LAB 2 INSTRUCTIONS

CKME 134 – BIG DATA ANALYTICS TOOLS

RYERSON UNIVERSITY

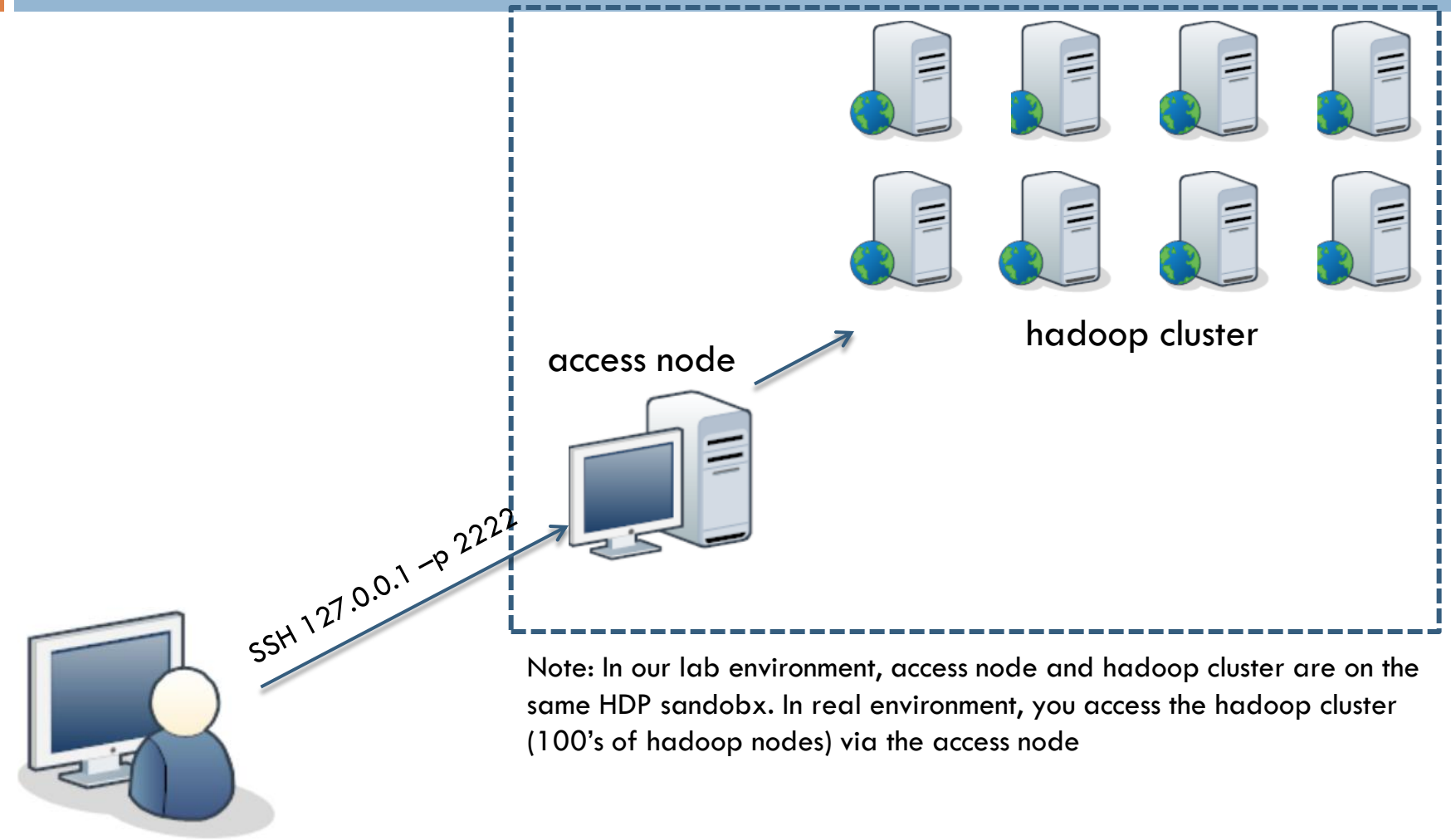
SPRING 2015

Instructor: Shaohua Zhang

Session 2 - Lab & Assignments

- Lab Computer
  - Username: datastudent
  - Password: datastudent
- HDP Sandbox
  - Username: root
  - Password: hadoop
- Lab 2
  - Download Geo-tagged Tweets data
  - Basic Linux Commands
  - Word count in linux

Lab Environment



Note: In our lab environment, access node and hadoop cluster are on the same HDP sandobx. In real environment, you access the hadoop cluster (100's of hadoop nodes) via the access node

Lab 2 – Download Dataset

- Dataset: twitter geo-tagged tweets
  - Download URL: http://www.ark.cs.cmu.edu/GeoText/
    - File to download: GeoText.2010-10-12.tgz
  - File to work with: full_text.txt
- Approach 1
  - Direct download from access node using Curl command
    - If you don't have Internet access on the sandbox access node, then you'll need to try approach 1
  - You'll need to unzip the file by using the tar command in linux
- Approach 2
  - Download to local computer and then upload via sftp
    - Usually the hadoop access node may not have direct web access. So you have to download the file to local computer first and then upload to access node or the cluster directly
  - After downloading the tgz file, you can use 7-zip tool to unzip the file first

Lab 2 – Before You Get Started…

- Open VBox and Start HDP Hadoop Sandbox
- Use putty to connect to the sandbox
- Create a directory '/home/lab' → we will put the data in this lab folder

```
[root@sandbox ~]# ll
total 32
-rw-------. 1 root    root    2143 Dec 16 18:13 anaconda-ks.cfg
-rw-r--r--. 1 root    root    9436 Dec 16 18:13 install.log
-rw-r--r--. 1 root    root    3314 Dec 16 18:12 install.log.syslog
drwxr-xr-x  8 root    root    4096 Dec 16 19:33 ranger_tutorial
lrwxrwxrwx  1 root    root      48 Dec 16 19:15 start_ambari.sh -> /usr/lib/hue/tools/start_scripts/start_ambari.sh
lrwxrwxrwx  1 root    root      47 Dec 16 19:17 start_hbase.sh -> /usr/lib/hue/tools/start_scripts/start_hbase.sh
-rwxrwxrwx  1 vagrant vagrant  241 Dec 16 19:15 start_solr.sh
-rwxrwxrwx  1 vagrant vagrant   63 Dec 16 19:17 stop_solr.sh
[root@sandbox ~]# pwd
/root
[root@sandbox ~]# cd /home
[root@sandbox home]# mkdir lab
[root@sandbox home]# cd lab
[root@sandbox lab]# pwd
/home/lab
[root@sandbox lab]# ll
total 0
[root@sandbox lab]#
```

→ list files in the current directory

→ show the current directory path

→ enter the home directory

→ create a directory called lab

→ enter the directory 'lab'

Lab 2

Approach 1 – download directly from access node

Note: if you get error message using the curl command, you may not have access to internet on your access node… in this case, try approach 1

```
[root@sandbox ~]# pwd
/root
[root@sandbox ~]# cd /home/lab
[root@sandbox lab]# ll
total 55804
-rw-r--r-- 1 root root 57139942 Jan 16 15:34 full_text.txt
[root@sandbox lab]# curl -O http://www.ark.cs.cmu.edu/GeoText/GeoText.2010-10-12.tgz
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 58.1M  100 58.1M    0     0  2731k      0  0:00:21  0:00:21 --:--:-- 2894k
[root@sandbox lab]# l
-bash: l: command not found
[root@sandbox lab]# ll
total 115352
-rw-r--r-- 1 root root 60973289 Jan 16 21:47 GeoText.2010-10-12.tgz
-rw-r--r-- 1 root root 57139942 Jan 16 15:34 full_text.txt
[root@sandbox lab]# tar -xzvf GeoText.2010-10-12.tgz
GeoText.2010-10-12/
GeoText.2010-10-12/full_text.txt
GeoText.2010-10-12/geo_eval/
GeoText.2010-10-12/preproc/
GeoText.2010-10-12/processed_data/
GeoText.2010-10-12/README.txt
GeoText.2010-10-12/processed_data/data.mat
GeoText.2010-10-12/geo_eval/stats.py
[root@sandbox lab]# ll
total 115356
drwxr-xr-x 5 nagios games     4096 Oct 12  2010 GeoText.2010-10-12
-rw-r--r-- 1 root   root  60973289 Jan 16 21:47 GeoText.2010-10-12.tgz
-rw-r--r-- 1 root   root  57139942 Jan 16 15:34 full_text.txt
[root@sandbox lab]#
```

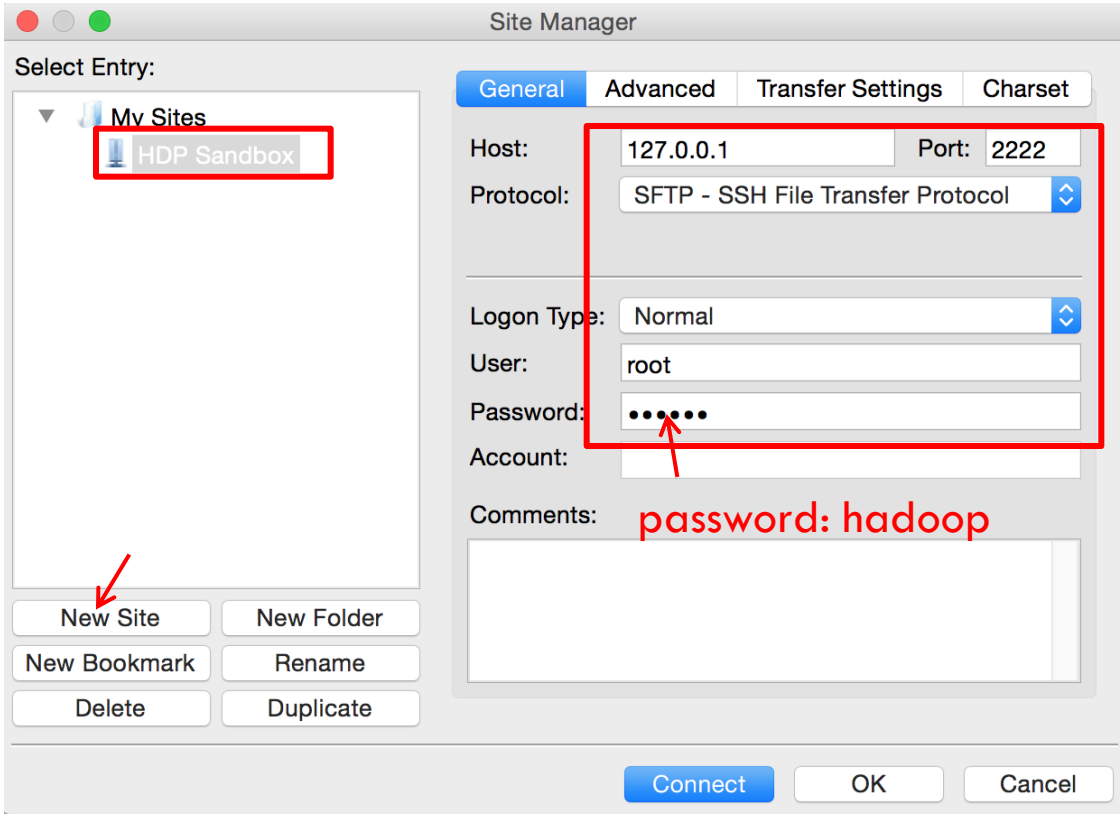
download the file to access node in the Sandbox

decompress the file

Lab 2

Approach 2 – download to local and upload via ftp

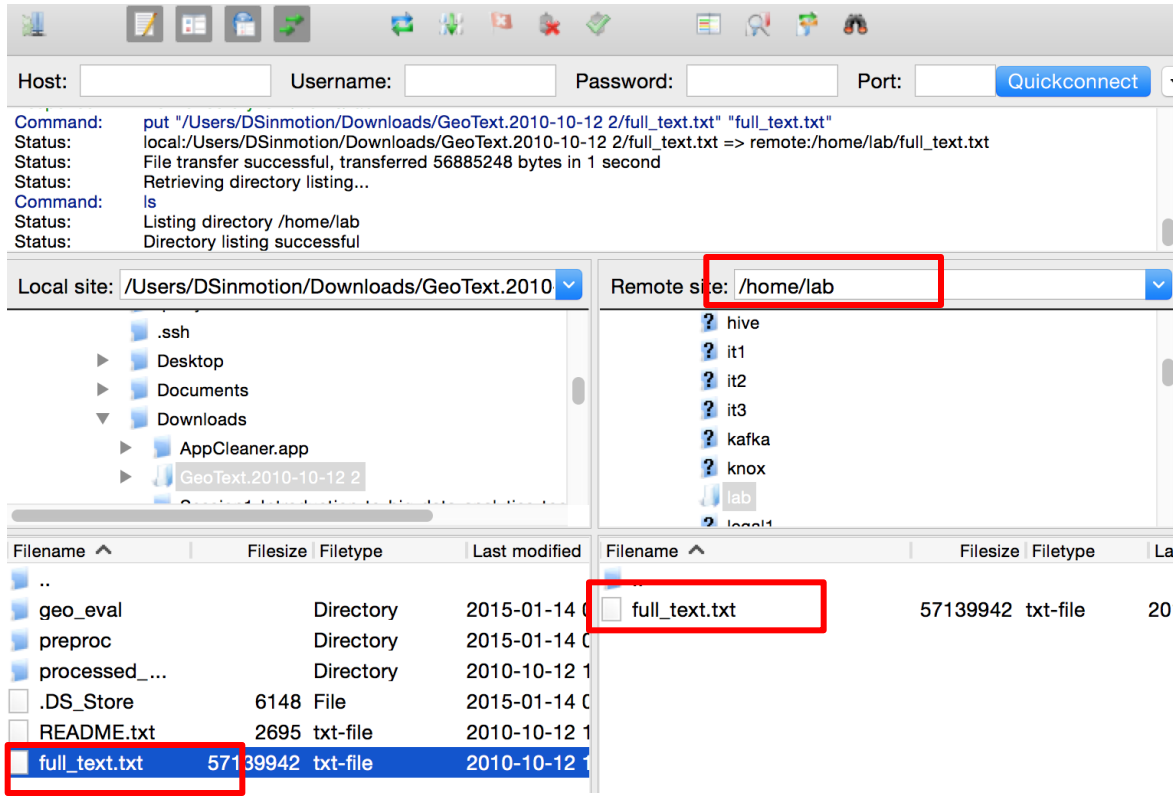
- Download and install FTP tool FileZilla
  - https://filezilla-project.org/download.php?type=client
- Create new FTP site and connect to Sandbox
  - Open FileZilla
  - Go to File → Site Manager
  - Click 'New Site' and name the site
  - In the 'General' tab, enter IP address and SSH port of the HDP Sandbox →



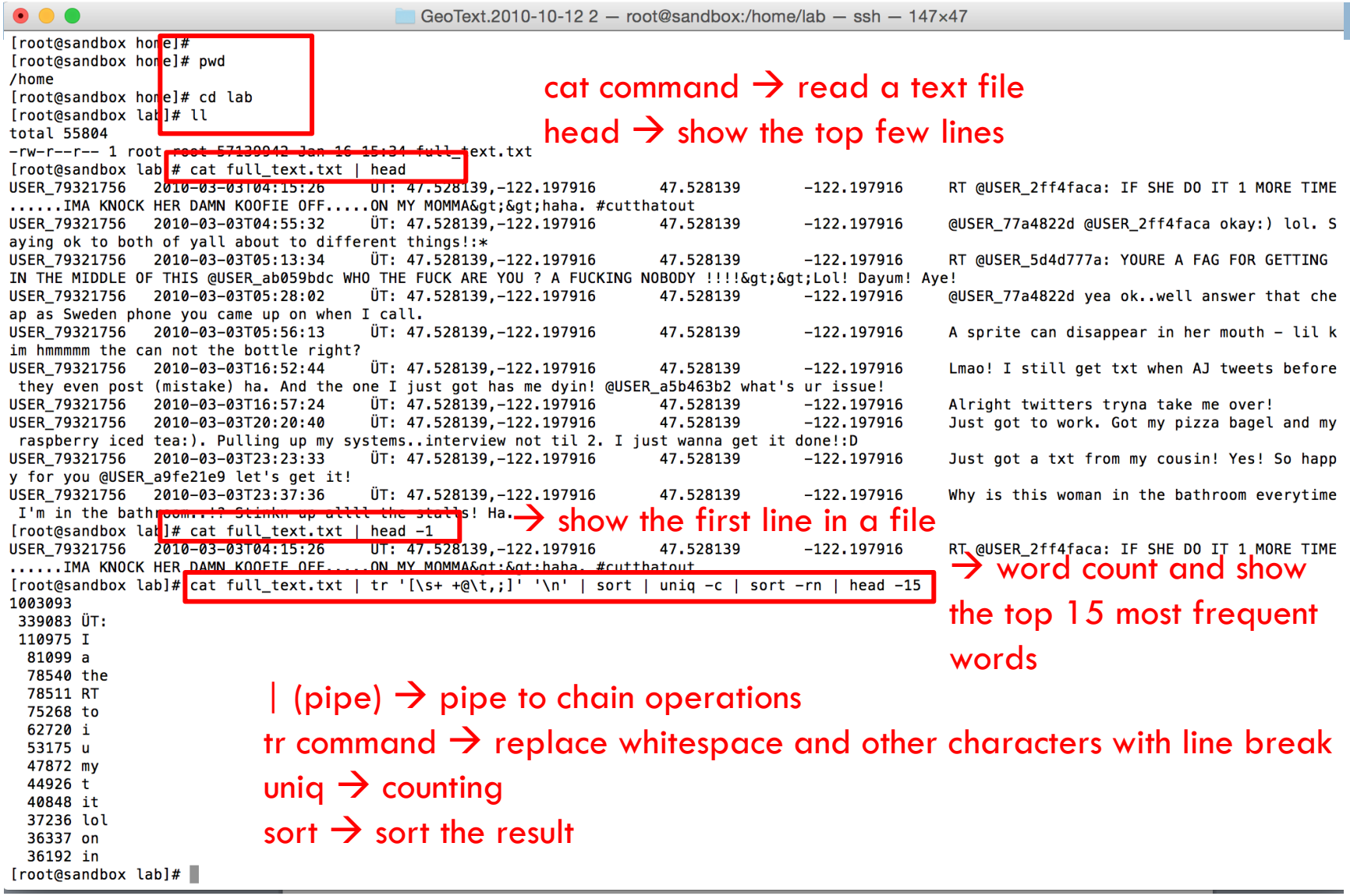
Lab 2

Approach 2 – download to local and upload via ftp

- After connecting to the Sandbox accessnode in Filezilla…
  - The left portion shows directories of your local computer
  - The right portion shows directories of your remote machine
    - In this case the HDP Sandbox
- Upload full_text.txt to Sandbox
  - Find the full_text.txt file you downloaded and drag to a folder on the right to the Sandbox
  - To see the full-text.txt file, you need to unzip the tgz file you downloaded by using the 7-zip tool



Lab 2 – Word Count in Linux



cat command → read a text file

head → show the top few lines

→ show the first line in a file

→ word count and show the top 15 most frequent words

| (pipe) → pipe to chain operations

tr command → replace whitespace and other characters with line break

uniq → counting

sort → sort the result