Hadoop 1.2.1 installation

Chun-Chen Tu
timtu@umich.edu

Before installation

- Where to get hadoop
  - http://ftp.twaren.net/Unix/Web/apache/hadoop/common/hadoop-1.2.1/
  - or my ftp://hadoop:hahahadoop@140.113.114.104
  - Please download: hadoop-1.2.1.tar.gz
- GUI mode may help for typing commands.
- List of commands are also on ftp:
  - the file cmd
- In this ppt, commands will be shown in italic and purple color
  - *mkdir hadoop*

After login, install required packages first

sudo apt-get install libssl-dev rsync g++

type “y” when asked

```
hadoop@ubuntu:~$ ls
Desktop  Documents  Downloads  Music  Pictures  Public  Templates  Videos
hadoop@ubuntu:~$ sudo apt-get install libssl-dev rsync g++
[sudo] password for hadoop:
Reading package lists... Done
Building dependency tree
Reading state information... Done
rsync is already the newest version.
The following extra packages will be installed:
  g++-4.6 libssl-doc libstdc++6-4.6-dev zlib1g-dev
Suggested packages:
  g++-multilib g++-4.6-multilib gcc-4.6-doc libstdc++6-4.6-dbg
  libstdc++6-4.6-doc
The following NEW packages will be installed:
  g++ g++-4.6 libssl-dev libssl-doc libstdc++6-4.6-dev zlib1g-dev
0 upgraded, 6 newly installed, 0 to remove and 61 not upgraded.
Need to get 11.4 MB of archives.
After this operation, 33.5 MB of additional disk space will be used.
Do you want to continue [Y/n]?
```

Download files:

cd Downloads

wget ftp://hadoop:hahahadoop@140.113.114.104/hadoop-1.2.1.tar.gz

wget ftp://hadoop:hahahadoop@140.113.114.104/jdk-7u45-linux-x64.gz

```
Connecting to 140.113.114.104:21... connected.
Logging in as hadoop ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD not needed.
==> SIZE hadoop-1.2.1.tar.gz ... 63851630
==> PASV ... done.    ==> RETR hadoop-1.2.1.tar.gz ... done.
Length: 63851630 (61M) (unauthoritative)

100%[=====================================>] 63,851,630  81.2M/s   in 0.8s

2013-12-29 12:32:57 (81.2 MB/s) - `hadoop-1.2.1.tar.gz' saved [63851630]

hadoop@ubuntu:~/Downloads$ wget ftp://hadoop:hahahadoop@140.113.114.104/jdk-7u45
-linux-x64.gz
--2013-12-29 12:34:05--  ftp://hadoop:*password*@140.113.114.104/jdk-7u45-linux-
x64.gz
           => `jdk-7u45-linux-x64.gz'
Connecting to 140.113.114.104:21... connected.
Logging in as hadoop ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD not needed.
==> SIZE jdk-7u45-linux-x64.gz ... 138094686
==> PASV ... done.    ==> RETR jdk-7u45-linux-x64.gz ... done.
Length: 138094686 (132M) (unauthoritative)

100%[=====================================>] 138,094,686 85.3M/s   in 1.5s

2013-12-29 12:34:07 (85.3 MB/s) - `jdk-7u45-linux-x64.gz' saved [138094686]

hadoop@ubuntu:~/Downloads$
```

Install java : reference website
(Under Downloads folder)
tar -zxvf jdk-7u45-linux-x64.gz
sudo mkdir /usr/lib/jdk
sudo cp -r jdk1.7.0_45 /usr/lib/jdk/

Edit profile:
sudo vim /etc/profile
(add four lines in at the end of profile)

```
# The default umask is now handled by pam_umask.
# See pam_umask(8) and /etc/login.defs.

if [ -d /etc/profile.d ]; then
  for i in /etc/profile.d/*.sh; do
    if [ -r $i ]; then
      . $i
    fi
  done
  unset i
fi

export JAVA_HOME=/usr/lib/jdk/jdk1.7.0_45
export JRE_HOME=/usr/lib/jdk/jdk1.7.0_45/jre
export PATH=$JAVA_HOME/bin:$JAVA_HOME/jre/bin:$PATH
export CLASSPATH=$CLASSPATH:.:$JAVA_HOME/lib:$JAVA_HOME/jre/lib
```

export JAVA_HOME=/usr/lib/jdk/jdk1.7.0_45
export JRE_HOME=/usr/lib/jdk/jdk1.7.0_45/jre
export PATH=$JAVA_HOME/bin:$JAVA_HOME/jre/bin:$PATH
export CLASSPATH=$CLASSPATH:.:$JAVA_HOME/lib:$JAVA_HOME/jre/lib

Config java:
sudo update-alternatives --install /usr/bin/java java /usr/lib/jdk/jdk1.7.0_45/bin/java 300
sudo update-alternatives --install /usr/bin/javac javac /usr/lib/jdk/jdk1.7.0_45/bin/javac 300
sudo update-alternatives --config java
sudo update-alternatives --config javac

Test it with version
java –version
You will see the version information if success.

```
hadoop@ubuntu:~/Downloads$ java -version
java version "1.7.0_45"
Java(TM) SE Runtime Environment (build 1.7.0_45-b18)
Java HotSpot(TM) 64-Bit Server VM (build 24.45-b08, mixed mode)
hadoop@ubuntu:~/Downloads$
```

SSH setting: SSH setting is optional but is recommended if you don't want to enter password every time.

Generate RSA key

ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa

Copy public key

cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

```
hadoop@ubuntu:~/Downloads$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/hadoop/.ssh'.
Your identification has been saved in /home/hadoop/.ssh/id_rsa.
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub.
The key fingerprint is:
eb:28:46:7d:29:87:08:63:cd:d0:f5:a8:76:82:13:2a hadoop@ubuntu
The key's randomart image is:
+--[ RSA 2048]----+
|    . ..          |
|   . .  o         |
|   .+  . .        |
| .+oo.            |
|E.oo+o..S.        |
|.  ooo+ +.        |
|    .   +.        |
|     o  o         |
|    . .. .        |
+-----------------+
hadoop@ubuntu:~/Downloads$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hadoop@ubuntu:~/Downloads$ _
```

SSH test:

ssh hadoop@localhost

remember to exit

exit

You will be asked for the authenticity for the first time. After this connection, no more inquiring.

```
hadoop@ubuntu:~/Downloads$ ssh hadoop@localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is 7a:7b:0a:46:e7:d3:89:4d:78:3e:d6:54:36:9d:dc:57.
Are you sure you want to continue connecting (yes/no)? y
Please type 'yes' or 'no': yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 12.04.3 LTS (GNU/Linux 3.8.0-29-generic x86_64)

 * Documentation:  https://help.ubuntu.com/

  System information as of Sun Dec 29 12:54:06 CST 2013

  System load:  0.03               Processes:           83
  Usage of /:   20.0% of 18.25GB   Users logged in:     1
  Memory usage: 20%                IP address for eth0: 10.0.2.15
  Swap usage:   0%

  Graph this data and manage this system at https://landscape.canonical.com/

*** System restart required ***
Last login: Sat Dec 28 20:15:40 2013
hadoop@ubuntu:~$ exit
```

If you fail the setting, you will need to enter password.

```
hadoop@ubuntu:~/.ssh$ ssh hadoop@localhost
hadoop@localhost's password:
```

Install hadoop:

tar -zxvf hadoop-1.2.1.tar.gz

mv hadoop-1.2.1 ~/hadoop move it under home directory for convenience

vim ~/hadoop/conf/hadoop-env.sh edit hadoop environment shell script

export JAVA_HOME=/usr add this line

```
# The maximum amount of heap to use, in MB. Default is 1000.
# export HADOOP_HEAPSIZE=2000

# Extra Java runtime options.  Empty by default.
# export HADOOP_OPTS=-server

# Command specific options appended to HADOOP_OPTS when specified
export JAVA_HOME=/usr
export HADOOP_NAMENODE_OPTS="-Dcom.sun.management.jmxremote $HADOOP_NAMENODE_OPT
S"
export HADOOP_SECONDARYNAMENODE_OPTS="-Dcom.sun.management.jmxremote $HADOOP_SEC
ONDARYNAMENODE_OPTS"
export HADOOP_DATANODE_OPTS="-Dcom.sun.management.jmxremote $HADOOP_DATANODE_OPT
S"
export HADOOP_BALANCER_OPTS="-Dcom.sun.management.jmxremote $HADOOP_BALANCER_OPT
S"
export HADOOP_JOBTRACKER_OPTS="-Dcom.sun.management.jmxremote $HADOOP_JOBTRACKER
_OPTS"
# export HADOOP_TASKTRACKER_OPTS=
# The following applies to multiple commands (fs, dfs, fsck, distcp etc)
# export HADOOP_CLIENT_OPTS

# Extra ssh options.  Empty by default.
# export HADOOP_SSH_OPTS="-o ConnectTimeout=1 -o SendEnv=HADOOP_CONF_DIR"

# Where log files are stored.  $HADOOP_HOME/logs by default.
# export HADOOP_LOG_DIR=${HADOOP_HOME}/logs

# File naming remote slave hosts.  $HADOOP_HOME/conf/slaves by default.
-- INSERT --                                                    21,22         38%
```

Set environment PATH:

sudo vim ~/.bashrc configure bash setting

export PATH=/home/hadoop/hadoop/bin:$PATH add this line so we can use command hadoop everywhere

logout and then
re-login
the setting should take effect

type "*hadoop*" to try

```
# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

export PATH=/home/hadoop/hadoop/bin:$PATH

# If not running interactively, don't do anything
[ -z "$PS1" ] && return

# don't put duplicate lines or lines starting with space in the history.
# See bash(1) for more options
HISTCONTROL=ignoreboth

# append to the history file, don't overwrite it
shopt -s histappend

# for setting history length see HISTSIZE and HISTFILESIZE in bash(1)
HISTSIZE=1000
HISTFILESIZE=2000

# check the window size after each command and, if necessary,
# update the values of LINES and COLUMNS.
shopt -s checkwinsize

# If set, the pattern "**" used in a pathname expansion context will
# match all files and zero or more directories and subdirectories.
#shopt -s globstar

# make less more friendly for non-text input files, see lesspipe(1)
-- INSERT --                                                    5,42
```

Standalone mode: test if hadoop is available ref: website

cd ~/hadoop

mkdir input

cp conf/.xml input*

hadoop jar hadoop-examples-1.2.1.jar grep input output 'dfs[a-z.]+'

cat output/part-00000

```
13/12/29 12:42:20 INFO mapred.JobClient: Counters: 21
13/12/29 12:42:20 INFO mapred.JobClient:   File Input Format Counters
13/12/29 12:42:20 INFO mapred.JobClient:     Bytes Read=123
13/12/29 12:42:20 INFO mapred.JobClient:   File Output Format Counters
13/12/29 12:42:20 INFO mapred.JobClient:     Bytes Written=23
13/12/29 12:42:20 INFO mapred.JobClient:   FileSystemCounters
13/12/29 12:42:20 INFO mapred.JobClient:     FILE_BYTES_READ=610049
13/12/29 12:42:20 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=782159
13/12/29 12:42:20 INFO mapred.JobClient:   Map-Reduce Framework
13/12/29 12:42:20 INFO mapred.JobClient:     Map output materialized bytes=25
13/12/29 12:42:21 INFO mapred.JobClient:     Map input records=1
13/12/29 12:42:21 INFO mapred.JobClient:     Reduce shuffle bytes=0
13/12/29 12:42:21 INFO mapred.JobClient:     Spilled Records=2
13/12/29 12:42:21 INFO mapred.JobClient:     Map output bytes=17
13/12/29 12:42:21 INFO mapred.JobClient:     Total committed heap usage (bytes)=
452608000
13/12/29 12:42:21 INFO mapred.JobClient:     CPU time spent (ms)=0
13/12/29 12:42:21 INFO mapred.JobClient:     Map input bytes=25
13/12/29 12:42:21 INFO mapred.JobClient:     SPLIT_RAW_BYTES=108
13/12/29 12:42:21 INFO mapred.JobClient:     Combine input records=0
13/12/29 12:42:21 INFO mapred.JobClient:     Reduce input records=1
13/12/29 12:42:21 INFO mapred.JobClient:     Reduce input groups=1
13/12/29 12:42:21 INFO mapred.JobClient:     Combine output records=0
13/12/29 12:42:21 INFO mapred.JobClient:     Physical memory (bytes) snapshot=0
13/12/29 12:42:21 INFO mapred.JobClient:     Reduce output records=1
13/12/29 12:42:21 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=0
13/12/29 12:42:21 INFO mapred.JobClient:     Map output records=1
hadoop@ubuntu:~/hadoop$ cat output/part-00000
1       dfsadmin
hadoop@ubuntu:~/hadoop$
```

Pseudo-distributed configuration: reference website

edit 3 .xml files under conf folder

core-site.xml, hdfs-site.xml, mapred-site.xml

these files may also download from ftp

cd ~/hadoop

vim conf/core-site.xml

```
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

vim conf/hdfs-site.xml

```
<configuration>
 <property>
 <name>dfs.replication</name>
<value>1</value>
</property>
</configuration>
```

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>

  </property>
</configuration>
```

vim conf/mapred-site.xml

```
<configuration>
 <property>
<name>mapred.job.tracker</name>
<value>localhost:9001</value>
 </property>
</configuration>
```

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>mapred.job.tracker</name>
    <value>localhost:9001</value>
  </property>

</configuration>
```

HDFS format:

hadoop namenode -format

```
STARTUP_MSG:   java = 1.7.0_45
************************************************************/
13/12/29 12:50:56 INFO util.GSet: Computing capacity for map BlocksMap
13/12/29 12:50:56 INFO util.GSet: VM type       = 64-bit
13/12/29 12:50:56 INFO util.GSet: 2.0% max memory = 1013645312
13/12/29 12:50:56 INFO util.GSet: capacity      = 2^21 = 2097152 entries
13/12/29 12:50:56 INFO util.GSet: recommended=2097152, actual=2097152
13/12/29 12:50:57 INFO namenode.FSNamesystem: fsOwner=hadoop
13/12/29 12:50:57 INFO namenode.FSNamesystem: supergroup=supergroup
13/12/29 12:50:57 INFO namenode.FSNamesystem: isPermissionEnabled=true
13/12/29 12:50:57 INFO namenode.FSNamesystem: dfs.block.invalidate.limit=100
13/12/29 12:50:57 INFO namenode.FSNamesystem: isAccessTokenEnabled=false accessK
eyUpdateInterval=0 min(s), accessTokenLifetime=0 min(s)
13/12/29 12:50:57 INFO namenode.FSEditLog: dfs.namenode.edits.toleration.length
= 0
13/12/29 12:50:57 INFO namenode.NameNode: Caching file names occuring more than
10 times
13/12/29 12:50:57 INFO common.Storage: Image file /tmp/hadoop-hadoop/dfs/name/cu
rrent/fsimage of size 112 bytes saved in 0 seconds.
13/12/29 12:50:57 INFO namenode.FSEditLog: closing edit log: position=4, editlog
=/tmp/hadoop-hadoop/dfs/name/current/edits
13/12/29 12:50:57 INFO namenode.FSEditLog: close success: truncate to 4, editlog
=/tmp/hadoop-hadoop/dfs/name/current/edits
13/12/29 12:50:57 INFO common.Storage: Storage directory /tmp/hadoop-hadoop/dfs/
name has been successfully formatted.
13/12/29 12:50:57 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at ubuntu/127.0.1.1
************************************************************/
hadoop@ubuntu:~/hadoop/conf$
```

Start hadoop in pseudo-distributed mode:

start-all.sh

```
hadoop@ubuntu:~/hadoop/conf$ start-all.sh
starting namenode, logging to /home/hadoop/hadoop/libexec/../logs/hadoop-hadoop-
namenode-ubuntu.out
localhost: starting datanode, logging to /home/hadoop/hadoop/libexec/../logs/had
oop-hadoop-datanode-ubuntu.out
localhost: starting secondarynamenode, logging to /home/hadoop/hadoop/libexec/..
/logs/hadoop-hadoop-secondarynamenode-ubuntu.out
starting jobtracker, logging to /home/hadoop/hadoop/libexec/../logs/hadoop-hadoo
p-jobtracker-ubuntu.out
localhost: starting tasktracker, logging to /home/hadoop/hadoop/libexec/../logs/
hadoop-hadoop-tasktracker-ubuntu.out
hadoop@ubuntu:~/hadoop/conf$ _
```

type *jps* to see what's working

```
hadoop@ubuntu:~/hadoop/conf$ jps
5706 NameNode
6258 JobTracker
6182 SecondaryNameNode
6659 Jps
6491 TaskTracker
5938 DataNode
hadoop@ubuntu:~/hadoop/conf$
```

If you need to enter password, it's fine just inconvenient.
To solve this, please refer to SSH setting in previous slides.

```
hadoop@ubuntu:~$ start-all.sh
starting namenode, logging to /home/hadoop/hadoop/libexec/../logs/hadoop-hadoop-
namenode-ubuntu.out
hadoop@localhost's password:
localhost: starting datanode, logging to /home/hadoop/hadoop/libexec/../logs/had
oop-hadoop-datanode-ubuntu.out
hadoop@localhost's password:
localhost: starting secondarynamenode, logging to /home/hadoop/hadoop/libexec/..
/logs/hadoop-hadoop-secondarynamenode-ubuntu.out
starting jobtracker, logging to /home/hadoop/hadoop/libexec/../logs/hadoop-hadoo
p-jobtracker-ubuntu.out
hadoop@localhost's password:
localhost: starting tasktracker, logging to /home/hadoop/hadoop/libexec/../logs/
hadoop-hadoop-tasktracker-ubuntu.out
hadoop@ubuntu:~$
```

Let's run an example! There is an example jar file under ~/hadoop
hadoop-examples-1.2.1.jar

hadoop jar hadoop-examples-1.2.1.jar : to get more information

Now suppose we want to run the wordcount example.
First, put the input data on HDFS (you have to create your own input.txt first)
hadoop dfs –put input.txt /input.txt

Next, execute the wordcount example
hadoop jar hadoop-examples-1.2.1.jar wordcount /input.txt /test_out

Finally, get the results
hadoop dfs –get /test_out test_out

The result file part-r-00000 show up in the directory test_out

Run hadoop with C

- We need to use pipes provided by hadoop.
- Really slow!

Recompile library: website

vim ~/hadoop/src/c++/pipes/impl/HadoopPipes.cc

#include <unistd.h>

```
#include "hadoop/Pipes.hh"
#include "hadoop/SerialUtils.hh"
#include "hadoop/StringUtils.hh"

#include <unistd.h>
#include <map>
#include <vector>

#include <errno.h>
#include <netinet/in.h>
#include <stdint.h>
-- INSERT --                                    23,20          Top
```

cd ~/hadoop/src/c++/utils
chmod 755 configure
./configure
make install

cd ~/hadoop/src/c++/pipes
export LIBS=-lcrypto
chmod 755 configure
./configure
make install

New library will appear in ~/hadoop/src/c++/install

Compile wordcount example : website

wget –r –np –nH ftp://hadoop:hahahadoop@140.113.114.104/wordcount

make wordcount

hadoop dfs –mkdir test

hadoop dfs –put wordcount test/wordcount

hadoop dfs –put testdata.txt test/testdata.txt

hadoop pipes -D hadoop.pipes.java.recordreader=true -D hadoop.pipes.java.recordwriter=true -input test/testdata.txt -output test/output -program test/wordcount

hadoop dfs -get test/output output

cat output/part-00000

```
13/12/30 06:19:39 INFO mapred.JobClient:     Combine output records=0
13/12/30 06:19:39 INFO mapred.JobClient:     Physical memory (bytes) snapshot=50
7523072
13/12/30 06:19:39 INFO mapred.JobClient:     Reduce output records=12
13/12/30 06:19:39 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=292
3032576
13/12/30 06:19:39 INFO mapred.JobClient:     Map output records=13
hadoop@ubuntu:~/hadoop/Code/wordcount$ hadoop dfs -get test/output output
hadoop@ubuntu:~/hadoop/Code/wordcount$ ls
input.txt  Makefile  output  wordcount  wordcount.cpp
hadoop@ubuntu:~/hadoop/Code/wordcount$ cd output
hadoop@ubuntu:~/hadoop/Code/wordcount/output$ ls
_logs  part-00000  _SUCCESS
hadoop@ubuntu:~/hadoop/Code/wordcount/output$ cd ..
hadoop@ubuntu:~/hadoop/Code/wordcount$ ls
input.txt  Makefile  output  wordcount  wordcount.cpp
hadoop@ubuntu:~/hadoop/Code/wordcount$ cat output/part-00000
Hello,  1
The     1
first   1
five    1
for     1
give    1
hadoop  2
hello,  1
me      1
program.        1
success.        1
the     1
hadoop@ubuntu:~/hadoop/Code/wordcount$
```

Congratulation!!