**Week 5 Assignment**

Monroe College

CS 675: Big Data Management and Analytics

Madeline Lin

Professor Syed Farid

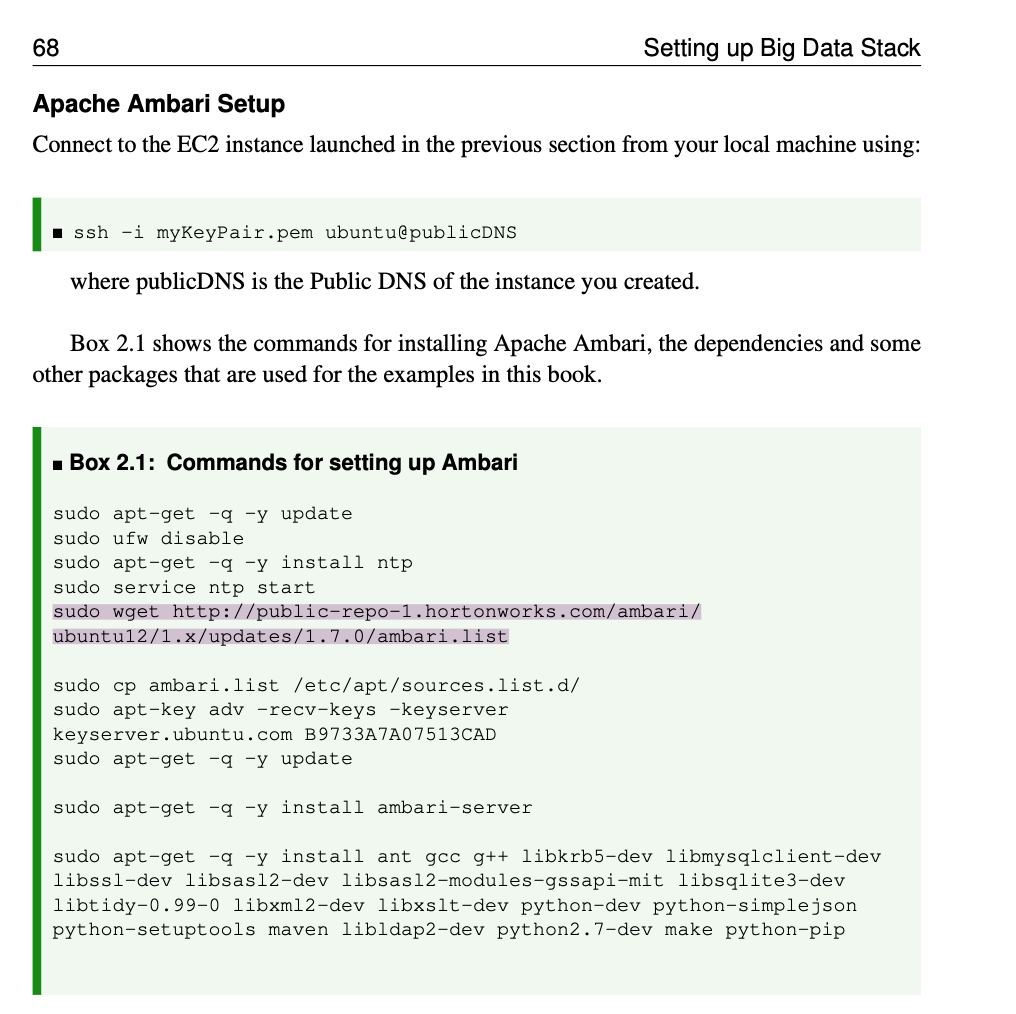
October 9th, 2022

*After reading the material and continue to build your Hadoop framework follow the instructions on the worked examples. Create a document with your results using Hadoop frame work. Also the section with the lab you will be working with Anaconda frame work to perform some of the statistical analysis, also create a file with your results. Upload the file using this section. Please save your files as PDF then upload them. Next week we will have a test base on statistics analysis.*

**ANS:**

Our group had a discussion and were still unable to resolve the issue we are facing with installation. As a team, we tried on different machines but still get the same error. After installing the Ambari server build fails each time with different errors after resolving previous errors (mvn -B clean install jdeb:jdeb -DnewVersion=2.7.6.0.0 -DbuildNumber=388e072381e71c7755673b7743531c03a4d61be8 -DskipTests -Dpython.ver="python >= 2.6" -Drat.skip=true -X).

Below attached is a screenshot of an error. We tried installing it in both the EC2 Instance and the Virtual box (ubuntu). We were unable to run the highlighted command. We have sent you an email regarding the above issue. Could you please suggest helping to resolve this issue?





Additionally, please see the below summaries of the reading materials.

**1. Big Data Analysis Ch4 Lab (Python and Hadoop)**

This file firstly discusses about the installations of Python, Anaconda, Jupyter Notebook and Conda. Then it discusses about how to use the above tools and then install packages to access Hadoop (HDFS). After that, it uses a real example to illustrate how to analyze the data in a Hadoop cluster. When it comes to the data analysis part, it firstly imports the pandas library and then use HDFS to load the data into a pandas data frame. After that, it gives examples of how to merge, join, remove duplicate data, etc.

Python is a popular programming language for data science. Anaconda is an all-in-one installer with an emphasis on science; it includes Python, the standard library, and many useful third-party libraries. Conda supports multiple environments that can coexist so we can set up different versions of Python.

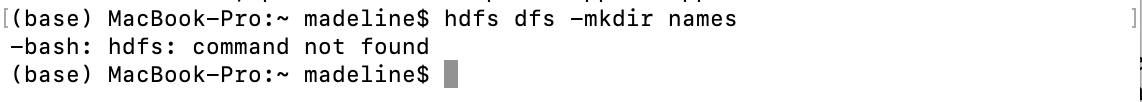
I previously have used Python via Spyder in my objective oriented software class; therefore I have both Python and Anaconda installed in my laptop. In terms of Jupyter Notebook, we start using this platform in my data science course this semester. As I’ve had Anaconda, I just need to open Jupyter via Anaconda.

**2. Worked Examples Assignment**

This file uses four examples to demonstrate how we use Hadoop as a new model for big data processing. Various methods for making data available to several Hadoop tools were outlined. The examples included copying files directly to HDFS, importing CSV files to Apache Hive and Spark, and importing JSON files into HIVE with Spark.

*Example 1: Commands for loading CSV data into a HIVE table*

As discussed, we had issues of installing Hadoop, thus when I typed the commands in terminal, it shows “command not found”.

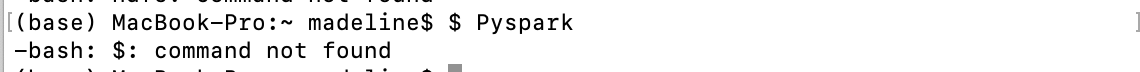


To my understanding, this example shows us how to create a name in Hadoop, move a name in Hadoop, how to create an external HIVE table, and check if names are there. After creating the HIVE table, we can play around with it such as copy data externally to internally, create partitioned table, select names, add data, save table, etc.

*Example 2: Commands to enter data CSV/JSON data into Spark using PySpark*

Apache Spark is a platform for batch processing, stream processing and machine learning. PySpark is the Python API for Apache Spark, an open source, distributed computing framework and set of libraries for real-time, large-scale data processing. PySpark is a good language to learn to create more scalable analyses and pipelines.

I am unable to run Pyspark as indicated below.



To my understanding, this example shows us how to use PySpark and import needed modules, create the sql context, then show the data frame, use HIVE commands, and read HIVE table from Spark. After that, it also demonstrates how to read the JSON file and load the data.

*Example 3: Using Apache Sqoop*

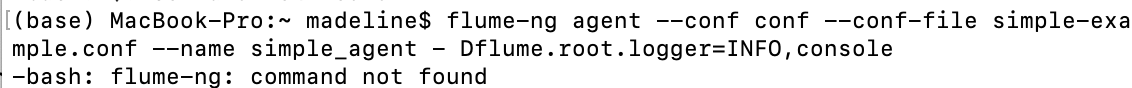
Based on my research, Apache Sqoop is a common-line interface application for transferring data between relational databases and Hadoop. The Apache Sqoop was retired in June 2021 and moved to Apache Attic.

This example gives instructions on how to extract the database table from MySQL to HDFS and then back to MySQL. Firstly, we need to install MySQL. Next, we are supposed to load the database into MySQL. Next, we should add Sqoop user permissions for local machine and cluster. After that, we should import data using Sqoop. Finally, we use Sqoop to list the tables, make directory for data in HDFS, check the actual data file, and check the resulting file, etc. Finally, we need to export data from HDFS to MySQL.

*Example 4: Using Apache Flume*

Apache Flume is a distributed, reliable, and available software for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows.Apache Flume was presented as tool for capturing and transporting continuous data, such as web logs, into HDFS.

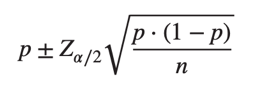
I think due to the fact that we haven’t installed Hadoop, we are not able to use the Flume for Simple Flume Test nor Web Log Example.



The example 4 has two sub-examples. First, it shows us how to perform a simple flume test by firstly starting a flume agent and then connect the agent in another window. Second, it shows us how to record the weblogs from the local machine and place them into HDFS using flume.

**3. Statistical Analysis Ch5**

This file contains the theory of statistics, specifically regarding how to use statistical approaches to estimation and prediction. Below is a brief summary.

Simply speaking, we are using observed sample statistics to estimate unknown population parameters and want to select the estimate with highest confidence. The sample error (i.e. | statistic – parameter | or |x − 𝜇| can be regarded as the distance between the observed sample mean and the unknown population mean. We want to choose the least sample error as our estimation. Moreover, in terms of the confidence interval estimation of the mean, we introduce a concept of the confidence interval, which is the point estimate ± margin of error, where the margin of error is a measure of the precision of the interval estimate. Smaller margins of error indicate greater precision. *x* ± *t𝛼*∕2(*s* ∕ *n*) is the formula of the margin of error. Usually, we find the 95% t-interval, which is 1.96. In terms of the confidence interval estimation of the proportion, we introduce the formula, , where the sample proportion p is the point estimate of pi and the quantity represents the margin of error. For 90% confidence, *Z𝛼*∕2 equals to 1.645; for 95% confidence, *Z𝛼*∕2 equals to 1.96; for 99% confidence, *Z𝛼*∕2 equals to 2.576. Finally, the file talks about the hypothesis testing. The null hypothesis is always rejecting H0. We want to find the p value to be small as possible so that the result will be more accurate.