Hadoop Cluster

installation

Outline

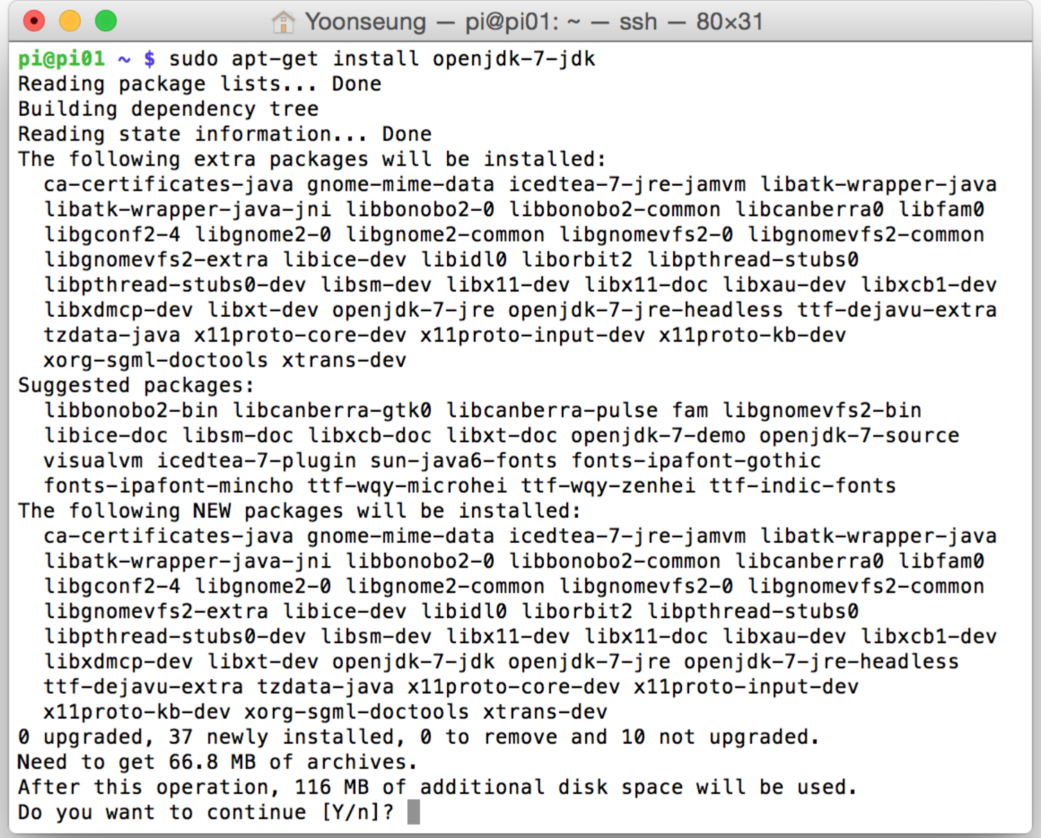
- For Raspberry Pi
  - Install JAVA & HADOOP
  - Set Path
  - Wordcount Example
- For Ubuntu 14.04

For Raspberry Pi

Hadoop MapReduce Cluster installation

Install JAVA Developer Kit

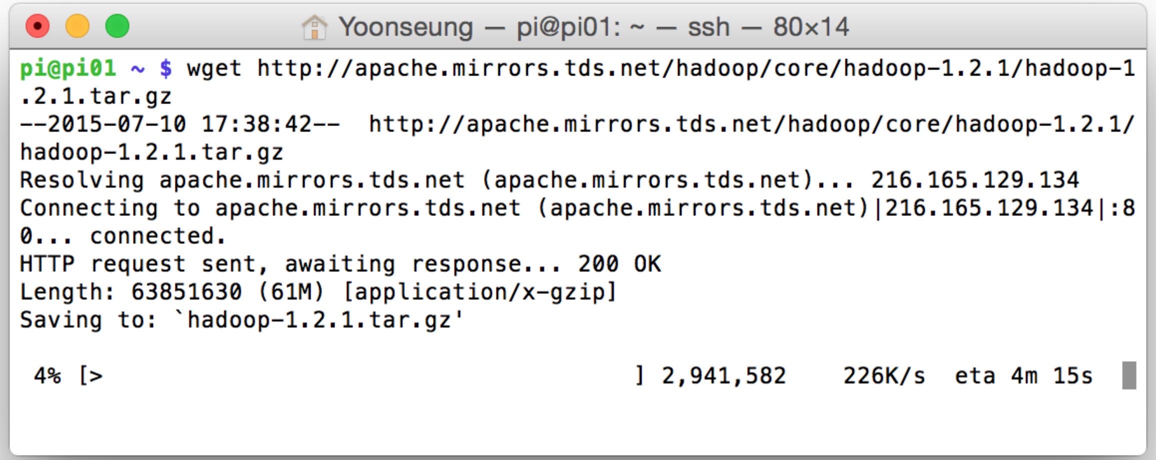
- sudo apt-get install openjdk-7-jdk



```
pi@pi01 ~ $ sudo apt-get install openjdk-7-jdk
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following extra packages will be installed:
  ca-certificates-java gnome-mime-data icedtea-7-jre-jamvm libatk-wrapper-java
  libatk-wrapper-java-jni libbonobo2-0 libbonobo2-common libcanberra0 libfam0
  libgconf2-4 libgnome2-0 libgnome2-common libgnomevfs2-0 libgnomevfs2-common
  libgnomevfs2-extra libice-dev libidl0 liborbit2 libpthread-stubs0
  libpthread-stubs0-dev libsm-dev libx11-dev libx11-doc libxau-dev libxcb1-dev
  libxdmcp-dev libxt-dev openjdk-7-jre openjdk-7-jre-headless ttf-dejavu-extra
  tzdata-java x11proto-core-dev x11proto-input-dev x11proto-kb-dev
  xorg-sgml-doctools xtrans-dev
Suggested packages:
  libbonobo2-bin libcanberra-gtk0 libcanberra-pulse fam libgnomevfs2-bin
  libice-doc libsm-doc libxcb-doc libxt-doc openjdk-7-demo openjdk-7-source
  visualvm icedtea-7-plugin sun-java6-fonts fonts-ipafont-gothic
  fonts-ipafont-mincho ttf-wqy-microhei ttf-wqy-zenhei ttf-indic-fonts
The following NEW packages will be installed:
  ca-certificates-java gnome-mime-data icedtea-7-jre-jamvm libatk-wrapper-java
  libatk-wrapper-java-jni libbonobo2-0 libbonobo2-common libcanberra0 libfam0
  libgconf2-4 libgnome2-0 libgnome2-common libgnomevfs2-0 libgnomevfs2-common
  libgnomevfs2-extra libice-dev libidl0 liborbit2 libpthread-stubs0
  libpthread-stubs0-dev libsm-dev libx11-dev libx11-doc libxau-dev libxcb1-dev
  libxdmcp-dev libxt-dev openjdk-7-jdk openjdk-7-jre openjdk-7-jre-headless
  ttf-dejavu-extra tzdata-java x11proto-core-dev x11proto-input-dev
  x11proto-kb-dev xorg-sgml-doctools xtrans-dev
0 upgraded, 37 newly installed, 0 to remove and 10 not upgraded.
Need to get 66.8 MB of archives.
After this operation, 116 MB of additional disk space will be used.
Do you want to continue [Y/n]?
```

Install Hadoop

- wget <link for hadoop>
- sudo tar vxzf <filename> -C /usr/local



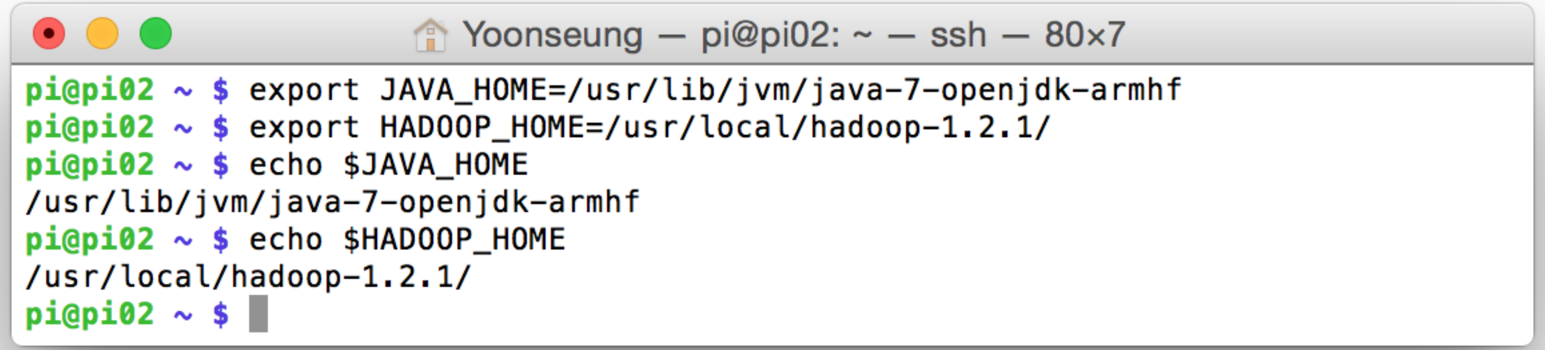
```
pi@pi01 ~ $ wget http://apache.mirrors.tds.net/hadoop/core/hadoop-1.2.1/hadoop-1
.2.1.tar.gz
--2015-07-10 17:38:42--  http://apache.mirrors.tds.net/hadoop/core/hadoop-1.2.1/
hadoop-1.2.1.tar.gz
Resolving apache.mirrors.tds.net (apache.mirrors.tds.net)... 216.165.129.134
Connecting to apache.mirrors.tds.net (apache.mirrors.tds.net)|216.165.129.134|:8
0... connected.
HTTP request sent, awaiting response... 200 OK
Length: 63851630 (61M) [application/x-gzip]
Saving to: `hadoop-1.2.1.tar.gz'

 4% [>                                        ] 2,941,582    226K/s  eta 4m 15s
```

ex >
wget https://archive.apache.org/dist/hadoop/common/hadoop-1.2.1/hadoop-1.2.1.tar.gz
sudo tar vxzf hadoop-1.2.1.tar.gz -C /usr/local

Set Path

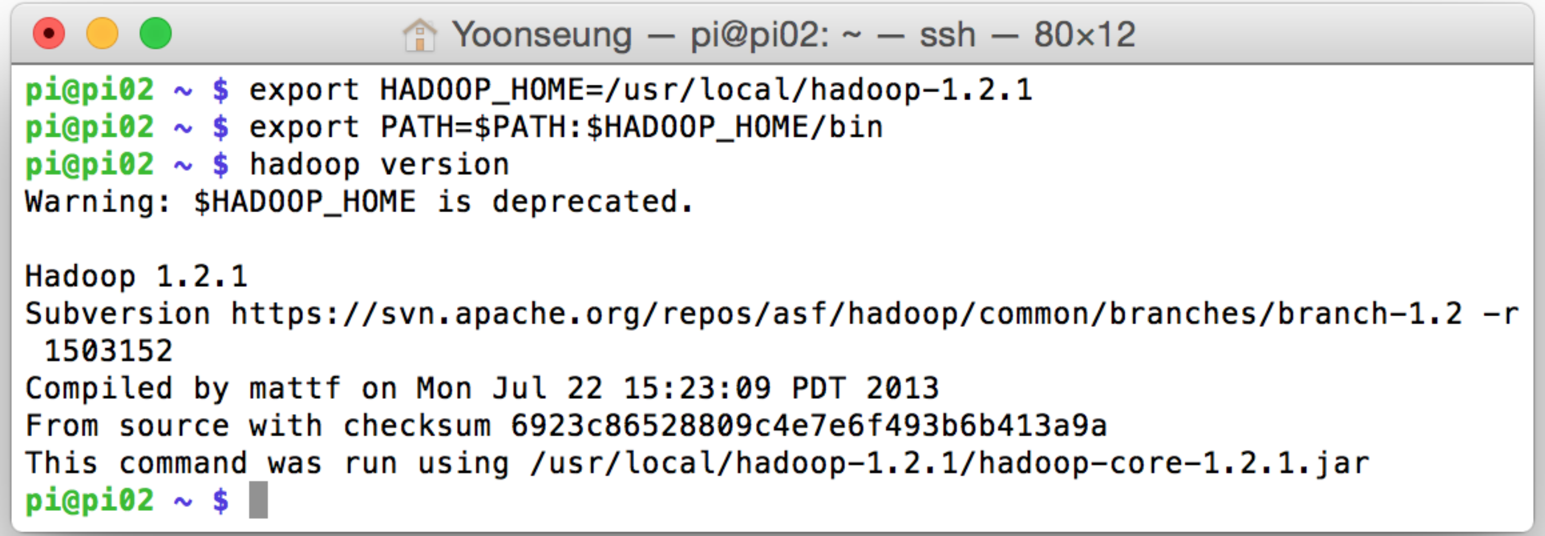
- export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-armhf
- export HADOOP_HOME=/usr/local/<hadoop ver>



```
pi@pi02 ~ $ export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-armhf
pi@pi02 ~ $ export HADOOP_HOME=/usr/local/hadoop-1.2.1/
pi@pi02 ~ $ echo $JAVA_HOME
/usr/lib/jvm/java-7-openjdk-armhf
pi@pi02 ~ $ echo $HADOOP_HOME
/usr/local/hadoop-1.2.1/
pi@pi02 ~ $ 
```

Set Path

- export PATH=$PATH:$HADOOP_HOME/bin



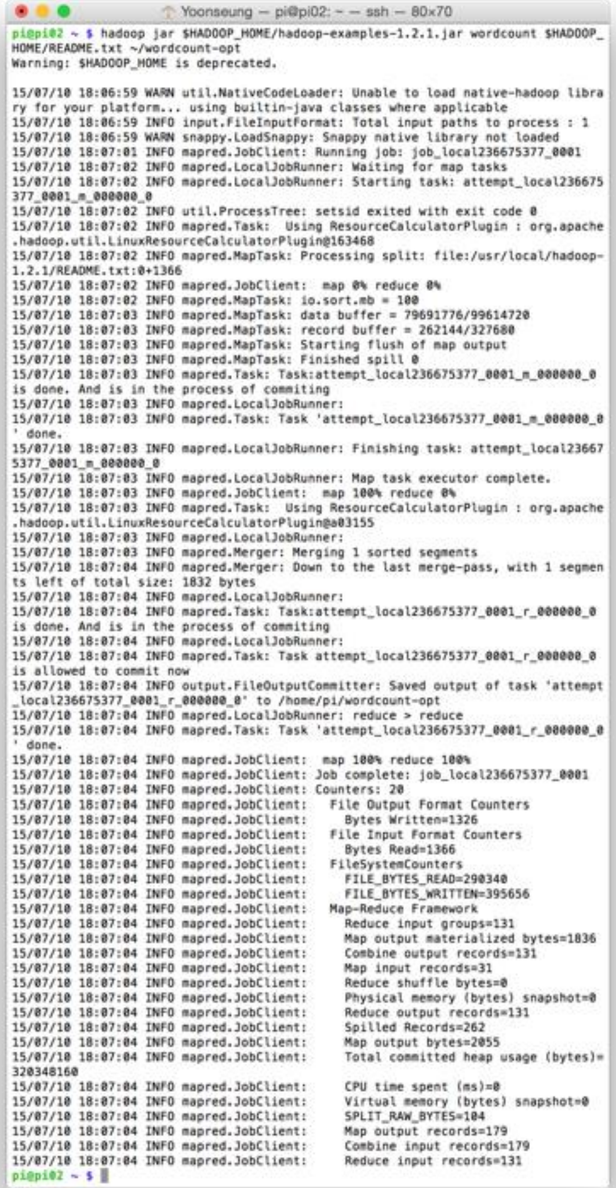
```
pi@pi02 ~ $ export HADOOP_HOME=/usr/local/hadoop-1.2.1
pi@pi02 ~ $ export PATH=$PATH:$HADOOP_HOME/bin
pi@pi02 ~ $ hadoop version
Warning: $HADOOP_HOME is deprecated.

Hadoop 1.2.1
Subversion https://svn.apache.org/repos/asf/hadoop/common/branches/branch-1.2 -r
 1503152
Compiled by mattf on Mon Jul 22 15:23:09 PDT 2013
From source with checksum 6923c86528809c4e7e6f493b6b413a9a
This command was run using /usr/local/hadoop-1.2.1/hadoop-core-1.2.1.jar
pi@pi02 ~ $ 
```

Wordcount example

for single node

- hadoop jar $HADOOP_HOME/hadoop-examples-1.2.1.jar wordcount $HADOOP_HOME/README.txt ~/wordcount-opt
- cat ~/wordcount-opt/part-r-00000



```
pi@pi02 ~ $ hadoop jar $HADOOP_HOME/hadoop-examples-1.2.1.jar wordcount $HADOOP_
HOME/README.txt ~/wordcount-opt
Warning: $HADOOP_HOME is deprecated.

15/07/10 18:06:59 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
15/07/10 18:06:59 INFO input.FileInputFormat: Total input paths to process : 1
15/07/10 18:06:59 WARN snappy.LoadSnappy: Snappy native library not loaded
15/07/10 18:07:01 INFO mapred.JobClient: Running job: job_local236675377_0001
15/07/10 18:07:02 INFO mapred.LocalJobRunner: Waiting for map tasks
15/07/10 18:07:02 INFO mapred.LocalJobRunner: Starting task: attempt_local236675
377_0001_m_000000_0
15/07/10 18:07:02 INFO util.ProcessTree: setsid exited with exit code 0
15/07/10 18:07:02 INFO mapred.Task:  Using ResourceCalculatorPlugin : org.apache
.hadoop.util.LinuxResourceCalculatorPlugin@163468
15/07/10 18:07:02 INFO mapred.MapTask: Processing split: file:/usr/local/hadoop-
1.2.1/README.txt:0+1366
15/07/10 18:07:02 INFO mapred.JobClient:  map 0% reduce 0%
15/07/10 18:07:02 INFO mapred.MapTask: io.sort.mb = 100
15/07/10 18:07:03 INFO mapred.MapTask: data buffer = 79691776/99614720
15/07/10 18:07:03 INFO mapred.MapTask: record buffer = 262144/327680
15/07/10 18:07:03 INFO mapred.MapTask: Starting flush of map output
15/07/10 18:07:03 INFO mapred.MapTask: Finished spill 0
15/07/10 18:07:03 INFO mapred.Task: Task:attempt_local236675377_0001_m_000000_0
is done. And is in the process of commiting
15/07/10 18:07:03 INFO mapred.LocalJobRunner:
15/07/10 18:07:03 INFO mapred.Task: Task 'attempt_local236675377_0001_m_000000_0
' done.
15/07/10 18:07:03 INFO mapred.LocalJobRunner: Finishing task: attempt_local23667
5377_0001_m_000000_0
15/07/10 18:07:03 INFO mapred.LocalJobRunner: Map task executor complete.
15/07/10 18:07:03 INFO mapred.JobClient:  map 100% reduce 0%
15/07/10 18:07:03 INFO mapred.Task:  Using ResourceCalculatorPlugin : org.apache
.hadoop.util.LinuxResourceCalculatorPlugin@a03155
15/07/10 18:07:03 INFO mapred.LocalJobRunner:
15/07/10 18:07:03 INFO mapred.Merger: Merging 1 sorted segments
15/07/10 18:07:04 INFO mapred.Merger: Down to the last merge-pass, with 1 segmen
ts left of total size: 1832 bytes
15/07/10 18:07:04 INFO mapred.LocalJobRunner:
15/07/10 18:07:04 INFO mapred.Task: Task:attempt_local236675377_0001_r_000000_0
is done. And is in the process of commiting
15/07/10 18:07:04 INFO mapred.LocalJobRunner:
15/07/10 18:07:04 INFO mapred.Task: Task attempt_local236675377_0001_r_000000_0
is allowed to commit now
15/07/10 18:07:04 INFO output.FileOutputCommitter: Saved output of task 'attempt
_local236675377_0001_r_000000_0' to /home/pi/wordcount-opt
15/07/10 18:07:04 INFO mapred.LocalJobRunner: reduce > reduce
15/07/10 18:07:04 INFO mapred.Task: Task 'attempt_local236675377_0001_r_000000_0
' done.
15/07/10 18:07:04 INFO mapred.JobClient:  map 100% reduce 100%
15/07/10 18:07:04 INFO mapred.JobClient: Job complete: job_local236675377_0001
15/07/10 18:07:04 INFO mapred.JobClient: Counters: 20
15/07/10 18:07:04 INFO mapred.JobClient:   File Output Format Counters
15/07/10 18:07:04 INFO mapred.JobClient:     Bytes Written=1326
15/07/10 18:07:04 INFO mapred.JobClient:   File Input Format Counters
15/07/10 18:07:04 INFO mapred.JobClient:     Bytes Read=1366
15/07/10 18:07:04 INFO mapred.JobClient:   FileSystemCounters
15/07/10 18:07:04 INFO mapred.JobClient:     FILE_BYTES_READ=290340
15/07/10 18:07:04 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=395656
15/07/10 18:07:04 INFO mapred.JobClient:   Map-Reduce Framework
15/07/10 18:07:04 INFO mapred.JobClient:     Reduce input groups=131
15/07/10 18:07:04 INFO mapred.JobClient:     Map output materialized bytes=1836
15/07/10 18:07:04 INFO mapred.JobClient:     Combine output records=131
15/07/10 18:07:04 INFO mapred.JobClient:     Map input records=31
15/07/10 18:07:04 INFO mapred.JobClient:     Reduce shuffle bytes=0
15/07/10 18:07:04 INFO mapred.JobClient:     Physical memory (bytes) snapshot=0
15/07/10 18:07:04 INFO mapred.JobClient:     Reduce output records=131
15/07/10 18:07:04 INFO mapred.JobClient:     Spilled Records=262
15/07/10 18:07:04 INFO mapred.JobClient:     Map output bytes=2055
15/07/10 18:07:04 INFO mapred.JobClient:     Total committed heap usage (bytes)=
320348160
15/07/10 18:07:04 INFO mapred.JobClient:     CPU time spent (ms)=0
15/07/10 18:07:04 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=0
15/07/10 18:07:04 INFO mapred.JobClient:     SPLIT_RAW_BYTES=104
15/07/10 18:07:04 INFO mapred.JobClient:     Map output records=179
15/07/10 18:07:04 INFO mapred.JobClient:     Combine input records=179
15/07/10 18:07:04 INFO mapred.JobClient:     Reduce input records=131
pi@pi02 ~ $
```

Distribute System [Master]

- sudo vi /etc/ssh/sshd_config
  PubkeyAuthentication yes AuthorizedKeysFile .ssh/authorized_keys
- mkdir ~/.ssh
- ssh-keygen -t rsa -P ""
- cp /home/pi/.ssh/id_rsa.pub /pi/stat/.ssh/authorized_keys

For Ubuntu 14.04

Hadoop MapReduce Cluster installation

configuration

- **Console mode booting**

  ```
  $ sudo vi /etc/default/grub
  ```

  -> changes these following lines

  ```
  GRUB_CMDLINE_LINUX_DEFAULT="”
  GRUB_CMDLINE_LINUX="text”
  ```

  -> after that, update conf & reboot

  ```
  $ sudo update-grub
  $ sudo reboot
  ```

- **Root password setting**

  ```
  $ sudo passwd root
  ```

configuration

- **Install java**

```
$ sudo add-apt-repository ppa:webupd8team/java
$ sudo apt-get update
$ sudo apt-get install oracle-jdk7-installer
```

- **Download & Install Hadoop**

```
$ wget <http link>
$ cp hadoop-x.x.x.tar.gz  /usr/local
$ rm hadoop-x.x.x.tar.gz
$ cd /usr/local
$ tar zxvf hadoop-x.x.x.tar.gz
```

configuration

- **PATH setting**

  ```
  $ sudo vi ~/.profile
  ```

  -> add these following lines

  ```
  export JAVA_HOME=/usr/lib/jvm/java-7-oracle
  export HADOOP_HOME=/usr/local/hadoop-1.2.1
  export PATH=$PATH:$HADOOP_HOME/bin
  ```

  -> add path by following command

  ```
  $ source ~/.profile
  ```

- **Check the correct PATH by following command**

  ```
  $ echo $HADOOP_HOME
  ```

configuration

- **Install ssh software**

```
$ sudo apt-get install ssh
$ sudo apt-get install rsync
```

- **Setup passphraseless ssh**

```
$ ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsa
$ cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys
```

MapReduce configuration

mapred-site.xml

- io.sort.factor (default value= 10)
  : The number of streams to merge at once while sorting files. This determines the number of open file handles.
- io.sort.mb (default value= 100)
  : The total amount of buffer memory to use while sorting files, in megabytes. By default, gives each merge stream 1MB, which should minimize seeks.