Analyzing 10Gb of Yelp Reviews Data

For this project, you will be tasked with provisioning a Spark Cluster on AWS EMR for loading and running some analysis on Yelp’s Reviews and Businesses dataset (about 10gb) from [**Kaggle**](https://www.kaggle.com/yelp-dataset/yelp-dataset#yelp_academic_dataset_user.json). You will run your analysis via Jupyter Notebook and the expected output artifact is a **.ipynb** file deployed to Github.

[**Requirements**](#_7jhm4q2zsy5m) **2**

[**Artifacts**](#_gcvq0fmil4i) **2**

[Notebook File](#_rdv81kksecfa) 2

[README](#_j9huuhiwntdf) 2

[S3 Bucket](#_ff3jwc9chkq7) 4

[Extra Credit](#_xgxj1fa629vv) 4

[Submission](#_jjtwlm2mnm0o) 4

[**Assignment**](#_fjy4akwrp9sy) **4**

[Part I: Installation and Initial Setup](#_s2djl8qglt8i) 5

[Part II: Analyzing Categories](#_44vrcxnjnzj0) 5

[Part III: Do Yelp Reviews Skew Negative?](#_jqb7u1tarerj) 5

[Part IV: Should the Elite be Trusted? (Or, some other analysis of your choice)](#_yt9f987fth5h) 5

# 

# 

# 

# 

# 

# 

# Requirements

This project is very simple: you are to provision a Spark cluster on AWS EMR, connect it to a Jupyter Notebook and then run a series of queries (in python with DataFrame API or Spark SQL) that answer a few simple questions about the Yelp Data available.

In doing so, you are demonstrating your ability to configure and provision infrastructure using the AWS Elastic Map Reduce ecosystem. Also you are demonstrating your understanding of how to leverage transformations and actions (as per the Spark terminology) with **PySpark** in performing basic data analysis tasks on information sources that are too large to manage in memory.

This project is due **MONDAY MAY 4TH, MIDNIGHT**. **NO EXCEPTIONS.**

**PLEASE READ THE FOLLOWING INSTRUCTIONS CAREFULLY**. I will likely be automating some or most of the grading process and therefore your output artifact must match the spec I define below.

# Artifacts

You are to create a Github Repo with a well formatted title (not something like “Project II”). The Github repo must contain the following:

## Notebook File

The **ipynb** file that contains your analysis and **the outputs of the code you wrote** to arrive at your results. This is very important as this is the sole method of validation that you actually ran an EMR cluster successfully.

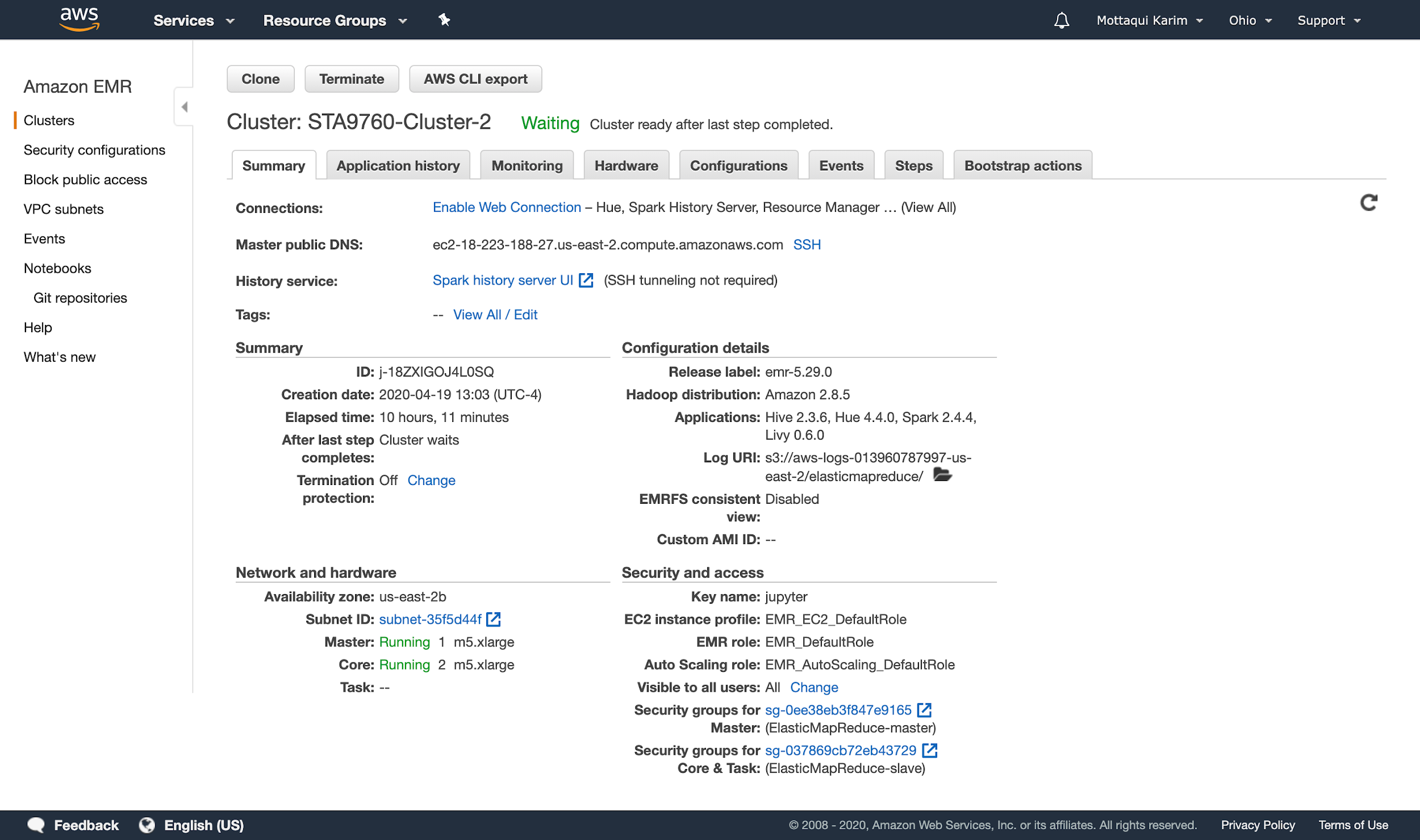
You **must** name your Notebook file **Analysis.ipynb**.

## README

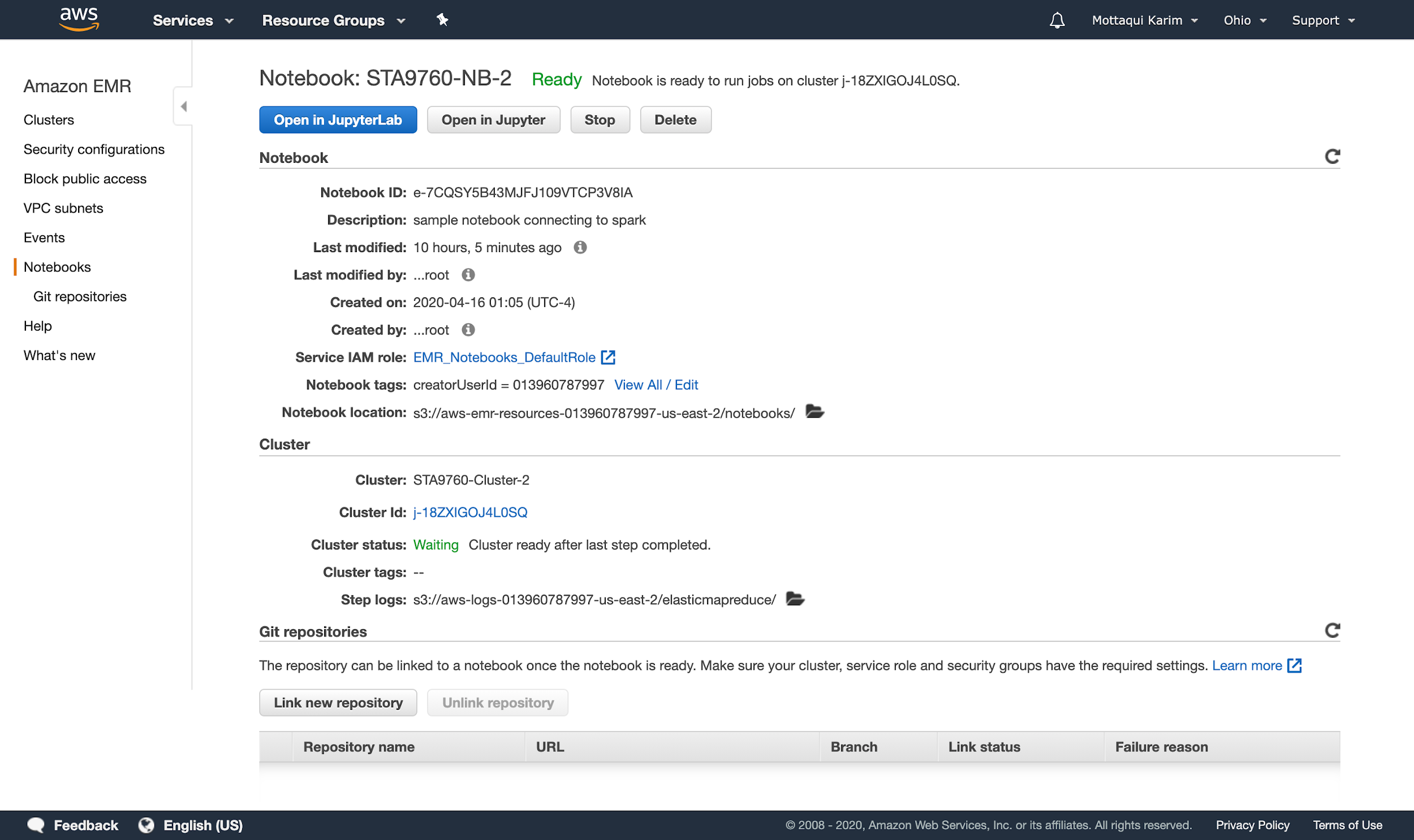
The **README**, in markdown, should contain a brief blurb describing the project and the technology leveraged to conduct your analysis. This ought to be brief and informational, in case folks in the future want to recreate your results.

**ALSO**, your README must contain screenshots of your EMR cluster configuration and Notebook configuration. Here are mine, shared below as reference

**Cluster Configuration**

****

**Notebook Configuration**

****

## S3 Bucket

You **must** read your Yelp data from **S3**. In order to do so, you must download the Yelp data and then upload it back to an S3 bucket that is accessible via EMR. Your **Analysis.ipynb** file must demonstrate that the data is being read from **S3** - this is largely as simple loading your DataFrame like so:

|  |
| --- |
| df = spark.read.json('s3:/[YOUR\_S3\_BUCKET\_NAME]/[SUBFOLDER\_NAME]/\*.json') |

But, you must also provide links in your **README** to at least one of your S3 hosted files. (This entails ensuring that the file is available for “public access” so that I can curl against it and ensure that it exists. If you have trouble with this step, you can also just share screen shots of the bucket management page in your README. I’ll leave that part up to you.

## Extra Credit

For extra credit, I would accept a python or bash script that leverages [**the kaggle API**](https://github.com/Kaggle/kaggle-api)moduleto download the dataset and the **AWS boto3** module to upload to S3. This script must run in a docker container and it should work for anyone’s AWS and Kaggle accounts for any dataset. (For example, if I wanted to download kaggle dataset ABC and then upload it to my S3 bucket DEF - I should be able to manage this. This script must not be specific to the datasets or accounts for this project).

This script should be uploaded to your project github repo under the **script** folder. Your dockerfile and all other dependencies must live inside this folder.

**Upon successful completion of this extra credit assignment, I will convert two of your lowest quiz grades to 100% each.**

But: please keep track of time and ensure that you can get your project done first.

## Submission

[Please submit your Github Repo URL here.](https://airtable.com/shr94AE0YIzAZewfT)

# Assignment

The actual analysis is broken into four parts - three which are guided and one that is freeform. I have published a [**sample github repo**](https://github.com/mottaquikarim/STA9760_Project2_Yelp_Data_Analysis)demonstrating this project.

Note that the output of the code written is provided as a means to give you **structure** as you write your analysis. For **Parts I, II & III**, you must fill in the blanks (implement the code however you want) to get the output provided in the file. (Mainly columns and aggregations, I don't care about the exact rows).

For **Parts III and IV,** you have more flexibility to take the analysis further however you see fit. Below, I expound a bit more about each part of analysis.

## Part I: Installation and Initial Setup

In this portion, you will import the necessary dependencies (**pandas** and **matplotlib**) and load your dataset as a pyspark dataframe.

## Part II: Analyzing Categories

For this part, you will take a stab at denormalizing the categories that are associated with each business (there may be more than one, presented as a string of comma separated identifiers) and then running some basic analysis on the result.

## Part III: Do Yelp Reviews Skew Negative?

For this next part, you will attempt to answer the question: are the (written) reviews generally more pessimistic or more optimistic as compared to the overall business rating. There are some required questions you must answer (see the analysis.ipynb file) which is the bare minimum. But, feel free to have fun with it and take your analysis as far as you’d like. Any additional work you do will be counted for up to 5 points of extra credit on your project grade, capped at 105.

## Part IV: Should the Elite be Trusted? (Or, some other analysis of your choice)

For this final part you may choose to either answer the question posed or explore the data in some other manner of your own choosing. The only requirements are:

* You must leverage the **users** dataset provided
* You must have at least one data visualization as part of your analysis