

SI 618 Lab 6: Getting set to run MapReduce jobs with Spark

Objectives:

1. Set up Great Lakes HPC cluster access, including remote terminal and file client programs.
2. Get familiar with some basic Unix command line utilities
4. Write a simple Map-Reduce job with interactive Jupyter + Spark

Note: Parts of this lab use, or are designed to follow, excerpts from the [Great Lakes User Guide](#).

1. Getting access to Great Lakes

To access the HPC cluster, you need to first obtain a Great Lakes account.

1a. Obtaining a Great Lakes account

Please see <https://arc.umich.edu/greatlakes/user-guide/> for a quick guide to accessing this.

1b. Preparing for 2-Factor Authentication using Duo

If you don't have 2-Factor authentication already set up, please follow these detailed instructions here: <http://documentation.its.umich.edu/2fa/enroll-smartphone-or-tablet-duo>

Note: If you are using a network other than MWireless, you will need to use a VPN (Virtual Private Network). If needed, follow the instructions here: <http://www.itcom.itd.umich.edu/vpn/>

2. Working with a remote server

For our Spark assignments, the computing and storage of datasets will no longer be done on your laptop. Instead, you have an account on a server cluster run by UM's Advanced Research Computing (ARC) center. You will need to upload the datasets and your code to the remote server, and run your code on Great Lakes cluster.

You will use your laptop (or other local client machine) for (a) logging into the cluster and then running commands in a terminal window, and (b) transferring files to/from the server cluster. For example, you might want to grab the output file from a script on the server so that you can upload it to Canvas from your laptop. Or you might edit your code in an editor on your laptop, and transfer it to the server when you want to run and test it remotely. The next two steps help you get set up to do that.

2a. Logging in

To login, you need to have a terminal window that provides a secure connection to the server.

- If you use a Mac, its Terminal app is all you need.
- If you use Windows, you can install Putty (<http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>).

Note: Logging into the remote server from off-campus. For security reasons, if you want to use your GreatLakes account from off-campus locations, you will need to connect via VPN (Virtual Private Networking). This requires you to install a VPN client if you don't already have one. Please see the University of Michigan instructions [here](http://www.itcom.itd.umich.edu/vpn/): <http://www.itcom.itd.umich.edu/vpn/>. Once you follow the instructions to connect using the VPN client, your laptop will be "virtually" on campus, and all other steps to connect, transfer files, etc. are the same.

Make sure that you can login by running in your terminal (replace uniqname):

```
ssh uniqname@greatlakes.arc-ts.umich.edu
```

On **Windows**, if you use Putty, launch Putty and enter greatlakes.arc-ts.umich.edu as the host name then click open.

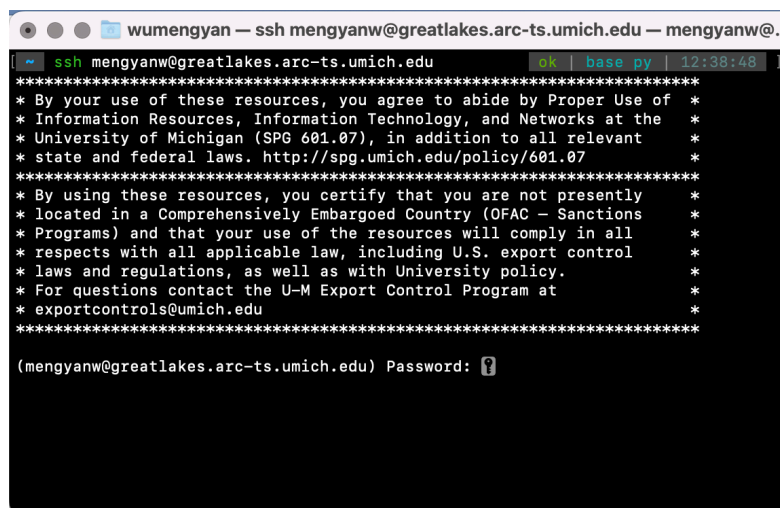
More detailed login instructions can be found here: [https://arc.umich.edu/greatlakes/user-guide/\(1.3 Getting Started \(Command Line\)\)](https://arc.umich.edu/greatlakes/user-guide/(1.3%20Getting%20Started%20(Command%20Line)))

When you're connecting for the first time, you might see a message similar to this one:

*The authenticity of host 'greatlakes.arc-ts.umich.edu (141.211.192.39)' can't be established.
RSA key fingerprint is 6f:8c:67:df:43:4f:e0:fc:80:5b:49:1a:eb:81:cc:54.
Are you sure you want to continue connecting (yes/no)?*

This is normal. By typing "yes" and pressing enter, you're accepting the public SSH key for the system. More detailed explanation could be found in the Great Lakes user guide.

If connected successfully, you should see a login prompt similar to this:



```
wumengyan — ssh mengyanw@greatlakes.arc-ts.umich.edu — mengyanw@..
ssh mengyanw@greatlakes.arc-ts.umich.edu
*****
* By your use of these resources, you agree to abide by Proper Use of *
* Information Resources, Information Technology, and Networks at the *
* University of Michigan (SPG 601.07), in addition to all relevant *
* state and federal laws. http://spg.umich.edu/policy/601.07 *
*****
* By using these resources, you certify that you are not presently *
* located in a Comprehensively Embargoed Country (OFAC – Sanctions *
* Programs) and that your use of the resources will comply in all *
* respects with all applicable law, including U.S. export control *
* laws and regulations, as well as with University policy. *
* For questions contact the U-M Export Control Program at *
* exportcontrols@umich.edu *
*****
(mengyanw@greatlakes.arc-ts.umich.edu) Password: 
```

Once you enter your UM password and pass the authentication stage, you should be successfully logged in, and see a message like this:

```
Passcode or option (1-3): 1
Success. Logging you in...
*      Advanced Research Computing - Technology Services      *
      University of Michigan
      arcts-support@umich.edu

      Welcome to Great Lakes

User documentation: http://arc-ts.umich.edu/greatlakes/user-guide/
The folders under /scratch are intended for data that is in active
use only. Please do not store data there for longer than 60 days.
For usage information, policies, and updates, please see:
https://arc-ts.umich.edu/document/greatlakes-policies/
-----
[mengyanw@gl-login2 ~]$
```

2b. File transfer between your laptop and server:

Download `uniquename_si618lab6.ipynb` from canvas. Copy this file to your home directory on the server. To do this, you can either use:

- 1) Terminal/Putty
- 2) Install an SFTP **client** like FileZilla (<https://filezilla-project.org/>) or CyberDuck.

When Terminal/Putty is running:

- Make sure you're **not** on the greatlakes cluster and running on your local
- Run command (replace path to your file with the path in your local computer and `uniquename` with your username):

Mac:

You could use this command to transfer file one by one:

```
scp path to your file uniquename@greatlakes-xfer.arc-ts.umich.edu:/home/uniquename
```

Windows:

```
pscp -scp path to your file uniquename@greatlakes-xfer.arc-ts.umich.edu:/home/uniquename
```

- You could use `scp -r` to transfer the whole directory together
- It'll ask you for your password and then to authenticate via duo

When FileZilla/CyberDuck is running:

- File -> Site Manager -> New Site
 - Select the "SFTP (SSH File Transfer Protocol) option"
 - Host: `greatlakes.arc-ts.umich.edu`
 - User: `youruniquename`
 - Logon Type: Interactive
- Click Connect.
- You should see your Linux home directory on the cluster as a Remote Site panel on the right.

On each side, the top panel helps you navigate the directory structure while the bottom panel shows you the file. The files starting with '.' (dot) are hidden files. Move the file to your home directory.

3. Obtain your first Spark Cluster on Great Lakes

To obtain a Spark Cluster on Great Lakes, you will need to:

1. Log in to <https://greatlakes.arc-ts.umich.edu/>.
2. Click *Interactive Apps* on the top of the page, create a Jupyter + Spark Basic session. **Use siads618f22_class as the slurm account.**

When choosing the number of hours, please **do not select more than 4 hours** for a single session. Apply for a new session only if you've canceled/completed the previous sessions. This is because if everybody requests lots of resources at the same time, some people might enter a long queue to start their sessions.

3. Wait 1 or 2 minutes. Once your session is running, you could connect to the Jupyter notebook on the Spark cluster.

4. Use PySpark to analyze the Google NGram dataset

The big data files we'll be using come from the [Google NGrams dataset](#). These datasets contain counted syntactic n-grams (sequences of one or more words) extracted from the English portion of the Google Books corpus. *(If you're interested, the datasets are described in the following [publication](#). The dataset format and organization are detailed in the [README file](#).)*

The ngrams dataset is tab-delimited and of the form (commas inserted for clarity):

ngram, year, total occurrence count, number of volumes ngram occurred in

As an example, the row

```
circumvallate    1978    335    91
```

tells us that in 1978, the word "circumvallate" occurred 335 times overall, in 91 distinct books of our sample. (Excerpted from [Google NGrams dataset](#))

For now, we'll just work with 1-grams (single words) that occurred in books dating back to 1500, and up to the present day.

The size of the datasets we are going to use in this lab is around 3GB. It is very hard to load and do complex analyses on such big datasets locally, and that is why we need great lakes. You will be able to see the power of Spark and Great Lakes through this lab.

We stored the dataset on Great Lakes under `/scratch/siads618f22_class_root/siads618f22_class/shared_data/lab6_data`, so that you do not need to download the dataset manually.

You can run this command to list all the files in our directory

```
ls /scratch/siads618f22_class_root/siads618f22_class/shared_data/lab6_data
```

You can run this command to view the content of the dataset

```
head -5 /scratch/siads618f22_class_root/siads618f22_class/shared_data/lab6_data/*
```

Todo #1: What are the characteristics of most of the 1grams contained in the datasets?

You can also dump the contents of files using the "cat" command. Here, I've added a pipe "|" symbol followed by the "grep" command. The pipe symbol means the output of the "cat" command (the file contents) will be piped to the input of the "grep" command, which searches the input for the given regular expression.

```
cat /scratch/siads618f22_class_root/siads618f22_class/shared_data/lab6_data/* | grep  
"^information_NOUN"
```

Todo #2: What year was *information* first mentioned (as a noun) in our dataset?

4. Run a sample PySpark job

Open `yourusername_si618lab6.ipynb` on the spark server. Follow the instructions in the notebook and complete the PySpark code. You will need to use PySpark to compute the average length of words in the dataset, per year.

Here is a high-level walk through the transformations used in the script:

1. Transform each line of the original dataset into an array of the form `[ngram, year, occurrences, volumes]`
2. Create a new dataset *length* with each row containing an array of the form `[year, total length of the words]`
3. Create a new dataset *count* with each row containing an array of the form `[year, total number of words]`
4. Create a new dataset *average_length*, dividing *length* by *words*, resulting in an array of the form `[year, average word length]`

Why create a new dataset at each stage? To ensure many processes can operate on the same dataset at once, without worrying that some other process might deliberately or accidentally change it, each object is **immutable** - it can *not* be modified once created. This is a common feature of highly **concurrent** programming languages (those designed to make it easy to do large-scale computing in parallel).

5. Looking at Spark output

Concatenate the output files to a single file and name it *username_ngrams_output.csv*, just like what we did last week.

Todo #3: Look at your result, did you notice any interesting patterns?

6. Copy from the server to local

```
scp username@greatlakes-xfer.arc-ts.umich.edu:remotefile localfile  
pscp -scp username@greatlakes-xfer.arc-ts.umich.edu:remotefile localfile
```

What to submit:

1. username_si618lab6.ipynb and username_si618lab6.html
2. username_ngrams_output.csv
2. a .txt file containing your response for the 3 todos.

Rubric:

Todo #1: 5pts

Todo #2: 5pts

Todo #3: 10pts

Spark job:

Step1: 10pts

Step2: 15pts

Step3: 15pts

Step4: 20pts

Step5: 10pts

Step6: 10pts

References:

Great Lakes User Guide

Spark Example