.

Connect the pc to the internet and configure the internet connection for Ubuntu server, the easy way is to plug the pc to a router with DHCP enabled.

For more comfort using Ubuntu install the Ubuntu 16.04 GUI (Graphic User interface).

To install Ubuntu GUI

Type

sudo apt-get update
sudo apt-get install ubuntu-desktop

After restarting Ubuntu will start in the GUI (Figure 2)

Figure 2

Step 2:

Commands used in this installation are obtained from" https://www.digitalocean.com/community" [8]

Install java

1- Click on the terminal windows as shown in (Figure 3)

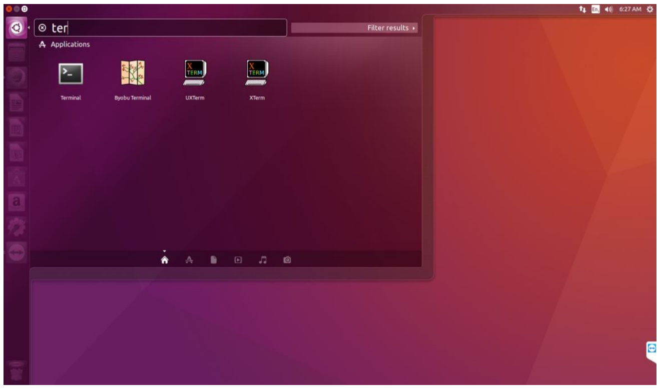


Figure 3

2- type each command line at a time

sudo apt-add-repository ppa:webupd8team/java
sudo apt-get update
sudo apt-get install oracle-java8-installer

3- check if java is correctly install

Type

java -version

If java is correctly installed the system will show

openjdk version "1.8.0_111"
OpenJDK Runtime Environment (build 1.8.0_111-8u111-b14-2ubuntu0.16.04.2-b14)
OpenJDK 64-Bit Server VM (build 25.111-b14, mixed mode)

Type

Javac –version

If javac is correctly installed the system will show

javac 1.8.0_111

Step 3:

Setting the JAVA_HOME Variable

Type

update-alternatives --config java

Copy the location of java by selecting from the line beginning with “/user/lib/…./” don’t copy the jre section

Type

gedit /etc/environment

An editor will open with a file containing a path section

add the following at the end

JAVA_HOME="/usr/lib/jvm/……"

Where “/usr/lib/jvm/……" is the line you copied form the command “update-alternatives --config java”

Save the file from the save bottom on the left hand side and exit, you will get back to the terminal window

Type

echo $JAVA_HOME

You will see the line added above

“/usr/lib/jvm/……”

Step 4:

Installing SSH

Type

apt-get install ssh
Press Y when it asks to and wait until ssh installation finishes

Type

ssh-keygen -t rsa -p""
Press enter

Type

cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys

Step 5:

Commands used in this installation are obtained from” https://www.ibm.com/developerworks/community/blogs” [9]

Installing Hadoop 2.7.3

Go to http://hadoop.apache.org/releases.html click 2.7.3 binary, download the gz file from the suggested mirror site at the beginning of the page.

Go to the download folder from files icon on desktop and go to the download folder ,Right click on the Hadoop .gz file and click extract her ,a folder will show with the name hadoop-2.7.3 like shown on (figure 4).

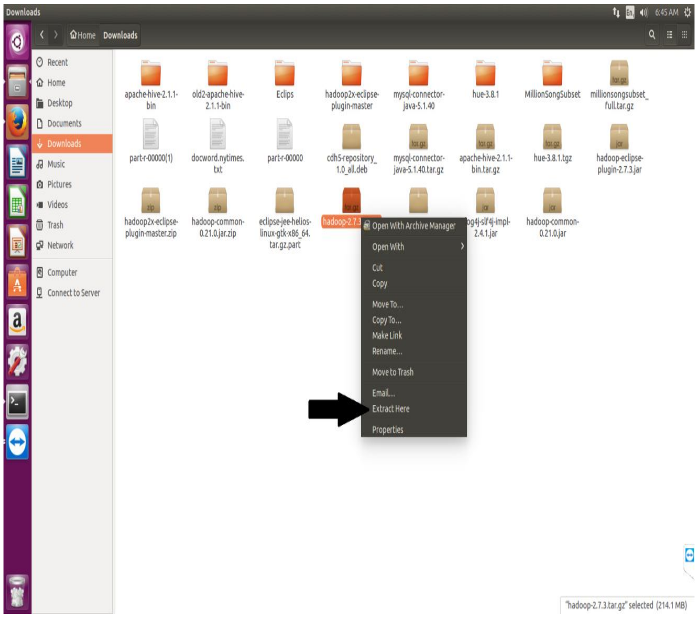

Downloads
6:45 AM
Home
Downloads
Recent
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
Trash
Network
Computer
Connect to Server
apache-hive-2.1.1-bin
old2-apache-hive-2.1.1-bin
Eclips
hadoop2x-eclipse-plugin-master
mysql-connector-java-5.1.40
hue-3.8.1
MillionSongSubset
millionsongsubset_full.tar.gz
part-r-00000(1)
docword.nytimes.txt
part-r-00000
cdh5-repository_1.0_all.deb
mysql-connector-java-5.1.40.tar.gz
apache-hive-2.1.1-bin.tar.gz
hue-3.8.1.tgz
hadoop-eclipse-plugin-2.7.3.jar
hadoop2x-eclipse-plugin-master.zip
hadoop-common-0.21.0.jar.zip
eclipse-jee-helios-linux-gtk-x86_64.tar.gz.part
hadoop-2.7.3
og4j-slf4j-impl-2.4.1.jar
hadoop-common-0.21.0.jar
Open With Archive Manager
Open With
Cut
Copy
Move To...
Copy To...
Make Link
Rename...
Move to Trash
Email...
Extract Here
Properties
"hadoop-2.7.3.tar.gz" selected (214.1 MB)

Figure 4

Copy this folder to your home folder as shown in (Figure 5)

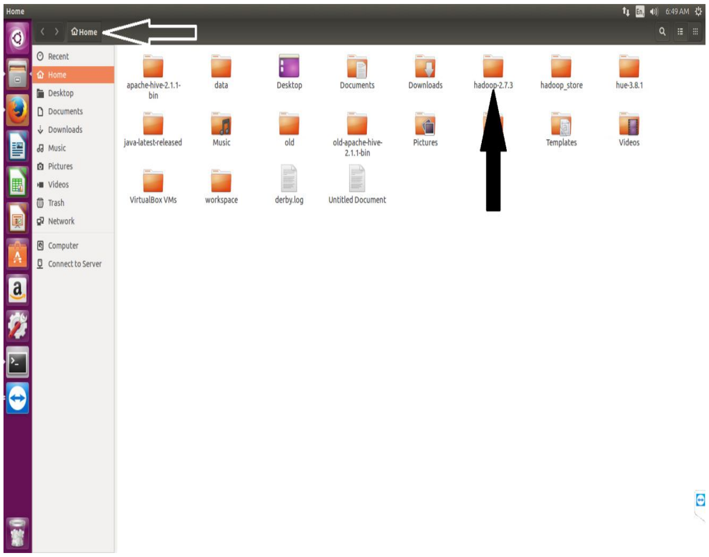


Figure 5

Configuration for the (.bashrc) file

In the terminal window

Type

```
gedit  ~/.bashrc
```

Add this to the end

```
#HADOOP VARIABLES START
export JAVA_HOME=/usr/lib/jvm/"java path when u installed
jave"
export PATH="$PATH:$JAVA_HOME/bin"
export HADOOP_INSTALL=/home/"YOUR USER
NAME"/hadoop-2.7.3
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export
HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INST
ALL/lib/native
export HADOOP_OPTS="-
Djava.library.path=$HADOOP_INSTALL/lib"
export HADOOP_HOME=/home/"YOUR USER
NAME"/hadoop-2.7.3
export PATH=$PATH:$HADOOP_HOME/bin
```

Type

```
source ~/.bashrc
```

To load the .bashrc changes

Go to Hadoop 2.7.3 directory

In the terminal window

Type

```
cd hadoop-2.7.3/etc/hadoop/
gedit hadoop-env.sh
```

At the end add

```
export JAVA_HOME=/usr/lib/jvm/"JAVA PATH WHEN YOU
INSTALLED JAVA"
```

Create new directory in your HOME FOLDER and name it hadoop_store

Make sure you're in your HOME FOLDER

Type

```
mkdir hadoop_store
```

To go to the directory

Type

```
cd hadoop_store
```

Create another directory inside Hadoop_store and name it hdfs

Type

```
mkdir hdfs
```

Go to hdfs directory

Type

```
cd hdfs
```

Create a directory inside hdfs and name it namenode

```
mkdir namenode
```

Don't change current directory to the namenode directory stay on hdfs folder

Create another directory name it datanode

```
mkdir datanode
```

Go to home directory, check the folders you created (Figure 6)

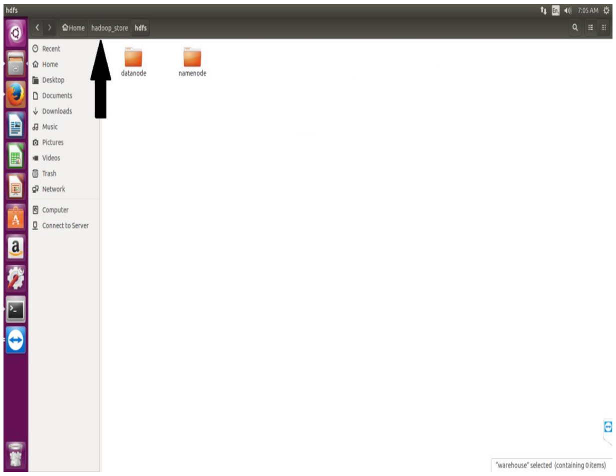


Figure 6

Type

```
cd hadoop-2.7.3/etc/hadoop/
edit hdfs-site.xml
```

Type

```
gedit hdfs-site.xml
```

Add the following between <configuration> add the lines here</configuration>

```
<property>
```

```
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/home/”YOUR USER NAME”/hadoop_store/hdfs/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/home/”YOUR USER NAME”/hadoop_store/hdfs/datanode</value>
</property>
<property>
<name>dfs.webhdfs.enabled</name>
<value>true</value>
</property>
<property>
<name>hadoop.proxyuser.hue.hosts</name>
<value>*</value>
</property>
<property>
<name>hadoop.proxyuser.hue.groups</name>
<value>*</value>
</property>
</configuration>
```

Save the file

create a tmp directory inside the Hadoop-2.7.3 folder

Go to terminal window

Type

```
mkdir Hadoop-2.7.3/tmp
```

Type

```
cd hadoop-2.7.3/etc/hadoop
```

configure a file called core-site.xml

Type

```
gedit core-site.xml
```

Add these lines to the file <configuration> </configuration>

```
<configuration>
<property>
<name>hadoop.tmp.dir</name>
<value>/home/”YOUR USER NAME”/hadoop-2.7.3/tmp</value>
<description>A base for other temporary directories.</description>
</property>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:54310</value>
</property>
<property>
<name>hadoop.proxyuser.hue.hosts</name>
<value>*</value>
</property>
<property>
<name>hadoop.proxyuser.hue.groups</name>
<value>*</value>
</property>
</configuration>
```

Save the file

copy a file called mapred-site.xml.template to mapred-site.xml inside the hadoop-2.7.3/etc/hadoop directory

Type

```
cp mapred-site.xml.template mapred-site.xml
```

Edit mapred-site.xml

Type

```
gedit mapred-site.xml
```

Add the following between <configuration> </configuration>

```
<configuration>
<property>
<name>mapred.job.tracker</name>
<value>localhost:54311</value>
<description>The host and port that the MapReduce job tracker runs
at.  If "local", then jobs are run in-process as a single map and reduce task.
</description>
</property>
</configuration>
```

format the Hadoop file system to make it usable for storing files,

Very Important Note: Hadoop file system (HDFS) is virtual it can be accessed with the Hadoop fs command.

Type

```
hadoop namenode –fromat
```

III. Results

After the installation of Hadoop 2.7.3 go to hadoop-2.7.3/sbin directory

Type

```
cd hadoop-2.7.2/sbin
```

start Hadoop

Type

start-all.sh

Enter your system password

Type

jps

if you see these lines on the screen then your configuration is correct

Jps

NameNode

SecondaryNameNode

DataNode

ResourceManager

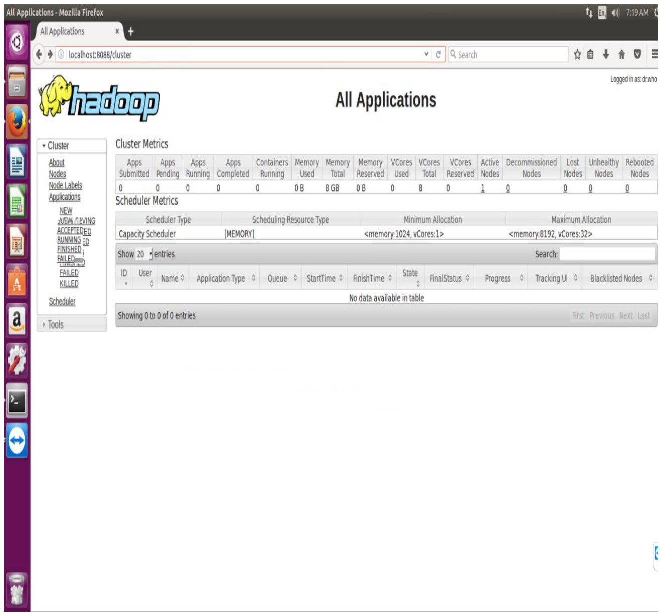
NodeManager

Go to your desktop

And start Firefox browser and navigate to

localhost:8088

You will see the following page figure (7).



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
0	0	0	0	0	0 B	8 GB	0 B	0	8	0	1	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:32>

Show 20 entries Search:

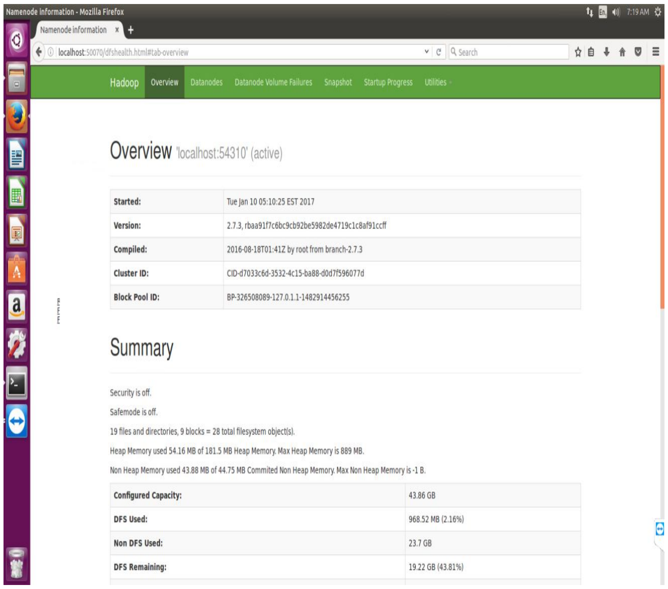
ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
No data available in table											

Showing 0 to 0 of 0 entries First Previous Next Last

Figure 7

To check HDFS navigate to *localhost:50070*

If **(Figure 8)** page is loaded then HDFS is working



Overview 'localhost:54310' (active)

Started:	Tue Jan 10 05:10:25 EST 2017
Version:	2.7.3, rbaa91f7c6bc9cb92be5982de4719c1c8af91ccff
Compiled:	2016-08-18T01:41Z by root from branch-2.7.3
Cluster ID:	CID-d7033c6d-3532-4c15-ba88-d0d7f596077d
Block Pool ID:	BP-326508089-127.0.1.1-1482914456255

Summary

Security is off.

Safemode is off.

19 files and directories, 9 blocks = 28 total filesystem object(s).

Heap Memory used 54.16 MB of 181.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 43.88 MB of 44.75 MB Commited Non Heap Memory. Max Non Heap Memory is -1 B.

Configured Capacity:	43.86 GB
DFS Used:	968.52 MB (2.16%)
Non DFS Used:	23.7 GB
DFS Remaining:	19.22 GB (43.81%)

Figure 8

IV. Recommendations

The absence of any required software or incorrect configuration of files will prevents Hadoop services from starting; error messages can give an indication about the missing software or the incorrect configuration.

For a best practice on running a Hadoop environment add more computers to the cluster (data nodes) this will give a clear understanding of the Hadoop multi node cluster and how each node performs its designated task, plus giving the ability to run larger testing datasets in the Hadoop environment[11].

A Hadoop ecosystem tools could also be implemented in the environment.

V. Conclusion

Hadoop has rapidly evolved during the last few years, major development projects improved the capabilities and the performance of the Hadoop framework, Hadoop can lead to new opportunities in processing of structured and unstructured massive data collected from difference resources [12].

References

1. Shvachko, Konstantin, et al. "The Hadoop distributed file system." Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. IEEE, 2010.

2. Monteith, J. Yates, John D. McGregor, and John E. Ingram. "Hadoop and its evolving ecosystem." 5th International Workshop on Software Ecosystems (IWSECO 2013). 2013.

3. http://thebigdatablog.weebly.com/blog/the-hadoop-ecosystem-overview

4. White, Tom. Hadoop: The definitive guide. " O'Reilly Media, Inc.", 2012.

5. Wang, Lizhe, et al. "G-Hadoop: MapReduce across distributed data centers for data-intensive computing." Future Generation Computer Systems 29.3 (2013): 739-750.

6. Vaidya, Madhavi. "Parallel Processing of cluster by Map Reduce." International journal of distributed and parallel systems 3.1 (2012): 167.

7. Apache™ Hadoop® 2.7.2 .https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html

8. https://www.digitalocean.com/community/tutorials/how-to-install-java-with-apt-get-on-ubuntu-16-04

9. https://www.ibm.com/developerworks/community/blogs/d9a07ec3-11e2-467d-b758-6861c4cb1d44/entry/How_to_install_Hadoop_2_7_0_in_ubuntu_16_04?lang=en

10. Hansen, Christer A. "Optimizing Hadoop for the cluster." Institue for Computer Science, University of Troms0, Norway, http://oss. csie. fju. edu. tw/~ tzu98/Optimizing Hadoop for the cluster. pdf, Retrieved online October (2012).

11. Bakshi, Kapil. "Considerations for big data: Architecture and approach." Aerospace Conference, 2012 IEEE. IEEE, 2012.

12. Agrawal, Divyakant, Sudipto Das, and Amr El Abbadi. "Big data and cloud computing: current state and future opportunities." Proceedings of the 14th International Conference on Extending Database Technology. ACM, 2011.