**Week 3 Assignment**

Monroe College

CS 675: Big Data Management and Analytics

Professor Syed Farid

September 27th, 2022

**Group 1**

Gita Gurung

Madeline (Zhiyang) Lin

Priyankaben Dashrathbhai Chaudhari

Sukriya Gurung

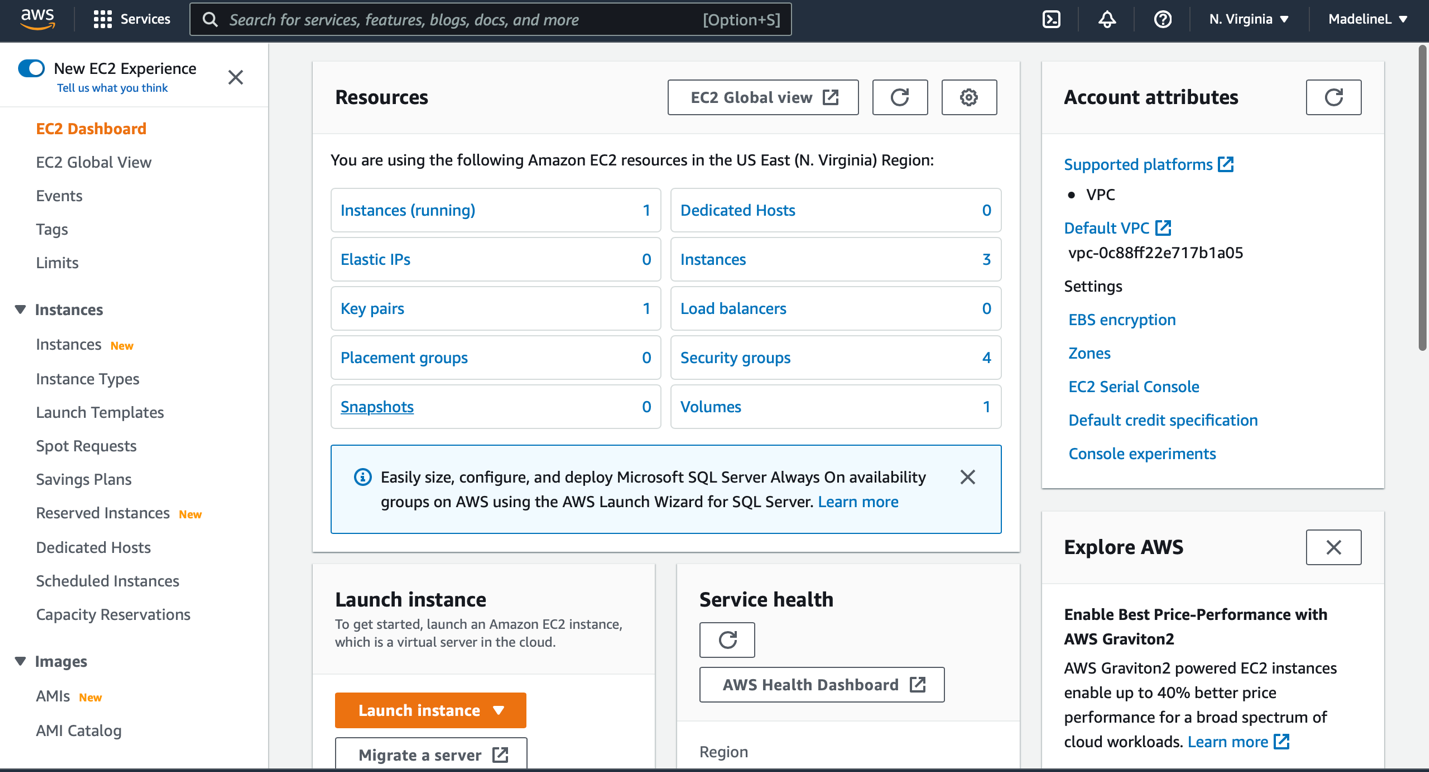
Following the lab file, me and my group tried launching an AWS EC2 instance running Ubuntu.

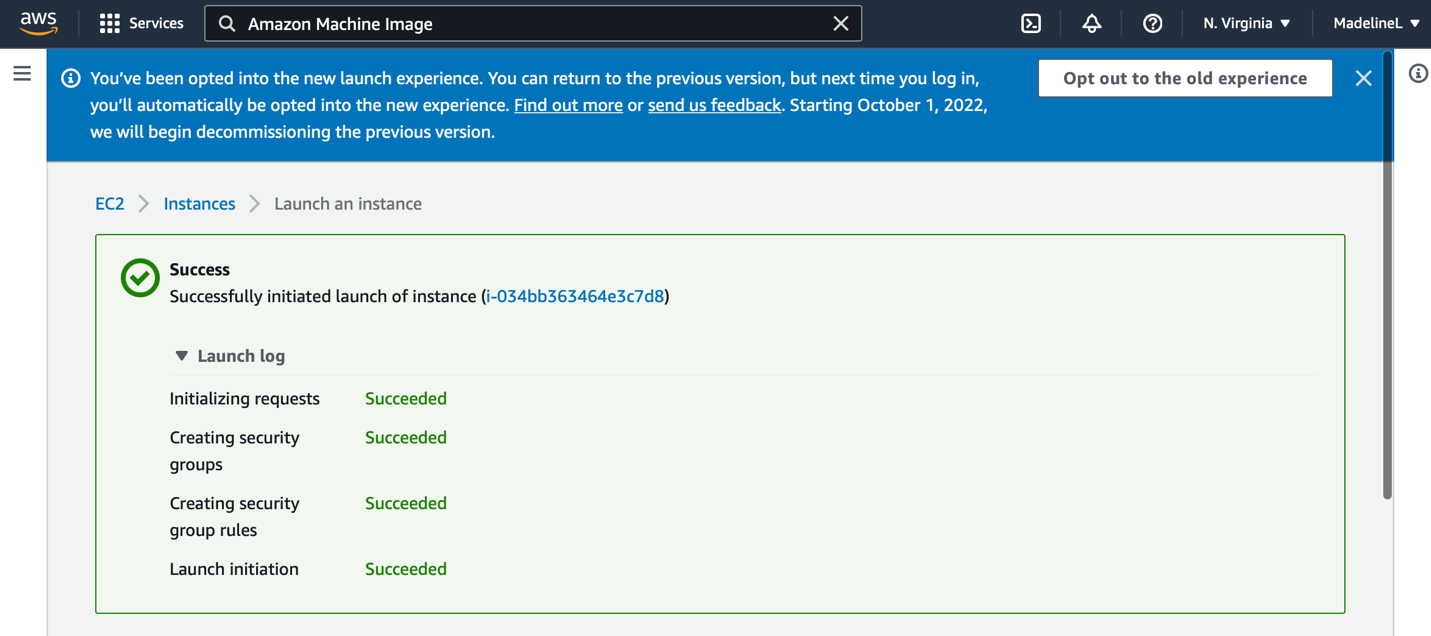
After creating my AWS account, we searched the EC2 in the search bar. Next, we went to EC2 Dashboard and selected “Launch instance”.

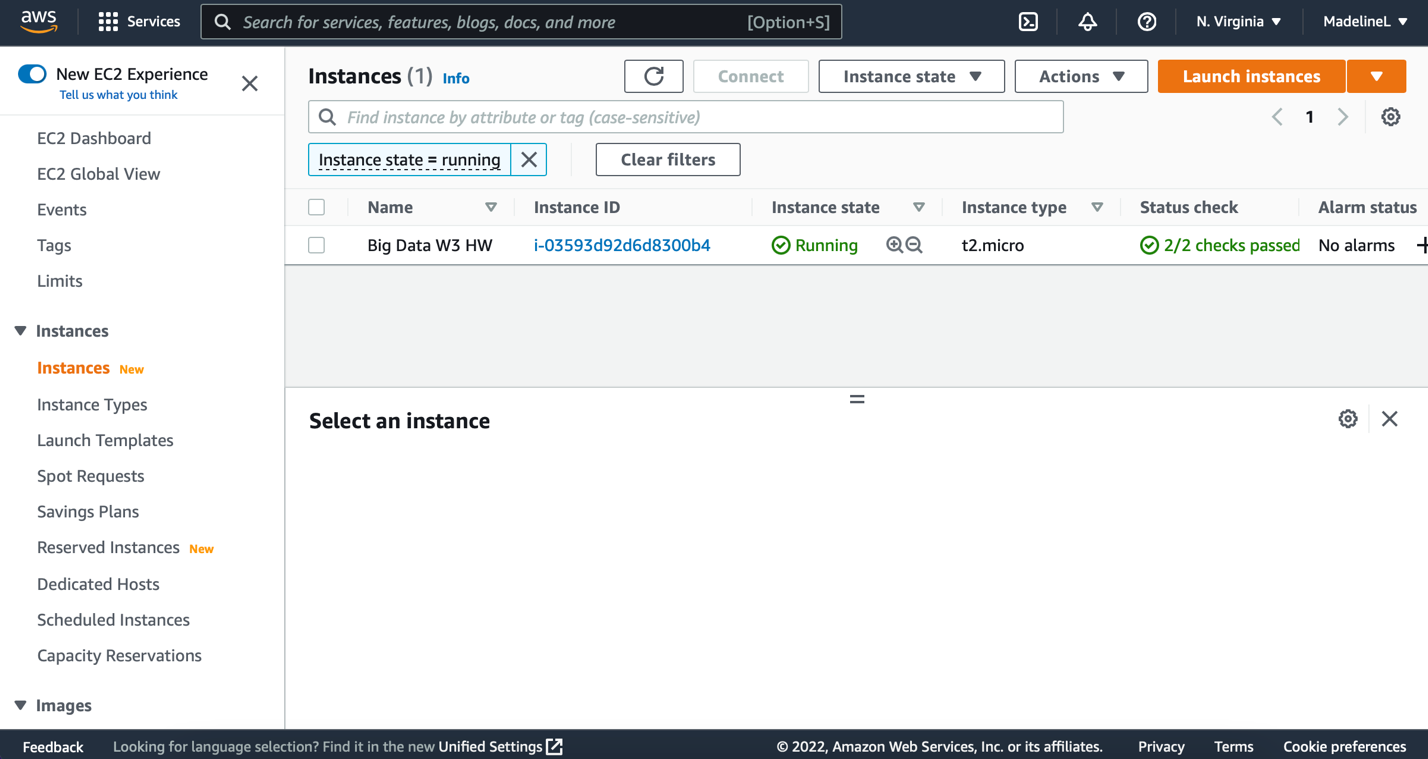
In the “Launch an instance” page, we created the name “Big Data W3 HW”, chose Ubuntu, 62-bit as Architecture, name the Key pair name to be “ML”, selected “Allow SSH traffic from”, “Allow HTTPs traffic from the internet”, and “Allow HTTP traffic from the internet”, used “30” GiB as Configure storage, and launched the instance.

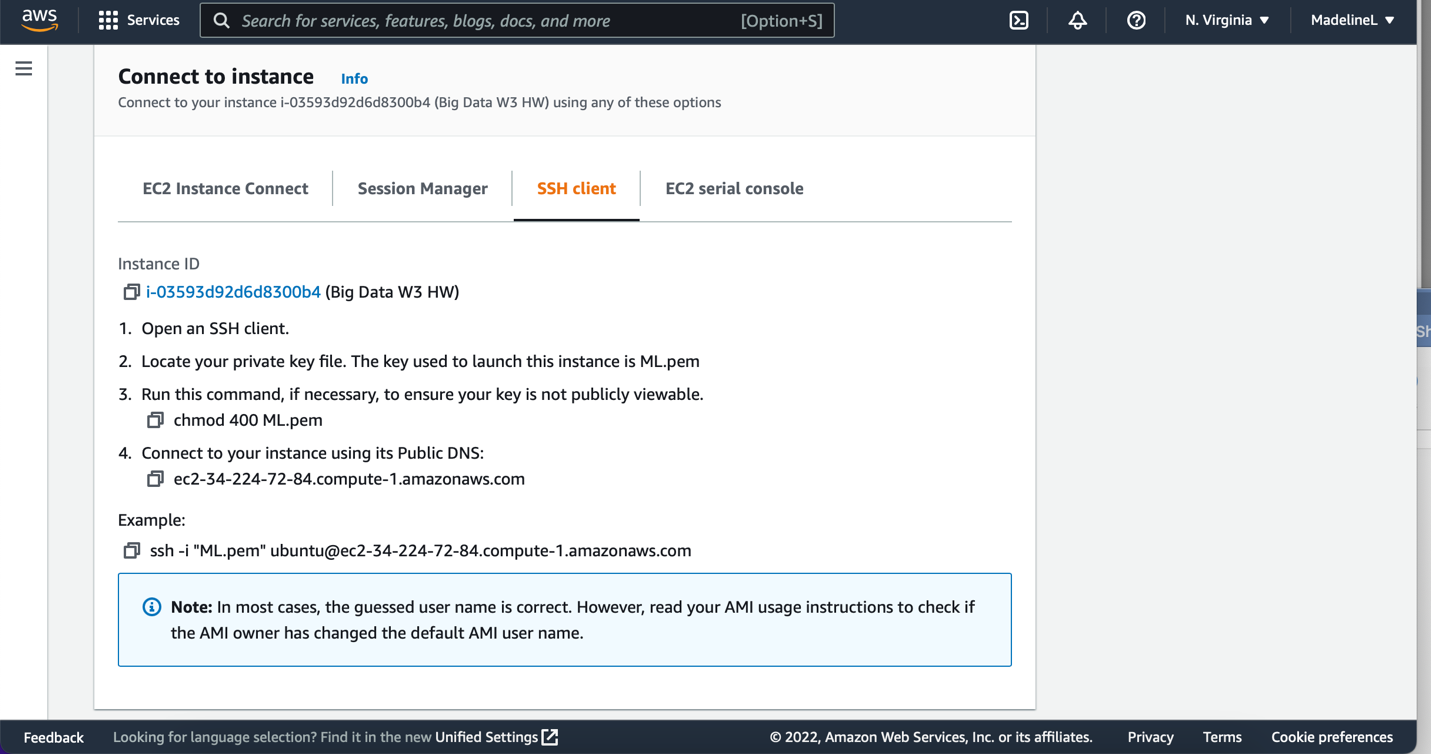
After launching the instance, we connected it to instance. We chose “SSH client”, located the private key file, run the command of “chmod 400 ML.pem” and connected to our instance using its Public DNS.

Below are the screenshots of what we’ve performed in AWS.

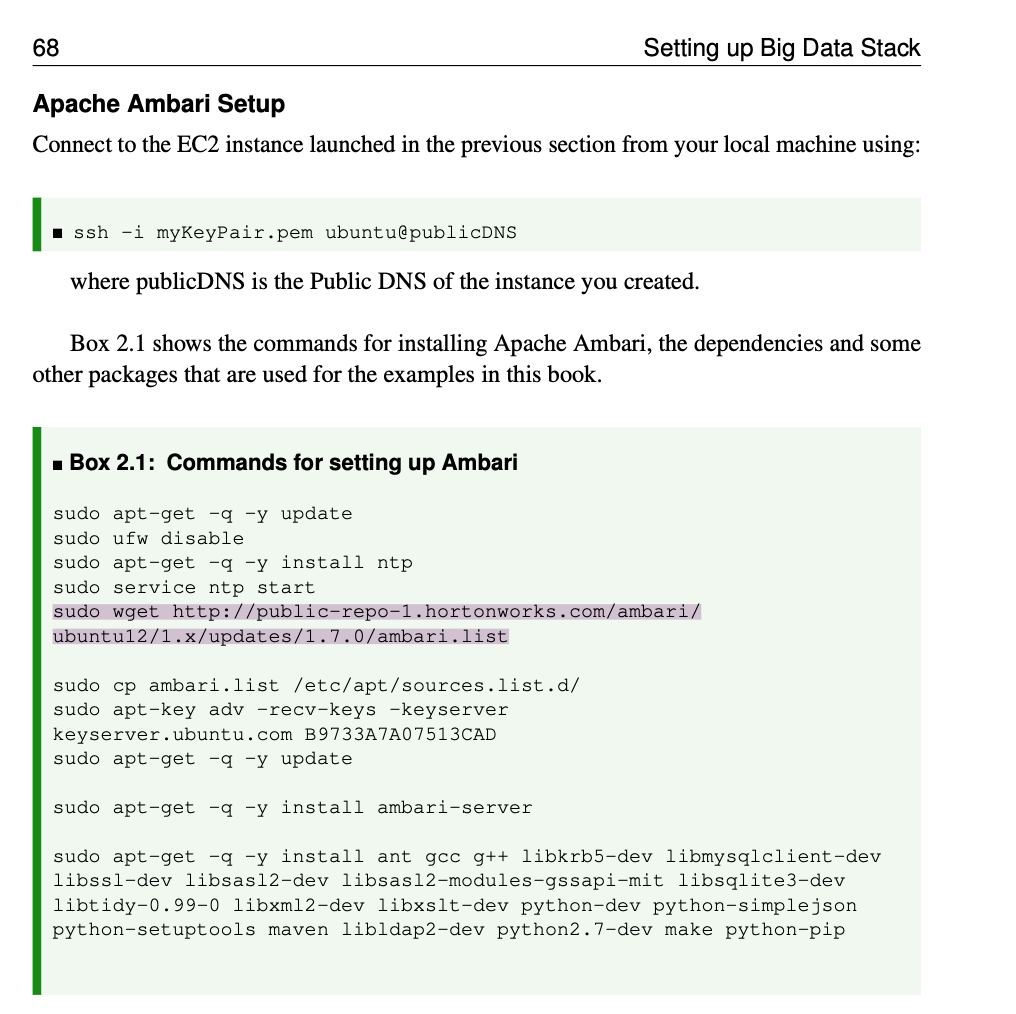








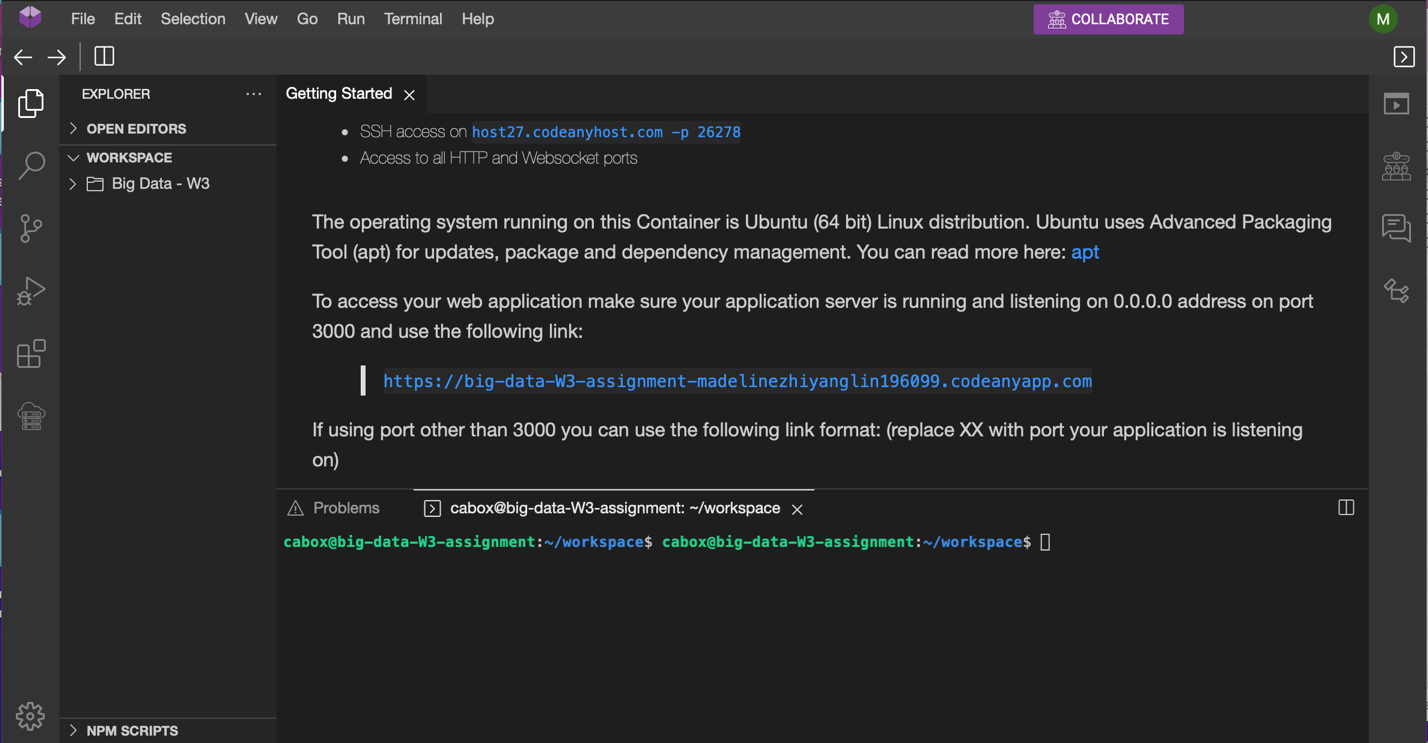
After all the above steps, we tried the Apache Ambari Setup and got stuck in the Box 2.1: Commands for setting up Ambari – the purple highlight line of code. It shows “403 forbidden” as indicated in below. The potential reason is that Cloudera has stopped freely offering the Ambari installation starting from this year based on our research.





Alternatively, we signed up for a Hadoop Online IDE (i.e. codeanywhere - <https://codeanywhere.com/languages/hadoop>) for this exercise. This operating system running on this container is Ubuntu (64 bit) Linux distribution as shown below.

By now, we were not able to figure it out how to use the Hadoop IDE to set up the source of the data and load that dataset. Please advise if possible.



Based on the website provided above, we selected “Red & White Wine” data which we are interested in (i.e. red wine - '<https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv>').

From our perspective, the first thing that we should do to feel about the data is to understand the data. In this case, this dataset contains 1600 pieces of information of the red wine with attributes including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality. We are supposed to have a basic understanding about each attribute. For instance, acids impart the sourness or tartness that is a fundamental feature in wine taste. Wines lacking in acid are “flat”. Chlorides add to the saltiness of a wine, which can contribute to or detract from the overall taste and quality of the wine. Without sulfites, wines would oxidize very quickly and, therefore, ruin the wine’s freshness and flavor.

Things we’ve learned about this exercise include: 1) Launched an AWS EC2 instance, 2) Signed up an online IDE for Hadoop, 3) What should be the first step when we receive a dataset.