Homework 5

Please complete all exercises in Jupyter notebook and turned in a knitted PDF with output included. Be sure to tune your Spark code with appropriate configuration parameters. During the execution of your code be sure to use the Spark UI to track your jobs and use available information from the UI to help you tune your jobs. If you don’t need all of the available memory on your PC to run the job be sure to tune it down.

1. Use Spark to read in the airport-codes-na and departuredelays datasets.
2. Join the airport codes dataset to the departuredelays dataset.
3. What type of Spark join would be best for this? Do you need to set any configuration parameters to complete the join?
4. What would a logical partition for this dataset be? You can use methods like aggregation and count to determine if there are any logical partitions.
5. Write the combined data back to your data directory using the partition key you determined in the previous step.
   1. Write the data in a directory using parquet.
   2. Write the data in a directory using ORC.
   3. What are the benefits/disadvantages of using ORC or Parquet for reading/writing data?
6. What if any configuration changes did you set to tune your job?
   1. As this process scales up, what parameters do you think you would need to optimize your job?