Running MapReduce Programs On Single Node Hadoop Cluster - Word Count/Word Frequency

Expt No: 3 March 06, 2019
Author: Subalakshmi Shanthosi S (186001008)

Aim

Implementation of MapReduce in Hadoop single node cluster.

Description

- Apache Hadoop
  - Large Scale,Open Source Software Framework.
  - Supports Three Projects:
    - Hadoop Common.
    - HDFS : Hadoop Distributed File System.
    - MapReduce.
- Hadoop MapReduce
  - Hadoop Programming Model and Software Framework.
  - Computational Processing:
    - Unstructured Data : File system
    - Structured Data : Database
  - MapReduce Layer has job and task tracker nodes.
  - Cluster nodes:
    - Single JobTracker per master.
    - Single TaskTracker per slave.
  - Fundamental Steps:
    - Map Step:
      - Master node slices problem input into several subproblems input.
      - Distributes data slices to worker nodes.
      - Worker nodes processes and hands over the control to master.
    - Reduce Step:
      - Master node takes the answers to the sub problems and combines them in a predefined way to get the output/answer to original problem.

Software's Used

- Ubuntu 16.04 LTS
- Hadoop 1.0.3

Hadoop MapReduce Architecture

- MapReduce Architecture:

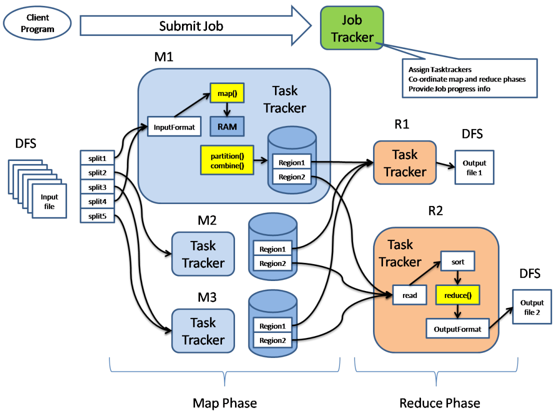
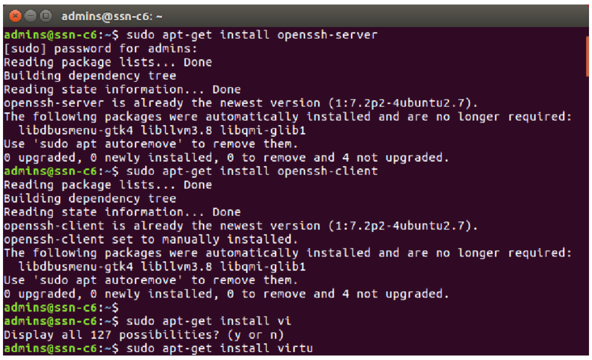


Figure 1: Hadoop MapReduce Architecture Diagram.

Procedure

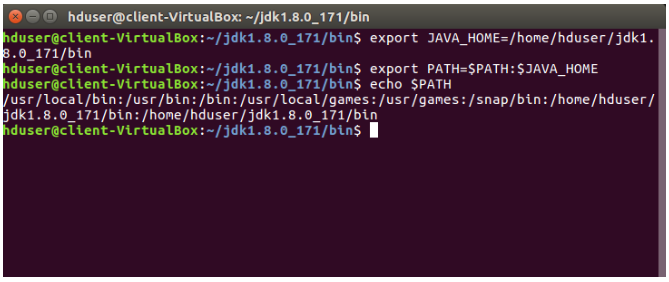
1. Launch Ubuntu 16.04 LTS.
2. Login to the OS with sudo permission and install the following packages using apt-get command
    - openssh-server
    - openssh-client
    - java jdk 8
    - javac compiler
    - hadoop 2.7.3

Output



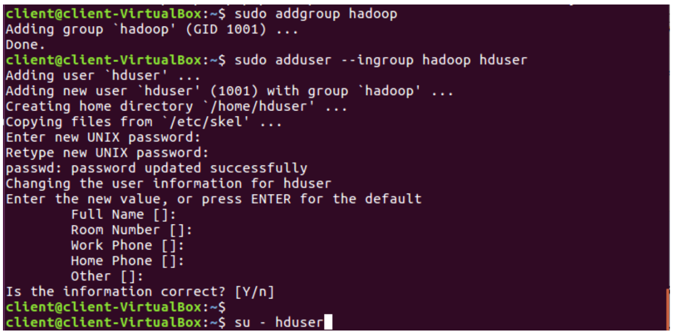
```
admins@ssn-c6:~$ sudo apt-get install openssh-server
[sudo] password for admins:
Reading package lists... Done
Building dependency tree
Reading state information... Done
openssh-server is already the newest version (1:7.2p2-4ubuntu2.7).
The following packages were automatically installed and are no longer required:
  libdbusmenu-gtk4 libllvm3.8 libqmi-glib1
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 4 not upgraded.
admins@ssn-c6:~$ sudo apt-get install openssh-client
Reading package lists... Done
Building dependency tree
Reading state information... Done
openssh-client is already the newest version (1:7.2p2-4ubuntu2.7).
openssh-client set to manually installed.
The following packages were automatically installed and are no longer required:
  libdbusmenu-gtk4 libllvm3.8 libqmi-glib1
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 4 not upgraded.
admins@ssn-c6:~$
admins@ssn-c6:~$ sudo apt-get install vi
Display all 127 possibilities? (y or n)
admins@ssn-c6:~$ sudo apt-get install virtu
```

Figure 2: Install openssh-server,openssh-client in Ubuntu OS.



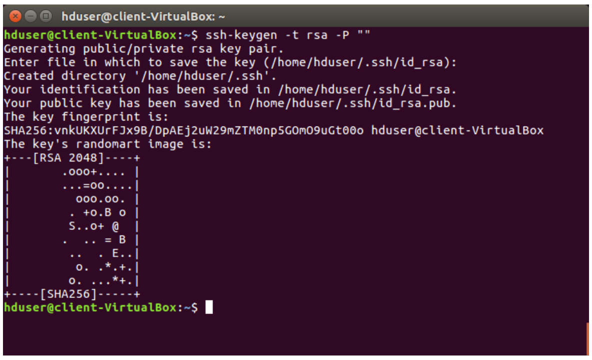
```
hduser@client-VirtualBox:~/jdk1.8.0_171/bin$ export JAVA_HOME=/home/hduser/jdk1.
8.0_171/bin
hduser@client-VirtualBox:~/jdk1.8.0_171/bin$ export PATH=$PATH:$JAVA_HOME
hduser@client-VirtualBox:~/jdk1.8.0_171/bin$ echo $PATH
/usr/local/bin:/usr/bin:/bin:/usr/local/games:/usr/games:/snap/bin:/home/hduser/
jdk1.8.0_171/bin:/home/hduser/jdk1.8.0_171/bin
hduser@client-VirtualBox:~/jdk1.8.0_171/bin$ 
```

Figure 3: Setting Java Home environment variable to the specified download path of JDK-1.7.



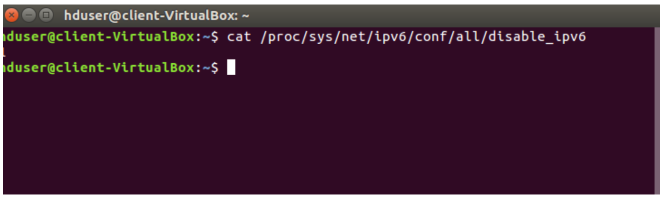
```
client@client-VirtualBox:~$ sudo addgroup hadoop
Adding group `hadoop' (GID 1001) ...
Done.
client@client-VirtualBox:~$ sudo adduser --ingroup hadoop hduser
Adding user `hduser' ...
Adding new user `hduser' (1001) with group `hadoop' ...
Creating home directory `/home/hduser' ...
Copying files from `/etc/skel' ...
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
Changing the user information for hduser
Enter the new value, or press ENTER for the default
        Full Name []:
        Room Number []:
        Work Phone []:
        Home Phone []:
        Other []:
Is the information correct? [Y/n]
client@client-VirtualBox:~$
client@client-VirtualBox:~$ su - hduser
```

Figure 4: Adding a dedicated hadoop system user.



```
hduser@client-VirtualBox: ~
hduser@client-VirtualBox:~$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hduser/.ssh/id_rsa):
Created directory '/home/hduser/.ssh'.
Your identification has been saved in /home/hduser/.ssh/id_rsa.
Your public key has been saved in /home/hduser/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:vnkUKXUrFJx9B/DpAEj2uW29mZTM0np5GOmO9uGt00o hduser@client-VirtualBox
The key's randomart image is:
+---[RSA 2048]----+
|        .ooo+.... |
|        ...=oo....|
|          ooo.oo. |
|         . +o.B o |
|         S..o+ @  |
|        .   .. = B |
|         ..   . E..|
|          o.  .*.+.|
|         o.  ...*+.|
+----[SHA256]-----+
hduser@client-VirtualBox:~$ 
```

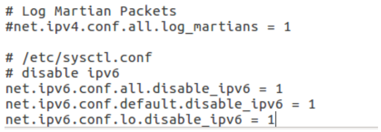
Figure 5: Configuring SSH in newly created user.



```
hduser@client-VirtualBox: ~
duser@client-VirtualBox:~$ cat /proc/sys/net/ipv6/conf/all/disable_ipv6

duser@client-VirtualBox:~$ 
```

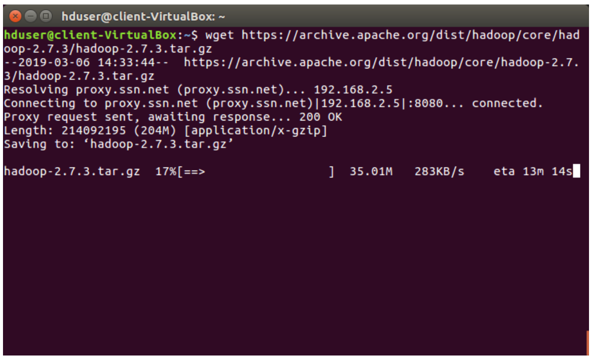
Figure 6: Disabling IPv6 in the newly created user account.



```
# Log Martian Packets
#net.ipv4.conf.all.log_martians = 1

# /etc/sysctl.conf
# disable ipv6
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1
```

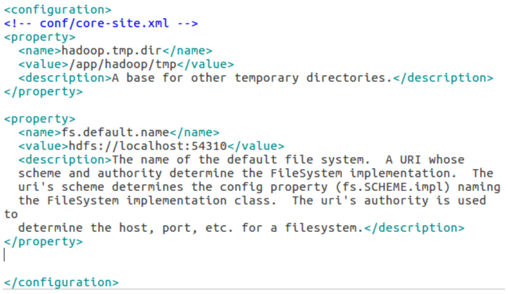
Figure 7: Disabling IPv6 in the newly created user account.



```
hduser@client-VirtualBox:~$ wget https://archive.apache.org/dist/hadoop/core/had
oop-2.7.3/hadoop-2.7.3.tar.gz
--2019-03-06 14:33:44--  https://archive.apache.org/dist/hadoop/core/hadoop-2.7.
3/hadoop-2.7.3.tar.gz
Resolving proxy.ssn.net (proxy.ssn.net)... 192.168.2.5
Connecting to proxy.ssn.net (proxy.ssn.net)|192.168.2.5|:8080... connected.
Proxy request sent, awaiting response... 200 OK
Length: 214092195 (204M) [application/x-gzip]
Saving to: ‘hadoop-2.7.3.tar.gz’

hadoop-2.7.3.tar.gz  17%[==>                ]  35.01M   283KB/s    eta 13m 14s
```

Figure 8: Installation of Hadoop 2.7.3 in new user login.

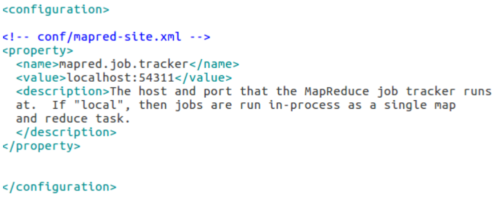


```
<configuration>
<!-- conf/core-site.xml -->
<property>
  <name>hadoop.tmp.dir</name>
  <value>/app/hadoop/tmp</value>
  <description>A base for other temporary directories.</description>
</property>

<property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:54310</value>
  <description>The name of the default file system.  A URI whose
  scheme and authority determine the FileSystem implementation.  The
  uri's scheme determines the config property (fs.SCHEME.impl) naming
  the FileSystem implementation class.  The uri's authority is used
to
  determine the host, port, etc. for a filesystem.</description>
</property>


</configuration>
```

Figure 9: Configuring hadoop core-site.xml .

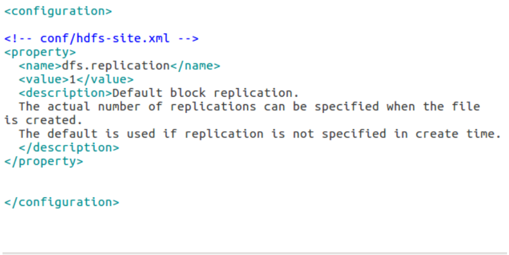


```
<configuration>

<!-- conf/mapred-site.xml -->
<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>
```

Figure 10: Configuring Hadoop MapReduce.

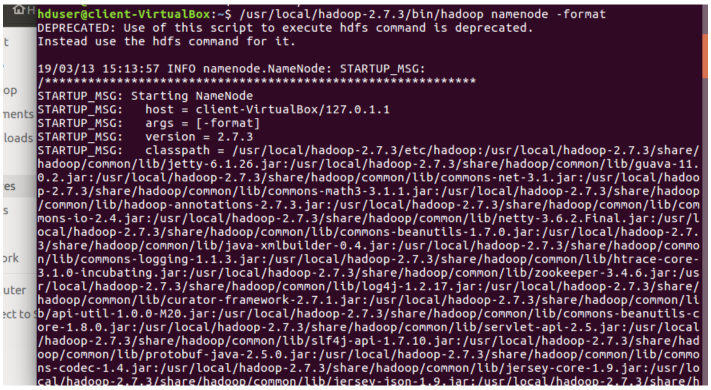


```
<configuration>

<!-- conf/hdfs-site.xml -->
<property>
  <name>dfs.replication</name>
  <value>1</value>
  <description>Default block replication.
  The actual number of replications can be specified when the file
is created.
  The default is used if replication is not specified in create time.
  </description>
</property>


</configuration>
```

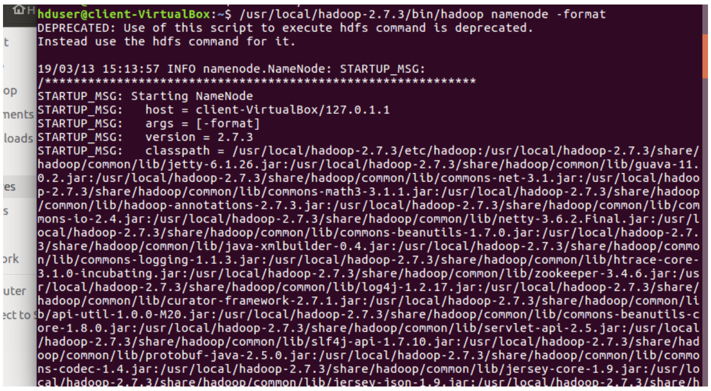
Figure 11: Configuring Hadoop HDFS Site.



```
hduser@client-VirtualBox:~$ /usr/local/hadoop-2.7.3/bin/hadoop namenode -format
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/03/13 15:13:57 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = client-VirtualBox/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.7.3
STARTUP_MSG:   classpath = /usr/local/hadoop-2.7.3/etc/hadoop:/usr/local/hadoop-2.7.3/share/
hadoop/common/lib/jetty-6.1.26.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/guava-11.
0.2.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/commons-net-3.1.jar:/usr/local/hadoo
p-2.7.3/share/hadoop/common/lib/commons-math3-3.1.1.jar:/usr/local/hadoop-2.7.3/share/hadoop
/common/lib/hadoop-annotations-2.7.3.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/com
mons-io-2.4.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/netty-3.6.2.Final.jar:/usr/l
ocal/hadoop-2.7.3/share/hadoop/common/lib/commons-beanutils-1.7.0.jar:/usr/local/hadoop-2.7.
3/share/hadoop/common/lib/java-xmlbuilder-0.4.jar:/usr/local/hadoop-2.7.3/share/hadoop/commo
n/lib/commons-logging-1.1.3.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/htrace-core-
3.1.0-incubating.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/zookeeper-3.4.6.jar:/us
r/local/hadoop-2.7.3/share/hadoop/common/lib/log4j-1.2.17.jar:/usr/local/hadoop-2.7.3/share/
hadoop/common/lib/curator-framework-2.7.1.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/li
b/api-util-1.0.0-M20.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/commons-beanutils-c
ore-1.8.0.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/servlet-api-2.5.jar:/usr/local
/hadoop-2.7.3/share/hadoop/common/lib/slf4j-api-1.7.10.jar:/usr/local/hadoop-2.7.3/share/had
oop/common/lib/protobuf-java-2.5.0.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/commo
ns-codec-1.4.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/jersey-core-1.9.jar:/usr/lo
cal/hadoop-2.7.3/share/hadoop/common/lib/jersey-json-1.9.jar:/usr/local/hadoop-2.7.3/share/h
```

Figure 12: Formatting HDFS file system via the NameNode.



```
hduser@client-VirtualBox:~$ /usr/local/hadoop-2.7.3/bin/hadoop namenode -format
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/03/13 15:13:57 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = client-VirtualBox/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.7.3
STARTUP_MSG:   classpath = /usr/local/hadoop-2.7.3/etc/hadoop:/usr/local/hadoop-2.7.3/share/
hadoop/common/lib/jetty-6.1.26.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/guava-11.
0.2.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/commons-net-3.1.jar:/usr/local/hadoo
p-2.7.3/share/hadoop/common/lib/commons-math3-3.1.1.jar:/usr/local/hadoop-2.7.3/share/hadoop
/common/lib/hadoop-annotations-2.7.3.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/com
mons-io-2.4.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/netty-3.6.2.Final.jar:/usr/l
ocal/hadoop-2.7.3/share/hadoop/common/lib/commons-beanutils-1.7.0.jar:/usr/local/hadoop-2.7.
3/share/hadoop/common/lib/java-xmlbuilder-0.4.jar:/usr/local/hadoop-2.7.3/share/hadoop/commo
n/lib/commons-logging-1.1.3.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/htrace-core-
3.1.0-incubating.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/zookeeper-3.4.6.jar:/us
r/local/hadoop-2.7.3/share/hadoop/common/lib/log4j-1.2.17.jar:/usr/local/hadoop-2.7.3/share/
hadoop/common/lib/curator-framework-2.7.1.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/li
b/api-util-1.0.0-M20.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/commons-beanutils-c
ore-1.8.0.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/servlet-api-2.5.jar:/usr/local
/hadoop-2.7.3/share/hadoop/common/lib/slf4j-api-1.7.10.jar:/usr/local/hadoop-2.7.3/share/had
oop/common/lib/protobuf-java-2.5.0.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/commo
ns-codec-1.4.jar:/usr/local/hadoop-2.7.3/share/hadoop/common/lib/jersey-core-1.9.jar:/usr/lo
cal/hadoop-2.7.3/share/hadoop/common/lib/jersey-json-1.9.jar:/usr/local/hadoop-2.7.3/share/h
```

Figure 13: Starting hadoop NameNode,Datanode,JobTracker and TaskTracker.

Result

Thus the hadoop single node cluster is sucessfully created in Ubuntu 16.04 OS version and required packages are installed.