Module 2: Big Data Workshop

# Goal

* Explore more Unix tools for moving and manipulating data (wget, curl, awk, and xargs/parallel)
* Getting data via AWS CLI (command line interface): [Use a free tier AWS account: <https://aws.amazon.com/free> ]
* Introduction to distributed computing and storage via Hadoop.
* Put data in a Hadoop standalone cluster (not distributed here; just for testing)
* Spark for (very basic) data manipulation

# Read

Top Hadoop analytics tools for 2021: <https://data-flair.training/blogs/hadoop-analytics-tools/>

# GitHub

We will take a few minutes in class to set up GitHub connections to manage and backup your code from the remote cluster. General instructions are here: <https://happygitwithr.com/rstudio-git-github.html>. These instructions rely on RStudio as a git client and interface to the Unix shell so we will tour RStudio access on the cluster as well (no installation or R coding required!).

# AWS

We will cover downloading data from AWS using the command line interface (aws cli). The target data can be accessed without a login, but you may want to set up an AWS account for other queries.

OPTIONAL info about AWS:

Big Data on AWS Tutorial - <https://aws.amazon.com/getting-started/hands-on/analyze-big-data/?trk=gs_card> \*Not our target but a good demo of moving these solutions to AWS.

Cost calculator: <https://calculator.aws/#/>

AWS Free Tier: <https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/billing-free-tier.html>

# Instructions: A Language Menagerie

***Setup***

Log into your BU cluster account: <https://scc-ondemand2.bu.edu/>

*Interactive Apps 🡪 RStudio 🡪 Request 4+ hours, 8 cores, and load modules ‘spark java/10.0.2’*

***AWS CLI***

On the command line enter:

“aws configure”

Enter (using your account details 🡪 MyAccount 🡪 My Security Credentials):s

***AWS Access Key ID: \*YourAccessKey\****

***AWS Secret Access Key: \*YourSecretKEY\****

***Default region name: us-east-1***

***Default output format:***

***GitHub for code management***

In-class we will demo Git/GitHub setup. Loosely based on these instructions:

<https://happygitwithr.com/rstudio-git-github.html>

1. Set up Git repository on github.com
2. Clone the new GitHub repository to your computer
3. Make changes (write stuff), save, commit. \*\*This is how you track changes/versions
4. Push local changes to central (online) repository.

***More Unix Fun***

Our aim in this section will be to use some Unix command line tools to access \*big-data\* from two sources. 1) The NYC Taxi data repository (using *AWS CLI*), and 2) The NYSE data using the AlphaVantage API and basic Unix tools for downloading and processing data.

With bigger datasets it pays to handle data with command line programs that are efficient and let us move and modify data without GUIs or downloading to a laptop. For example, the NYC Taxi data in its complete form is ~300 GB. Who has disk space for that? Well, the cloud does. So we’d better use tools that copy data directly to a remote resource.

*NYC Taxi:* Code dissection of

<https://github.com/rsh249/malthus/blob/main/bash/nyc_taxi.sh>

*AlphaVantage API:* Code dissection of

<https://github.com/rsh249/malthus/blob/main/bash/av_test.sh>

*Important Unix commands for today:*

The pipe |

wget

awk

sed

xargs

***A primer on Hadoop***

*Lecture from:* [*https://data-flair.training/blogs/hadoop-tutorial/*](https://data-flair.training/blogs/hadoop-tutorial/)

*Live Demo:*

* BU environment setup: bashrc configuration
* Basic hdfs commands
  + *‘hdfs -put folder’ #Add ‘folder’ to Hadoop file system*
  + *‘hdfs –ls’ # list files on Hadoop file system (in standalone mode this is the same as system ‘ls’*

***Spark for (very basic) extract, transform, load (ETL) operations:***

Code Dissection:

<https://github.com/rsh249/malthus/blob/main/spark_ex/spark_avdata.scala>

<https://github.com/rsh249/malthus/blob/main/spark_ex/spark_nyctaxi.scala>

Tutorial (watch after class): <https://youtu.be/QaoJNXW6SQo>

# Post-Class Assignment: Designing Testable Questions

OK so now we have some data and a few tools to start working with large datasets. What do we want to know?

*Watch:* [*https://youtu.be/Np1T7VNnfBk*](https://youtu.be/Np1T7VNnfBk)

*Read:* [*https://www.datasciencegraduateprograms.com/research-design/*](https://www.datasciencegraduateprograms.com/research-design/)

*Design:* Write a specific, testable set of questions that address a potentially interesting aspect of the two datasets we have targeted so far. Consider here how you might go about answering the question(s) in terms of data partitioning, modeling, stats, or visualizations.

*Example:* We now have a long record of stock exchange data and can pull down more. Obviously investors want to maximize returns by making good investments and there are probably many ways to achieve their goals. Can we reconstruct investment strategy histories and compare performance over time? (e.g., Which has a higher return: long term holdings vs. hand picking stocks in “hot” sectors?)

Post to Slack:

1. *What general topic interests you?*
2. *What specific question would you like to address?*
3. *How (given these data) are you going to design a test? (e.g., parse the data and build + compare models; extract data and perform stats; or visualize data).*

DUE: Sunday February 21, 2021 @ 11:59PM