

# LAB 2 INSTRUCTIONS

CIND 719 BIG DATA ANALYTICS TOOLS  
RYERSON UNIVERSITY

Instructor: Sebnem Sahin Kuzulugil

# Session 2 - Lab & Assignments

## □ Lab Computer

- ▣ Username: dsstudent
- ▣ Password: data

## □ Lab 2

- ▣ Download Geo-tagged Tweets data
- ▣ Word count in Linux
- ▣ Word count in MapReduce

# Lab Environment for Windows Users

Lab computer  
*Windows*

Access Node

*Hadoop cluster*

*Linux (redhat)*

(Azure or VM)  
HDP Sandbox

SSH 127.0.0.1 -p 2222

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

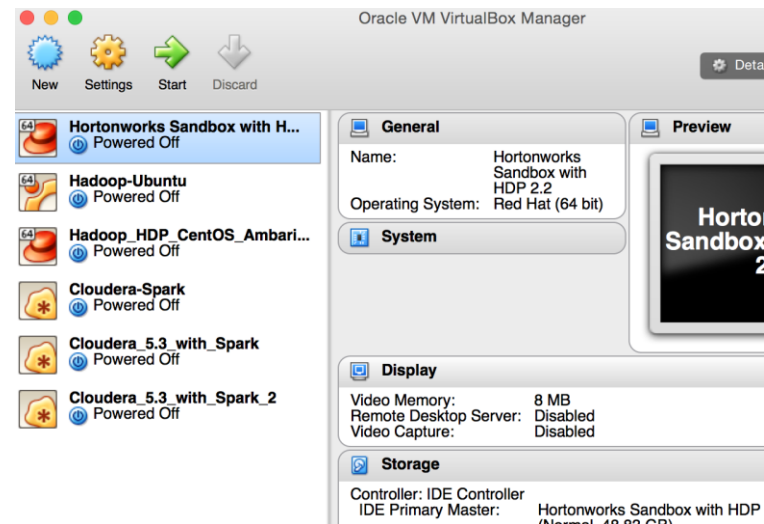




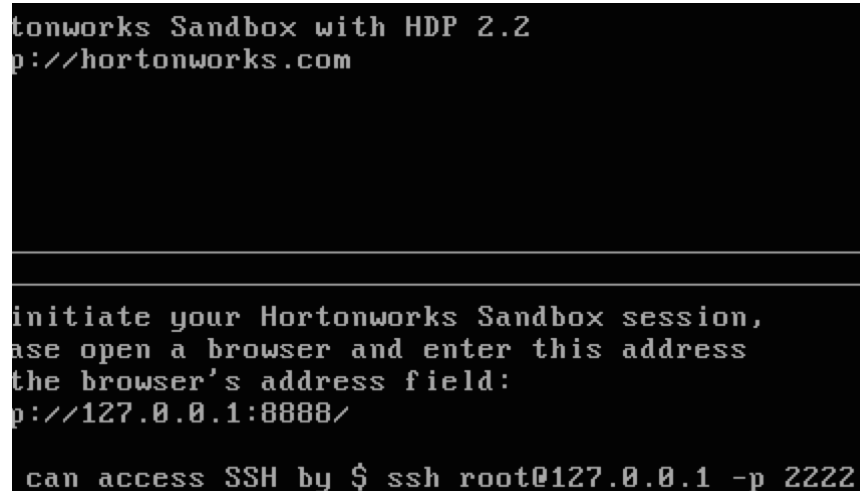
# Lab 2 – Accessing Hadoop

# Lab Environment for Windows Users

1. Log into Windows as dsstudent
2. Start Oracle Virtualbox
3. **Start** the HDP Sandbox (wait few minutes...)



This screen shows  
successful startup



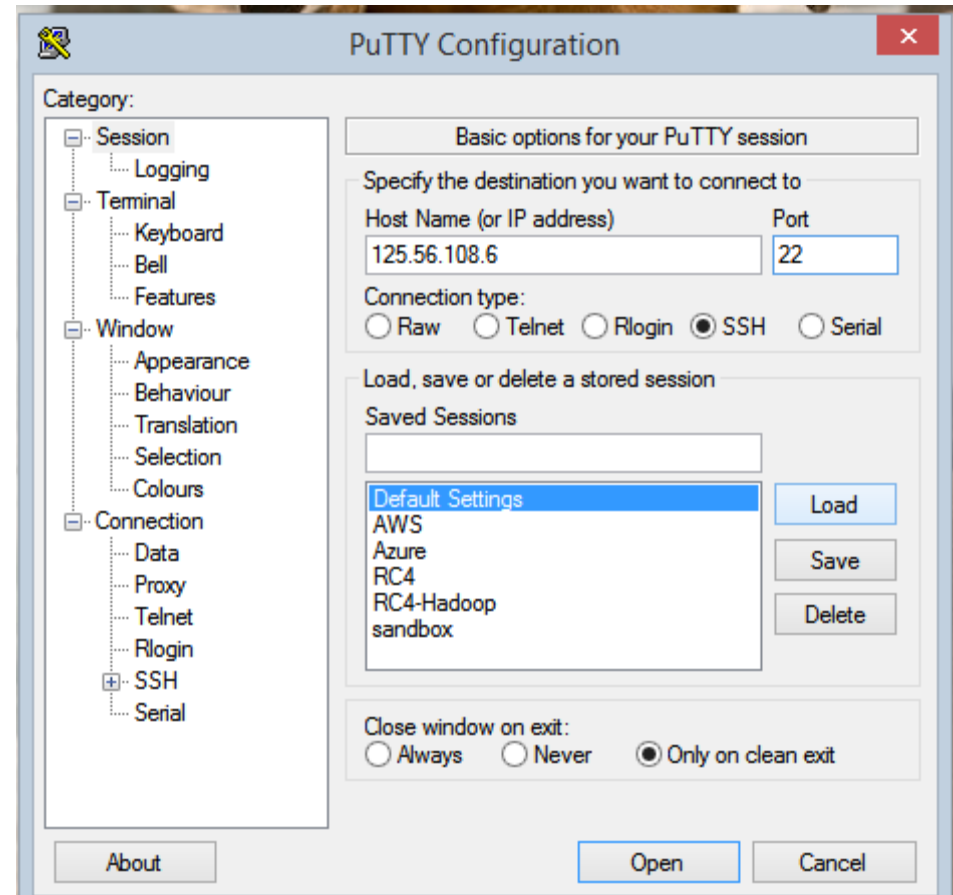
```
tonworks Sandbox with HDP 2.2
p://hortonworks.com

initiate your Hortonworks Sandbox session,
ase open a browser and enter this address
the browser's address field:
p://127.0.0.1:8888/

can access SSH by $ ssh root@127.0.0.1 -p 2222
```

# Lab Environment for Windows Users

- Open PuTTY in Windows and specify hostname, port; then click open to connect to HDP sandbox
  - ▣ Hostname: 127.0.0.1
  - ▣ Port: 2222
  - ▣ Session name: name your session and save so you won't have to enter the same info every time.
- Enter Sandbox username and password
  - ▣ Username: root
  - ▣ Password: hadoop



# Understanding Linux and HDFS



Lab computer  
Your PC/Mac

NOTE: Your client Linux file system and Hadoop file system (HDFS) are separate environments. First, you need to learn how to move files between the two file systems

You can talk to this in Linux!

access node



```
$ hadoop fs -put myFiles.txt /user/lab
```

```
$ hadoop fs -get /user/lab/myFile.txt
```

Access node talks to HDFS  
through “hadoop fs” commands



HDP Hadoop

# Lab 2 – Before You Get Started...

- Create a directory in Linux (HDP sandbox) `‘/home/lab’` → we will put the data in *this* lab folder

```
login as: cind719
cind719@40.86.231.79's password:
Last login: Tue Jun 21 00:51:10 2016 from 141.117.49.52
[cind719@sandbox ~]$ ll → list files in the current directory
total 16
-rw-rw-r-- 1 cind719 cind719 4208 2016-06-20 05:06 pig_1466399167789.log
-rw-rw-r-- 1 cind719 cind719 4331 2016-06-21 00:53 pig_1466470398676.log
[cind719@sandbox ~]$ pwd → show the current directory path
/home/cind719
[cind719@sandbox ~]$ mkdir lab → create a directory called lab
[cind719@sandbox ~]$ cd lab → enter the directory 'lab'
[cind719@sandbox lab]$ pwd
/home/cind719/lab
[cind719@sandbox lab]$ ll
total 0
[cind719@sandbox lab]$
```



# Hadoop FileSystem Shell

Hadoop Filesystem Command	Description
<code>hadoop fs -mkdir</code>	create a new directory in hdfs
<code>hadoop fs -ls</code>	list files in a directory
<code>hadoop fs -put</code>	to copy file from local to hdfs
<code>hadoop fs -cat</code>	to preview the content of an hdfs file
<code>hadoop fs -get</code>	to move file from hdfs to local
<code>hadoop fs -rmdir</code>	to delete a directory
<code>hadoop fs -cp</code>	to make a copy of an hdfs file
<code>hadoop fs -du</code>	to display the size of an hdfs file
<code>hadoop fs -mv</code>	to move hdfs files from source to destination
<code>hadoop fs -tail</code>	to print the last few lines of an hdfs file/directory
<code>hadoop fs -head</code>	to print the first few lines of an hdfs file
<code>hadoop fs -getmerge</code>	to merge several hdfs files into one single file and copy to local

To learn more about Hadoop shell commands, check out the documentations

<http://hadoop.apache.org/docs/r2.6.0/hadoop-project-dist/hadoop-common/FileSystemShell.html>



## Lab 2 – Download and upload dataset

# Upload Dataset to the Sandbox

- Download the dataset: Twitter geo-tagged tweets
  - ▣ Course Page/Content/Resources/Geo-tagged Tweets Dataset/full\_text.txt
- Upload the dataset via FileZilla
  - ▣ If you don't have FileZilla Client, download here: <https://filezilla-project.org>
  - ▣ Install FileZilla

# Upload Dataset to the Sandbox-FileZilla

The screenshot shows the FileZilla interface with the following components and annotations:

- Host:** `sftp://40.86.231.79` (Annotated with `sftp://127.0.0.1` pointing to the local site).
- Username:** `cind719` (Annotated with `root` pointing to the local site).
- Password:** Masked with dots (Annotated with `hadoop` pointing to the local site).
- Local Site (Left):** Shows the local file system. A red arrow points from the text **Your local machine** to the `Program Files (x86)` folder.
- Remote Site (Right):** Shows the remote file system. A red arrow points from the text **Your home on the access node (virtual machine)** to the `home` directory.
- Local Site Table:**

Filename	Filesize	Filetype	Last modified
..			
All Users		File folder	
AppData		File folder	2013-12-06 12:36:1...
Bebeem		File folder	2016-05-17 8:18:49...
Default		File folder	2014-01-09 11:34:4...
Default User		File folder	
Default.migrated		File folder	2014-01-09 11:48:2...
Public		File folder	2014-03-29 8:01:43...
Sebnem		System Folder	2016-06-19 11:50:2...
- Remote Site Table:**

Filename	Filesize	Filetype	Last modified
..			
lab		File folder	
.bash_history		File	
.bash_logout		File	
.bash_profile		File	
.bashrc		File	
.hivehistory		File	
.pig_history		File	
- Drag & Drop:** A red arrow points from the `Program Files (x86)` folder in the local site to the `home` directory in the remote site, with the text **Drag & Drop** in a red box.

# Word Count in Linux

```
[cind719@sandbox lab]$ cat full_text.txt | head
USER_79321756 2010-03-03T04:15:26 UT: 47.528139,-122.197916 47.528139 -122.197916 RT @USER_2ff4faca: IF SHE DO IT 1
MORE TIME.....IMA KNOCK HER DAMN KOOFIE OFF.....ON MY MOMMA&gt;&gt;haha. #cutthatout
USER_79321756 2010-03-03T04:55:32 UT: 47.528139,-122.197916 47.528139 -122.197916 @USER_77a4822d
@USER_2ff4faca okay:) lol. Saying ok to both of yall about to different things!:*
USER_79321756 2010-03-03T05:13:34 UT: 47.528139,-122.197916 47.528139 -122.197916 RT @USER_5d4d777a: YOURE A
FAG FOR GETTING IN THE MIDDLE OF THIS @USER_ab059bdc WHO THE FUCK ARE YOU ? A FUCKING NOBODY !!!!&gt;&gt;Lol! Dayum! Aye!
USER_79321756 2010-03-03T05:28:02 UT: 47.528139,-122.197916 47.528139 -122.197916 @USER_77a4822d yea ok..well
answer that cheap as Sweden phone you came up on when I call.
USER_79321756 2010-03-03T05:56:13 UT: 47.528139,-122.197916 47.528139 -122.197916 A sprite can disappear in her mouth -
lil kim hmmmmm the can not the bottle right?
USER_79321756 2010-03-03T16:52:44 UT: 47.528139,-122.197916 47.528139 -122.197916 Lmao! I still get txt when AJ tweets
before they even post (mistake) ha. And the one I just got has me dyin! @USER_a5b463b2 what's ur issue!
USER_79321756 2010-03-03T16:57:24 UT: 47.528139,-122.197916 47.528139 -122.197916 Alright twitters tryna take me over!
USER_79321756 2010-03-03T20:20:40 UT: 47.528139,-122.197916 47.528139 -122.197916 Just got to work. Got my pizza bagel
and my raspberry iced tea:). Pulling up my systems.interview not til 2. I just wanna get it done!D
USER_79321756 2010-03-03T23:23:33 UT: 47.528139,-122.197916 47.528139 -122.197916 Just got a txt from my cousin! Yes! So
happy for you @USER_a9fe21e9 let's get it!
USER_79321756 2010-03-03T23:37:36 UT: 47.528139,-122.197916 47.528139 -122.197916 Why is this woman in the bathroom
everytime I'm in the bathroom..!? Stinkn up allll the stalls! Ha.
[cind719@sandbox lab]$ cat full_text.txt | head -1
USER_79321756 2010-03-03T04:15:26 UT: 47.528139,-122.197916 47.528139 -122.197916 RT @USER_2ff4faca: IF SHE DO IT 1
MORE TIME.....IMA KNOCK HER DAMN KOOFIE OFF.....ON MY MOMMA&gt;&gt;haha. #cutthatout
[cind719@sandbox lab]$ cat full_text.txt | tr '[:space:]' '\n' | sort | uniq -c | sort -rn | head -15
339083 UT:
108259 I
77958 RT
75393 the
75005 to
62177 a
46331 my
```

**cat command → read a text file, head → show the top few lines**

**cat full\_text.txt | head -1 → show the first line in a file**

**cat full\_text.txt | tr '[:space:]' '\n' | sort | uniq -c | sort -rn | head -15 → word count and show the top 15 most frequent words in Linux**

**| (pipe) → pipe to chain operations**  
**tr command → replace whitespace and other characters with line break**  
**uniq → unique instances (c: count)**  
**sort → sort the result (r: reverse, n: numeric)**

# Loading Files into HDFS

```
[cind719@sandbox lab]$ hadoop fs -ls /user/
```

Found 11 items

drwxrwx---	- ambari-qa	hdfs	0	2016-02-29 17:57	/user/ambari-qa
drwxr-xr-x	- cind719	hdfs	0	2016-06-20 05:05	/user/cind719
drwxr-xr-x	- hcat	hdfs	0	2016-02-29 18:02	/user/hcat
drwxr-xr-x	- hive	hdfs	0	2016-02-29 18:02	/user/hive
drwxr-xr-x	- maria_dev	hdfs	0	2016-06-20 04:52	/user/maria_dev
drwxrwxr-x	- oozie	hdfs	0	2016-02-29 18:03	/user/oozie
drwxr-xr-x	- solr	hdfs	0	2016-02-29 18:11	/user/solr
drwxrwxr-x	- spark	hdfs	0	2016-02-29 17:59	/user/spark
drwxr-xr-x	- unit	hdfs	0	2016-02-29 18:05	/user/unit
drwxr-xr-x	- yarn	hdfs	0	2016-02-29 18:05	/user/yarn
drwxr-xr-x	- zeppelin	hdfs	0	2016-06-20 05:10	/user/zeppelin

← You don't see a /user/lab folder in HDFS yet

```
[cind719@sandbox lab]$ hadoop fs -mkdir /user/lab
```

```
[cind719@sandbox lab]$ hadoop fs -ls /user/
```

← Create the lab folder in HDFS

Found 12 items

drwxrwx---	- ambari-qa	hdfs	0	2016-02-29 17:57	/user/ambari-qa
drwxr-xr-x	- cind719	hdfs	0	2016-06-20 05:05	/user/cind719
drwxr-xr-x	- hcat	hdfs	0	2016-02-29 18:02	/user/hcat
drwxr-xr-x	- hive	hdfs	0	2016-02-29 18:02	/user/hive
drwxr-xr-x	- cind719	hdfs	0	2016-06-22 04:46	<b>/user/lab</b>
drwxr-xr-x	- maria_dev	hdfs	0	2016-06-20 04:52	/user/maria_dev
drwxrwxr-x	- oozie	hdfs	0	2016-02-29 18:03	/user/oozie
drwxr-xr-x	- solr	hdfs	0	2016-02-29 18:11	/user/solr
drwxrwxr-x	- spark	hdfs	0	2016-02-29 17:59	/user/spark
drwxr-xr-x	- unit	hdfs	0	2016-02-29 18:05	/user/unit
drwxr-xr-x	- yarn	hdfs	0	2016-02-29 18:05	/user/yarn
drwxr-xr-x	- zeppelin	hdfs	0	2016-06-20 05:10	/user/zeppelin

← Now you see it!!

```
[cind719@sandbox lab]$ hadoop fs -put /home/cind719/lab/full_text.txt /user/lab
```

```
[cind719@sandbox lab]$ hadoop fs -ls /user/lab
```

← Upload the full\_text.txt file to HDFS

Found 1 items

-rw-r--r--	3	cind719	hdfs	57139942	2016-06-22 04:48	/user/lab/full_text.txt
------------	---	---------	------	----------	------------------	-------------------------

```
[cind719@sandbox lab]$
```

# Loading Files into HDFS

```
[cind719@sandbox lab]$ hadoop fs -cat /user/lab/full_text.txt | head -n 3 ← Display the first few lines of HDFS file
USER_79321756 2010-03-03T04:15:26 ÜT: 47.528139,-122.197916 47.528139 -122.197916 RT @USER_2ff4faca: IF SHE DO IT 1
MORE TIME.....IMA KNOCK HER DAMN KOOFIE OFF....ON MY MOMMA&gt;&gt;haha. #cutthatout
USER_79321756 2010-03-03T04:55:32 ÜT: 47.528139,-122.197916 47.528139 -122.197916 @USER_77a4822d @USER_2ff4faca
okay:) lol. Saying ok to both of yall about to different things!:*
USER_79321756 2010-03-03T05:13:34 ÜT: 47.528139,-122.197916 47.528139 -122.197916 RT @USER_5d4d777a: YOURE A FAG
FOR GETTING IN THE MIDDLE OF THIS @USER_ab059bdc WHO THE FUCK ARE YOU ? A FUCKING NOBODY !!!!&gt;&gt;Lol! Dayum! Aye!
cat: Unable to write to output stream.
[cind719@sandbox lab]$ hadoop fs -cp /user/lab/full_text.txt /user/lab/full_text2.txt ← Make a copy of the HDFS file
[cind719@sandbox lab]$ hadoop fs -ls /user/lab
Found 2 items
-rw-r--r-- 3 cind719 hdfs 57139942 2016-06-22 04:48 /user/lab/full_text.txt
-rw-r--r-- 3 cind719 hdfs 57139942 2016-06-22 05:03 /user/lab/full_text2.txt
[cind719@sandbox lab]$ hadoop fs -rm /user/lab/full_text2.txt ← Delete a file in HDFS
16/06/22 05:03:58 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 360 minutes, Emptier interval = 0
minutes.
Moved: 'hdfs://sandbox.hortonworks.com:8020/user/lab/full_text2.txt' to trash at:
hdfs://sandbox.hortonworks.com:8020/user/cind719/.Trash/Current
[cind719@sandbox lab]$ hadoop fs -ls /user/lab
Found 1 items
-rw-r--r-- 3 cind719 hdfs 57139942 2016-06-22 04:48 /user/lab/full_text.txt
[cind719@sandbox lab]$ hadoop fs -get /user/lab/full_text.txt full_text3.txt ← Download a file from HDFS to access
[cind719@sandbox lab]$ ll
total 111612
-rw-r--r-- 1 cind719 cind719 57139942 2016-06-22 05:04 full_text3.txt
-rw-rw-r-- 1 cind719 cind719 57139942 2016-06-22 04:44 full_text.txt
[cind719@sandbox lab]$ rm full_text3.txt ← Delete a local file
[cind719@sandbox lab]$ ll
total 55808
-rw-rw-r-- 1 cind719 cind719 57139942 2016-06-22 04:44 full_text.txt
[cind719@sandbox lab]$
```

# Java MapReduce WordCount

## 1. Download the shakespeare.txt file

- ▣ Shakespeare dataset has been uploaded to the Course Page under “Contents/Resources/Other Datasets” section

## 2. Upload the file to Sandbox via FileZilla

## 3. Put the shakespeare.txt file into HDFS (/user/lab)

## 4. Find your mapreduce-example jar file

- ▣ `[root@sandbox ~]# find /usr -name *hadoop-mapreduce-example*`

## 5. Run java M/R wordcount example

- ▣ `[root@sandbox lab]# hadoop jar /usr/hdp/2.4.0.0-2041/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount /user/lab/shakespeare.txt /user/lab/shakespeare-wc-out`

Make sure this \*.jar file path matches with the result from previous step

## ▣ View results

- ▣ `[root@sandbox lab]# hadoop fs -ls /user/lab/shakespeare-wc-out`
- ▣ `[root@sandbox lab]# hadoop fs -cat /user/lab/shakespeare-wc-out/part-r-00000 | tail -n 50`



# Python Streaming WordCount

1. Upload the shakespeare.txt file to sandbox via FileZilla
2. Put the shakespeare.txt file into HDFS
3. Upload the `wc_mapper.py` and `wc_reducer.py` scripts to sandbox /home/lab folder *(scripts are available on course page at Labs/Session-2 for download)*

4. Find your mapreduce-example jar file

- ▣ `[root@sandbox lab]# find /usr -name *hadoop-streaming*`

Make sure this \*.jar file path  
↙ matches with the result from  
previous step

5. Run Hadoop streaming wordcount

- ▣ `[root@sandbox lab]# hadoop jar /usr/hdp/2.4.0.0-2041/hadoop-mapreduce/hadoop-streaming.jar -file /home/lab/wc_mapper.py -mapper /home/lab/wc_mapper.py -file /home/lab/wc_reducer.py -reducer /home/lab/wc_reducer.py -input /user/lab/shakespeare.txt -output /user/lab/shakespeare-wc-out-py`

6. View results

- ▣ `[root@sandbox lab]# hadoop fs -ls /user/lab/shakespeare-wc-out-py`
- ▣ `[root@sandbox lab]# hadoop fs -cat /user/lab/shakespeare-wc-out-py/part-00000 | tail -n 50`

# Summary

- The two python scripts perform the mapper and reducer tasks. The underlying data shuffling/sorting have been taken care of by the MapReduce framework
- If you don't like to program Java/Python, the “most” boring part ends here
- Now that you've practiced Hadoop file system commands and tried to run some MapReduce code, you're ready to move on to the fun part – Pig and Hive!