

Hola Hadoop

Peligro!

... please

Peligro!

... please be careful of what you are doing!

- Think twice before:

rm

mv

cp

kill

emacs/vim/... configuration files

Peligro!

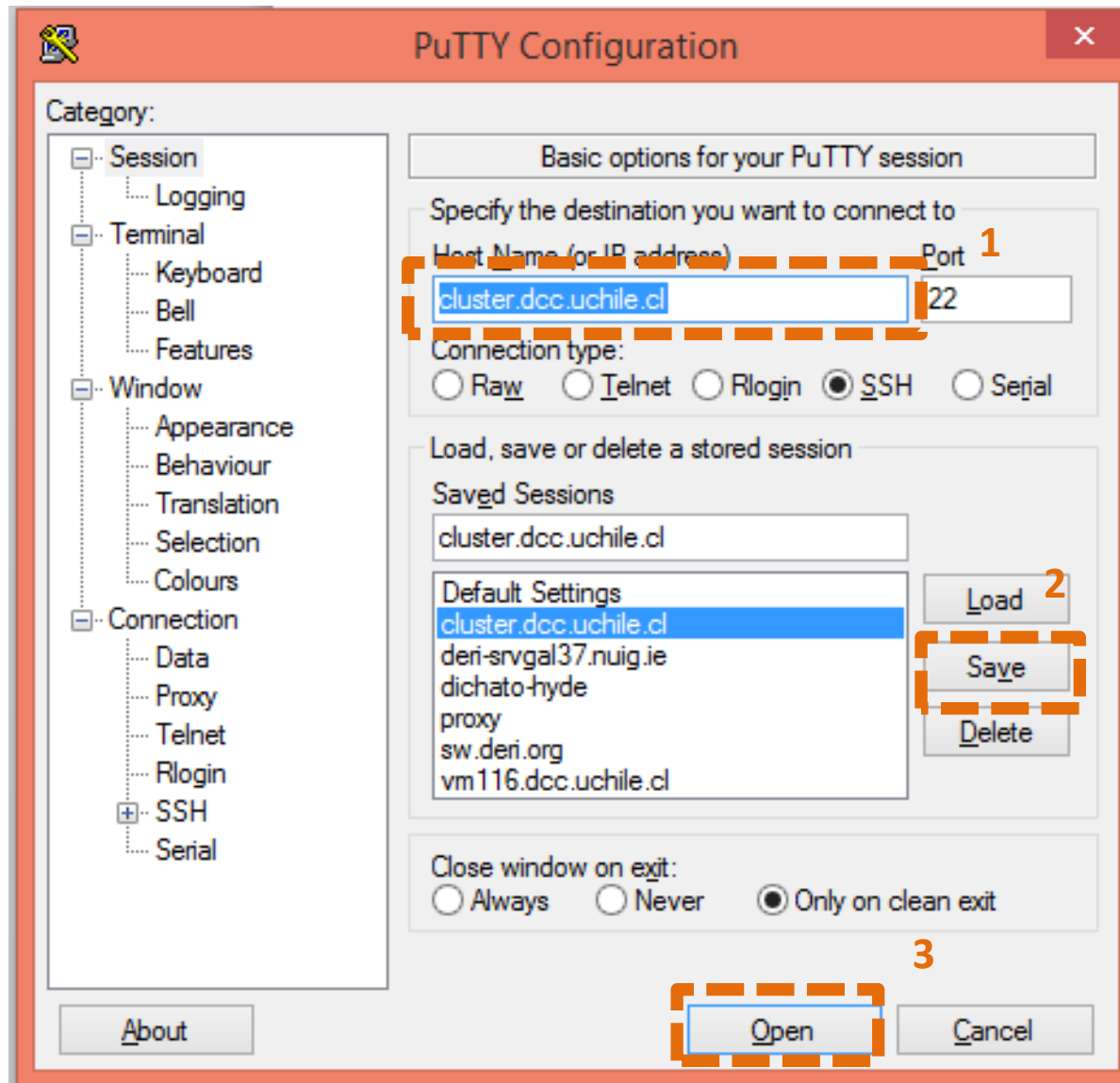
... please.

- `cluster.dcc.uchile.cl`

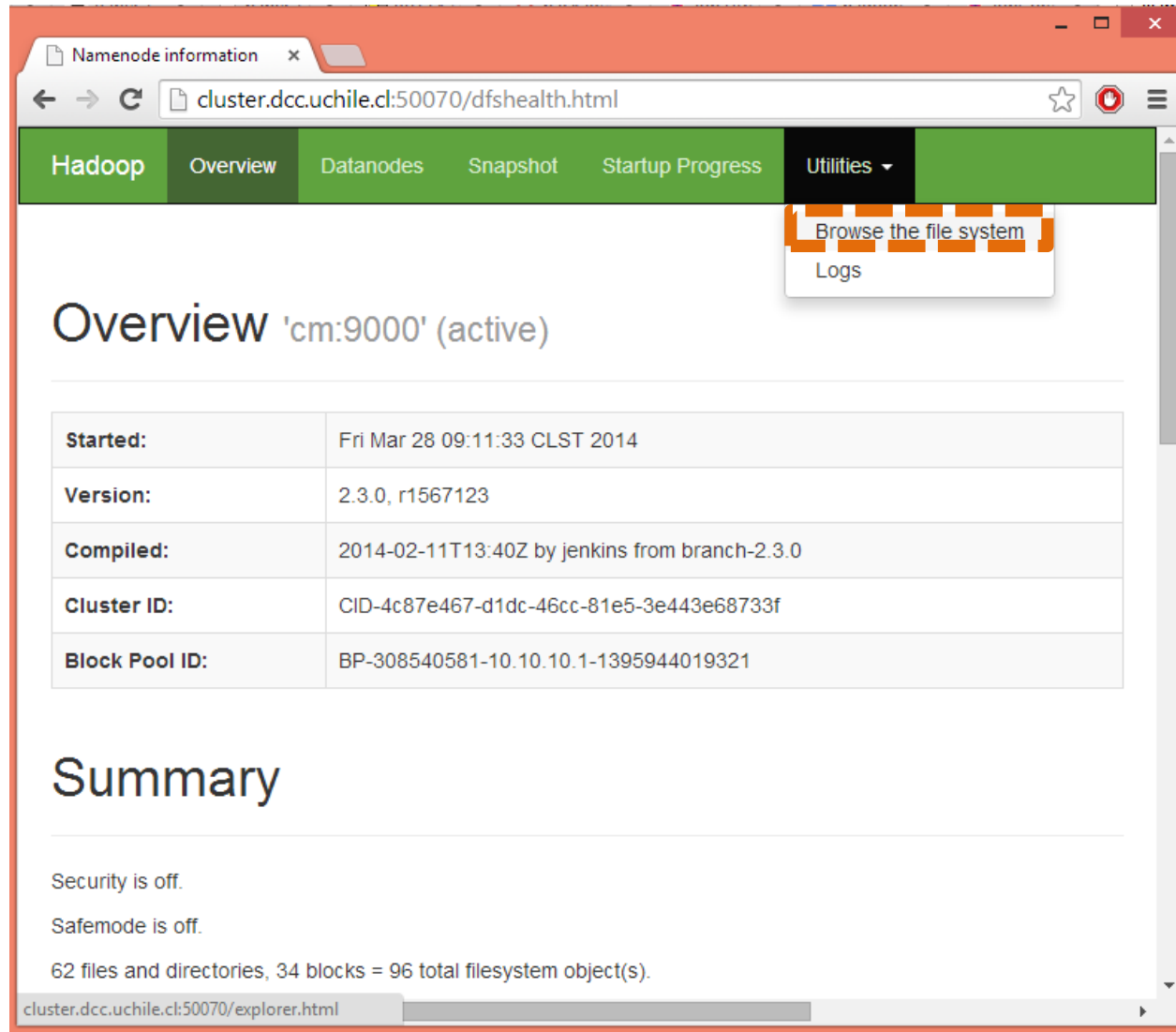
1. Download tools

- <http://aidanhogan.com/teaching/cc5212-1-2015/tools/>
- Unzip them somewhere you can find them

2. Log-in PuTTY



3. Open DFS Browser



The screenshot shows a web browser window with the URL `cluster.dcc.uchile.cl:50070/dfshealth.html`. The page title is "Namenode information". The navigation menu includes "Hadoop", "Overview", "Datanodes", "Snapshot", "Startup Progress", and "Utilities". The "Utilities" menu is open, showing "Browse the file system" and "Logs". The main content area displays the "Overview 'cm:9000' (active)" page. Below the title is a table with the following information:

Started:	Fri Mar 28 09:11:33 CLST 2014
Version:	2.3.0, r1567123
Compiled:	2014-02-11T13:40Z by jenkins from branch-2.3.0
Cluster ID:	CID-4c87e467-d1dc-46cc-81e5-3e443e68733f
Block Pool ID:	BP-308540581-10.10.10.1-1395944019321

Below the table is a "Summary" section with the following text:

Security is off.
Safemode is off.
62 files and directories, 34 blocks = 96 total filesystem object(s).

The browser's address bar shows `cluster.dcc.uchile.cl:50070/explorer.html`.

<http://cluster.dcc.uchile.cl:50070/>

4. PuTTY: See state of DFS

- `hdfs dfsadmin -report`

5. PuTTY: Create folder

- `hdfs dfs -ls /`
- `hdfs dfs -ls /uhadoop`
- `hdfs dfs -mkdir /uhadoop/[username]`
 - `[username]` = first letter first name, last name (e.g., “ahogan”)

6. PuTTY: Upload Data

- `cd /data/2014/uhadoop/shared/`
- Then
 - `hdfs dfs -copyFromLocal /data/2014/uhadoop/shared/es-wiki-abstracts.txt /uhadoop/[username]/`
 - OR
 - `hdfs dfs -copyFromLocal /data/2014/uhadoop/shared/es-wiki-abstracts.txt.gz /uhadoop/[username]/`

Note on namespace

- If you need to disambiguate local/remote files
- HDFS file
 - `hdfs://cm:9000/uhadoop/...`
- Local file
 - `file:///data/hadoop/...`

7. Let's Build Our First MapReduce Job

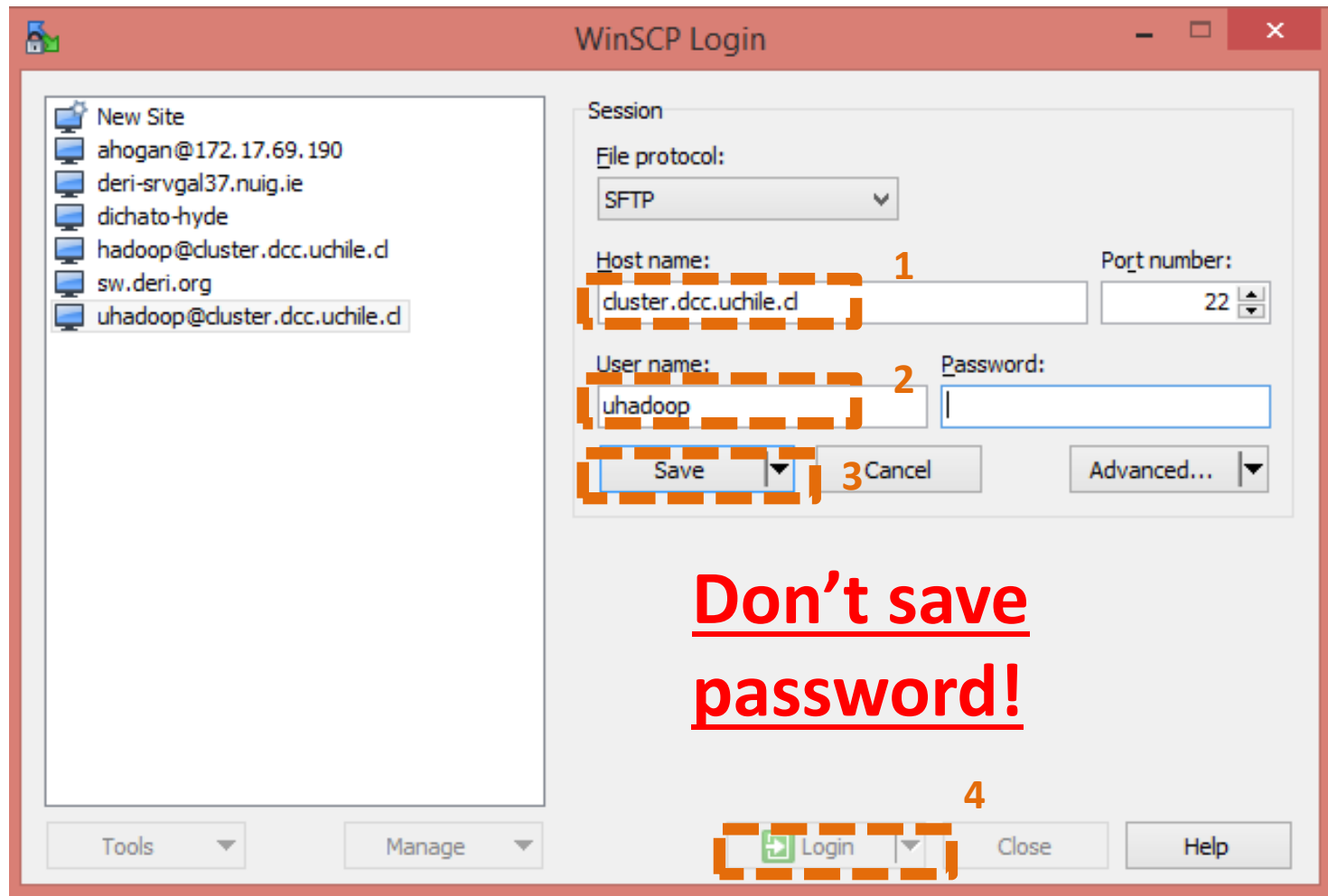
- Hint: Use Monday's slides for "inspiration"
 - <http://aidanhogan.com/teaching/cc5212-1-2016/>
 - Also copied in `CitationCount.java`
1. Implement `map(.,.,.,.)` method
 2. Implement `reduce(.,.,.,.)` method
 3. Implement `main(.)` method

8. Eclipse: Build jar

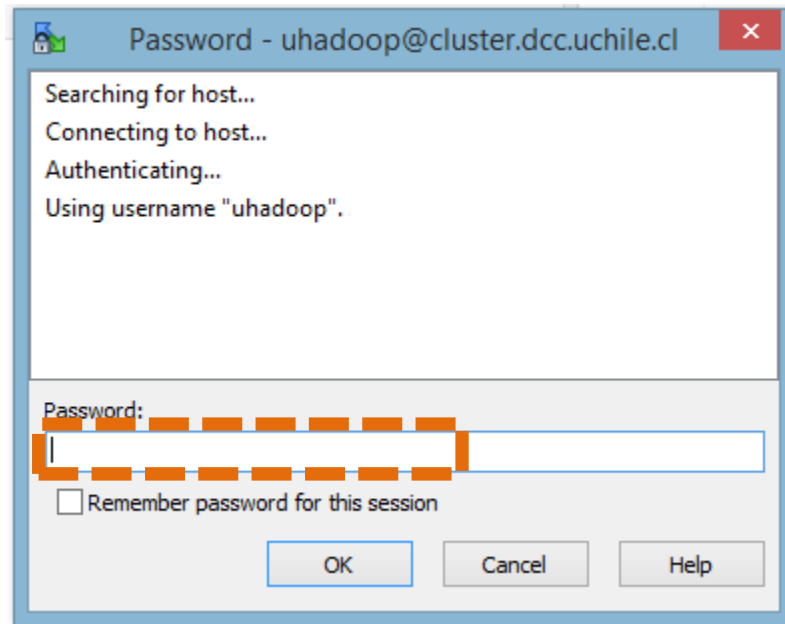
Right Click build.xml > dist

(Might need to make a dist folder)

9. WinSCP: Copy .jar to Master Server

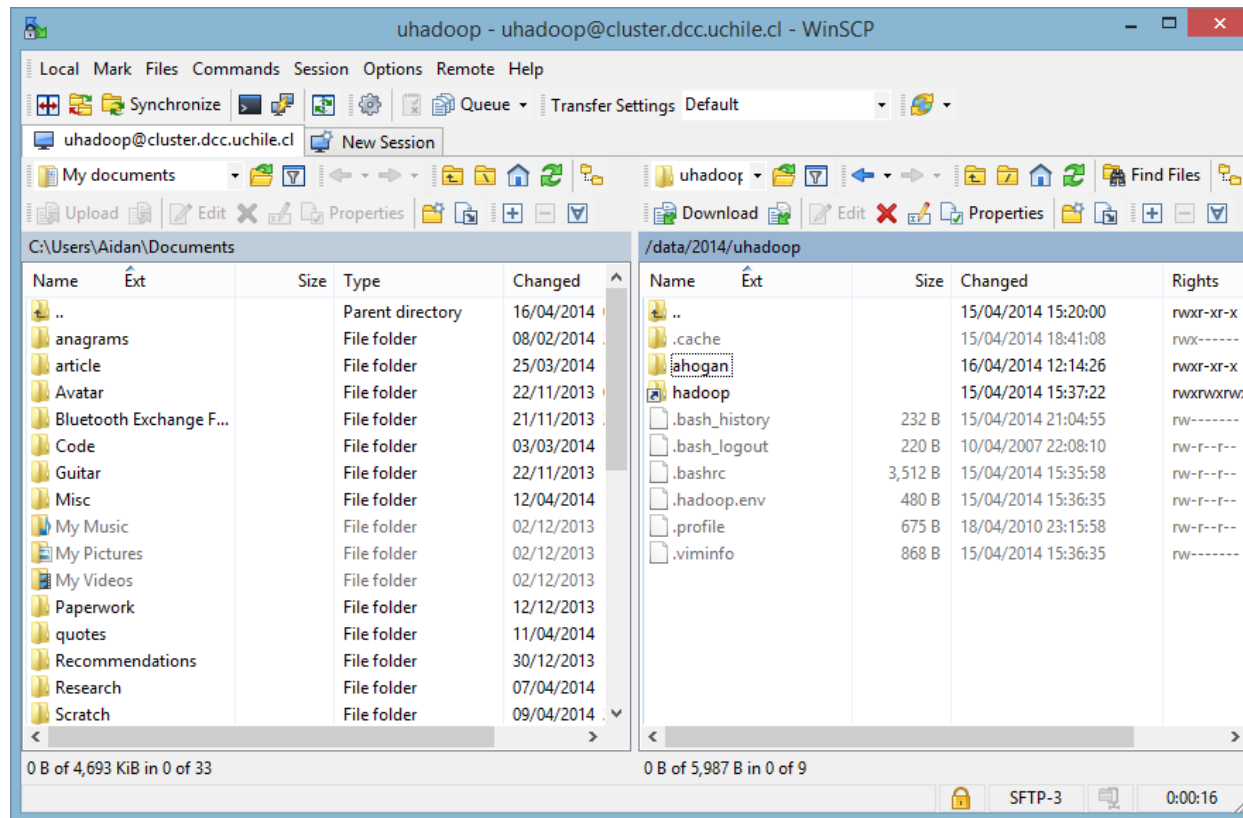


9. WinSCP: Copy .jar to Master Server



9. WinSCP: Copy .jar to Master Server

- Create dir: /data/2014/uhadoop/[username]/
- Copy your mdp-lab4.jar into it



10. PuTTY: Run Job

- `hadoop jar` **All one command!**
`/data/2014/uhadoop/[username]/mdp-`
`lab4.jar WordCount /uhadoop/[username]/es-`
`wiki-abstracts.txt.gz`
`/uhadoop/[username]/wc/`

11. PuTTY: Look at output

- `hdfs dfs -ls /uhadoop/[username]/wc/`
- `hdfs dfs -cat /uhadoop/[username]/wc/part-r-00000 | more`
All one command!
- `hdfs dfs -cat /uhadoop/[username]/wc/part-r-00000 | grep -P "^de\t" | more`

Look for "de" ... 4916432
occurrences in local run

12. Look at output through browser

Namenode information

cluster.dcc.uchile.cl:50070/dfshealth.html

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

Browse the file system
Logs

Overview 'cm:9000' (active)

Started:	Fri Mar 28 09:11:33 CLST 2014
Version:	2.3.0, r1567123
Compiled:	2014-02-11T13:40Z by jenkins from branch-2.3.0
Cluster ID:	CID-4c87e467-d1dc-46cc-81e5-3e443e68733f
Block Pool ID:	BP-308540581-10.10.10.1-1395944019321

Summary

Security is off.
Safemode is off.
62 files and directories, 34 blocks = 96 total filesystem object(s).

cluster.dcc.uchile.cl:50070/explorer.html

<http://cluster.dcc.uchile.cl:50070/>

HELLO-WORLD