CS561: Advanced Topics In Database Systems
Spring-2012

Instructions to Access Hadoop

Hadoop Account and Setup

The first step in this project is to make sure you can access Hadoop and your account is working. In order to do so, follow the following steps:

1- Hadoop Cluster

Hadoop cluster is installed over 4 machines:

cs-master.wpi.edu: The master node running the NameNode and JobTracker
compute-0-1, compute-0-2, compute-0-3: The data nodes

2- Connect to Hadoop

- You need to ssh to the cs-master.wpi.edu using your WPI Linux account

  ```
  ssh cs-master.wpi.edu  -l <wpi-userID>   (then use your wpi password)
  ```

- Check environment variables are set correctly
  JAVA_HOME=/usr/java/latest
  HADOOP_HOME=/usr/share/hadoop
  PATH contains /usr/java/jdk1.6.0_16/bin/ (if not add this path to your PATH environment variable

- Make sure you can compile and run any java program, e.g., "Hello World", (in this step you test your java and make sure it is working fine)

3- Testing Hadoop connection by uploading small file

- In Hadoop, every user X has a directory in HDFS "/user/X" under which you can create your data
- In cs-master.wpi.edu, you can run some Hadoop commands such as

  ```
  > hadoop fs –put <local file to read from>   <destination in HDFS>
  ```

  For example, to upload file abc.txt (in current local directory) to Hadoop, you can execute command (replace X with your username):

  ```
  >hadoop fs –put ./abc.txt  /user/X/
  ```

- You can check your file using command (replace X with your username):

  ```
  >hadoop fs –ls /user/X
  ```

- For more commands about Hadoop file system check:
  http://hadoop.apache.org/common/docs/r0.18.3/hdfs_shell.html

4- Accessing Hadoop using web browser

Hadoop cluster has a web interface though which users can navigate the file system (HDFS) and see the jobs they submit. Our cluster has a single access point, which is cs-master.wpi.edu. The other datanodes are in a private network, therefore, you need to follow the following steps to access all nodes.

- Install Firefox web browser

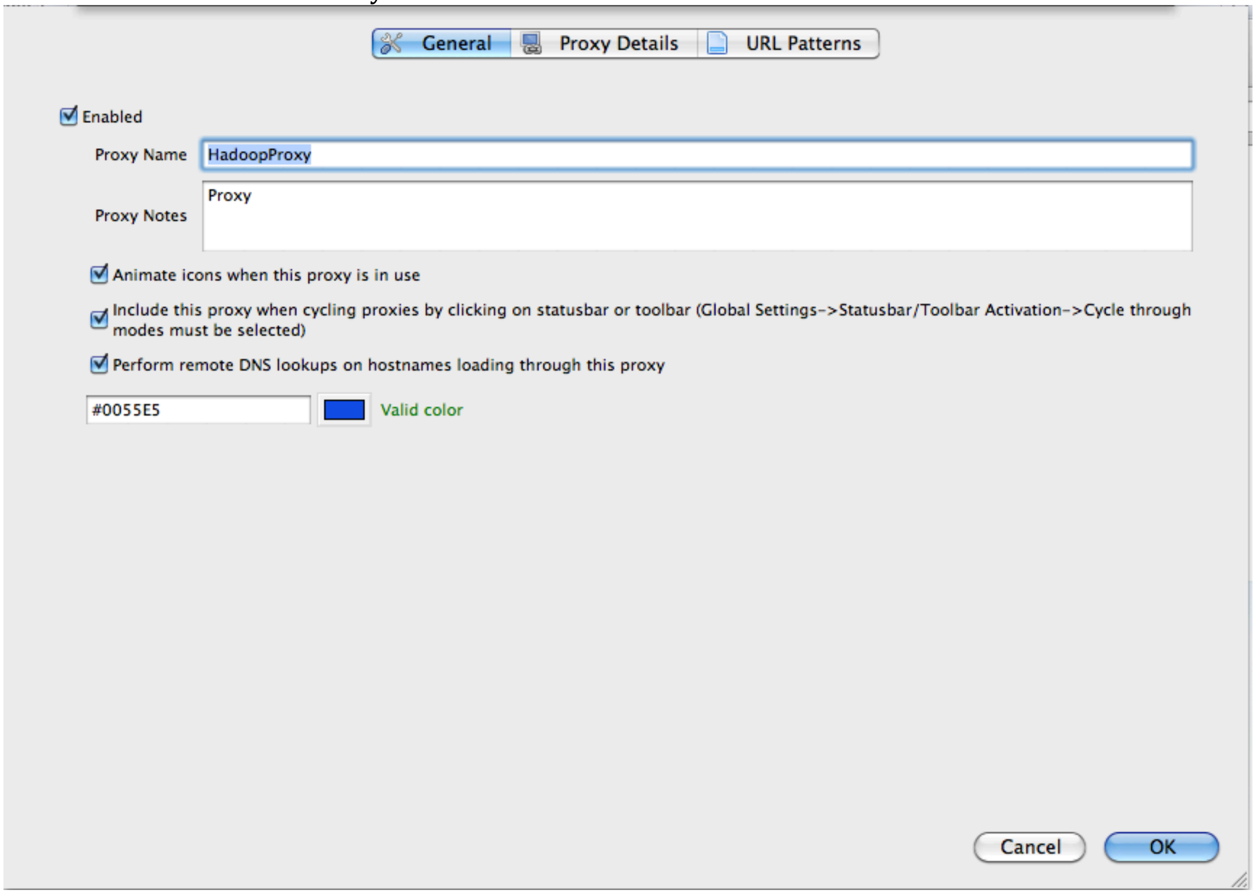
- The HDFS website is: http://cs-master.wpi.edu:50770/dfshealth.jsp

- The Jobs website is: http://cs-master.wpi.edu:50660/jobtracker.jsp

- At this point, you access the master node only. If you click on any of the data nodes to see their content, it will fail (because they are not directly accessible). You set the proxy as follows:

- Install FoxyProxy software in Firefox (it enables routing the HTTP requests) and follow these steps:

1- “Add New Proxy”



☑ Enabled

Proxy Name: HadoopProxy

Proxy Notes: Proxy

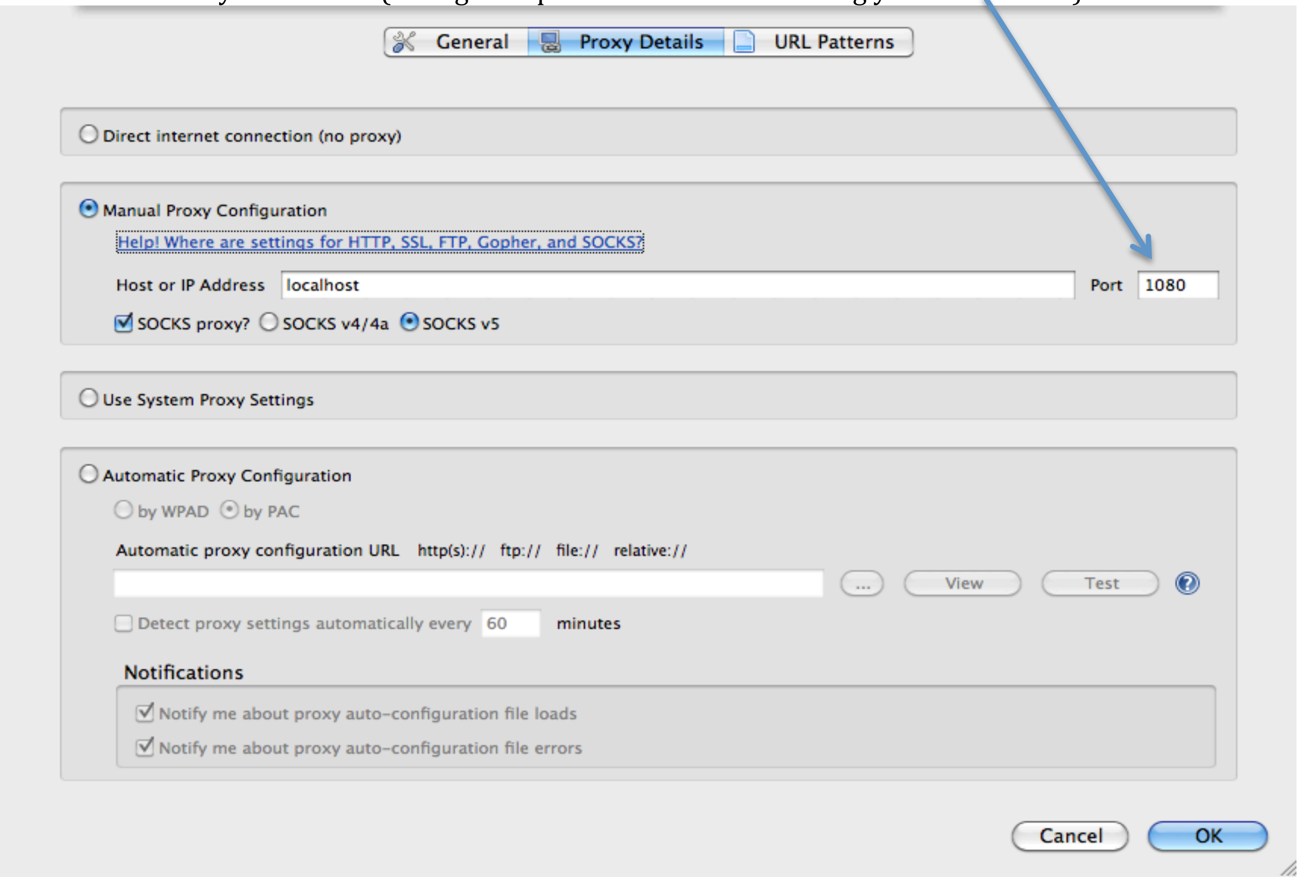
☑ Animate icons when this proxy is in use

☑ Include this proxy when cycling proxies by clicking on statusbar or toolbar (Global Settings->Statusbar/Toolbar Activation->Cycle through modes must be selected)

☑ Perform remote DNS lookups on hostnames loading through this proxy

#0055E5 🟦 Valid color

2- In Proxy Details tab (Change the port number to something you select XXXX)



○ Direct internet connection (no proxy)

● Manual Proxy Configuration

Help! Where are settings for HTTP, SSL, FTP, Gopher, and SOCKS?

Host or IP Address: localhost   Port: 1080

☑ SOCKS proxy? ○ SOCKS v4/4a ● SOCKS v5

○ Use System Proxy Settings

○ Automatic Proxy Configuration

○ by WPAD ● by PAC

Automatic proxy configuration URL   http(s)://   ftp://   file://   relative://

[…] View Test

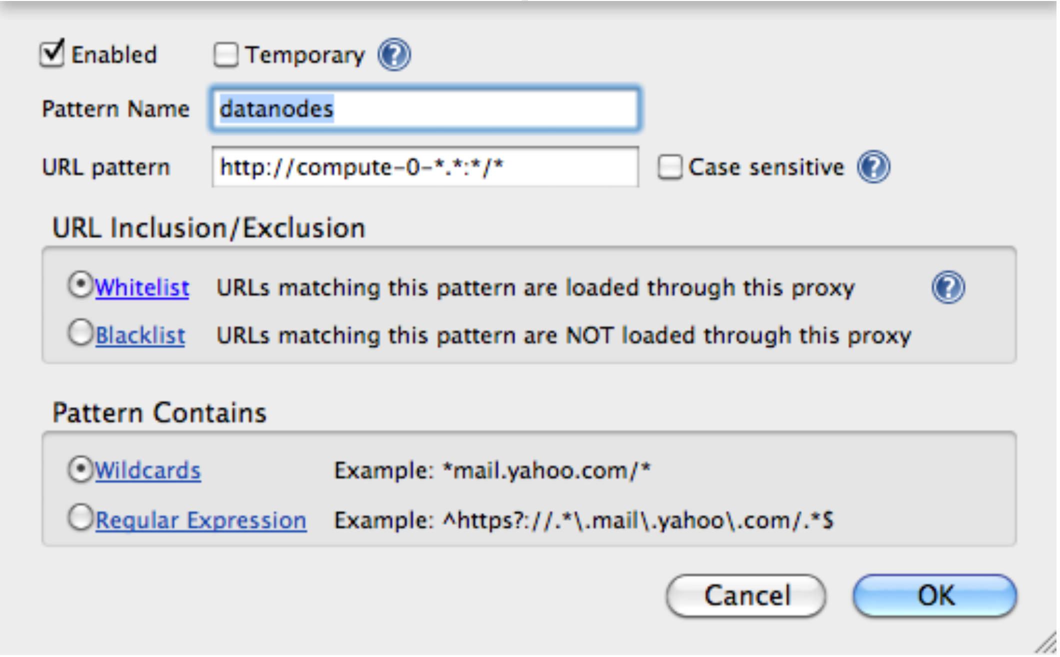
□ Detect proxy settings automatically every 60 minutes

Notifications

☑ Notify me about proxy auto-configuration file loads

☑ Notify me about proxy auto-configuration file errors

3- In URL Patterns → Add new pattern



☑ Enabled   □ Temporary

Pattern Name: datanodes

URL pattern: http://compute-0-*.*:*/*   □ Case sensitive

URL Inclusion/Exclusion

● Whitelist   URLs matching this pattern are loaded through this proxy

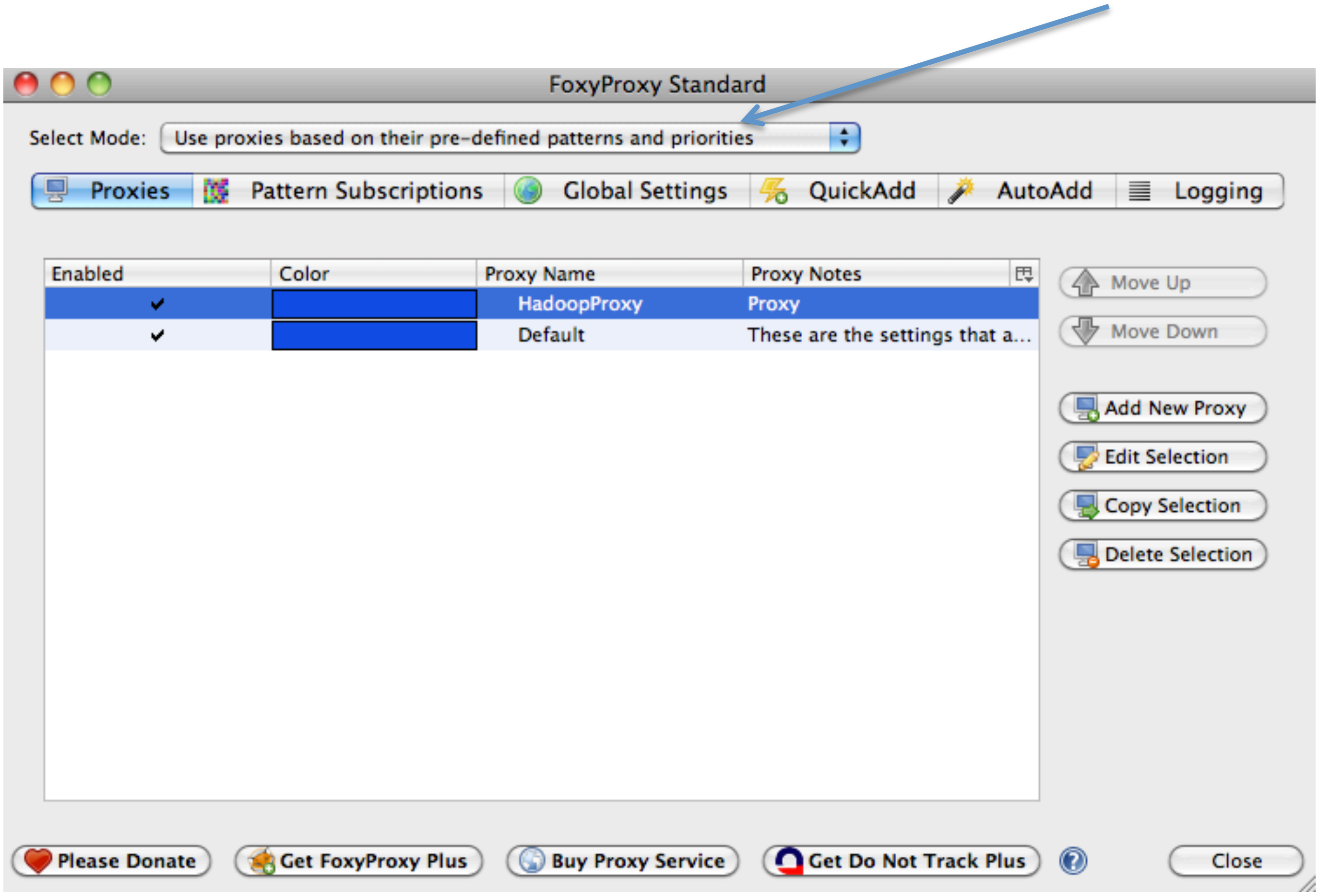
○ Blacklist   URLs matching this pattern are NOT loaded through this proxy

Pattern Contains

● Wildcards   Example: *mail.yahoo.com/*

○ Regular Expression   Example: ^https?://.*\.mail\.yahoo\.com/.*$

4- In the main page (where you selected Add New Pattern”) make sure you select this mode



Select Mode: Use proxies based on their pre-defined patterns and priorities

Enabled	Color	Proxy Name	Proxy Notes
✓	🟦	HadoopProxy	Proxy
✓	🟦	Default	These are the settings that a...

5- Finally, build a tunnel between your machine and the cs-master machine (using any terminal in Linux or Cgywin in Windows) using the following command

```
> ssh –D XXXX <your-user-name>@cs-master.wpi.edu  (enter your password)
```

You need to leave the terminal running. XXXX is the port-number you set in Setp 2 above.

➢ At this point, you access all nodes and see your directory /user/<your-user-name>, and under that you should see the file you uploaded.

5- Running a simple job on Hadoop

Over the small file that you uploaded, you can run a simple job "word count" that counts the occurrences of each word in the file. An example code can be found at:

http://hadoop.apache.org/common/docs/r0.17.0/mapred_tutorial.html

The above link has instructions on how to run the code. Two things need to be changed in the instructions:

- When compiling refer to the following jar file

```
>javac -classpath ${HADOOP_HOME}/hadoop-core-1.0.0.jar -d
wordcount_classes WordCount.java
```

- When calling the program for execution, make sure the paths you pass as arguments are under your HDFS directory (you do not have access to other directories)

The job should run successfully and you can check the job execution here:

http://cs-master.wpi.edu:50660/jobtracker.jsp