Analytics on the cloud projects

# PROJECT 1: Vertex AI

In this project, you will use GCPs Vertex AI Studio to build a classification model. The goal of this project is to classify news stories as being either positive, negative or neutral. Start with the dataset at <https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news> which contains headlines already tagged as positive, negative or neutral. Then build a fine tuned model on gemini using GCP’s Vertex AI Studio. Remember to create training and testing datasets. There is plenty of help available including GCP tutorials, YouTube videos, but the best place to start is at: <https://cloud.google.com/vertex-ai/docs/text-data/classification/prepare-data>

The model fine tuning, testing etc., must be done using Python (or Spark, but that may not be easy) and the API. Submit your python code (including creating and formatting the data) and your training and testing results.

# PROJECT 2: Google Big Query

In this project, you will play around with [Google Big Query](https://cloud.google.com/bigquery?hl=en). Big Query is a data warehouse that can handle very large (theoretically unlimited) databases. Experiment with Big Query and submit a report that contains the following:

1. Two pages that summarize what Big Query is, what its capabilities are, and its limitations
2. Connect Big Query to Apache Spark (https://cloud.google.com/bigquery/docs/connect-to-spark)
3. Create a database with at least two tables from the 311 dataset that is currently on our canvas
4. Run a few queries using Spark and this database (at least one query should use a join on your tables)

(Note: You can use Scala Spark or PySpark for this assignment. PySpark may be easier)

# PROJECT 3: Docker

This project is fairly straightforward. You’ll create a python flask app, containerize it using Docker, and deploy it on GCP. You can experiment with setting up a docker image at <https://docker-curriculum.com/> (please add Werkzeug==2.2.2 to your requirements.txt file before calling docker build). Then mosey over to <https://cloud.google.com/build/docs/build-push-docker-image> to deploy the container on GCP. Use this to test out the process but after that, you must build your own flask app that is more substantial (more linked pages) than the example! Submit a link to your running web app on GCP, the flask code, and your Dockerfile.

# PROJECT 4: Hadoop

Install hadoop <https://hadoop.apache.org/> (my understanding is that hadoop only works with Java 8 or 11. If you’re using a later version of Java, you might want to skip this project! Install it locally rather than on GCP (complicated!). Then, using the material at <https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html>, and the tutorial at <https://www.tutorialspoint.com/hadoop/index.htm> (unfortunately for a earlier hadoop version but most of it should be fine), load the 311 data file (if you run into space limitations, extract a few 1000 lines from the file and use that) onto HDFS. Then, use hadoop mapreduce to find the average time by agency. Submit your hadoop code and appropriate screenshots.

# PROJECT 5: White paper

For any of the four topics above, or for any cloud related topic that is of interest to you, write a 4-5 page paper that outlines the what, the why, and the how (without code). Imagine you’re convincing senior management to buy the product - that’s what your paper should focus on. This project is for those of you who’ve had it with coding and don’t want to do that anymore!

# PROJECT 6: ??

If you have an idea of your own, run it by me and we’ll figure something out.