CS 410/510 Cloud and Cluster Data Management

Spring 2018 Quarter

Assignment 4 Spark in the Cloud: PySpark on GCP Dataproc

Due: Thursday, 17 May 2018 at the beginning of class

You should do this assignment individually. You should only talk to the instructor and TA about this assignment. You may also post questions and comments to the course mailing list on Piazza.

Please turn in your completed assignments on paper. Put your last name, first name, the assignment number in that order in the first line of your assignment.

This assignment involves running PySpark jobs on Google Cloud Platform's Dataproc service. Google has donated GCP credit for use by the students of the course. Please refer to the MapReduce example posted by the TA for information about getting your account and using the GCP dashboard.

Please remember to stop or delete Dataproc clusters when you are not using them, so you don't use up your GCP credits.

CS410 students do Parts I and II; CS510 students do all parts.

Part I: Running the Word-Count Job (30 points)

There are instructions on Piazza for running a PySpark word-count example over the Shakespeare dataset on GCP. Run that example on a Dataproc cluster.

For this part, turn in

1. The first page of preview rows from your BigQuery output table.

Part II: Running on Your Own Data Files (40 points)

Modify the PySpark example to run the word-count job on a different input of your choice. You will have to modify the example in several ways. For one, you'll need to specify a different input dataset. You will likely need to modify the Spark statements themselves to use a different-named column in your dataset and to set the initial count

to 1. Note: You don't have to count "words" – it could be names, zip codes, URLs, etc., depending on the datasets. Also, be careful of datasets with NULL values in the columns you are accessing.

For this part, turn in

1. The name of the input dataset you used and a brief description of its content, particularly the columns you access.
2. A listing of your modified PySpark program.
3. The first page of preview rows from your BigQuery output table.

Part III: Different Analysis (30 points, CS 510 only)

Modify the PySpark example to compute a different analysis, which should include at least one Spark transformation not in the original code. You may use whichever dataset you wish.

For this part, turn in

1. The name of the input dataset you used and a brief description of its content, particularly the columns you access.
2. An explanation of what your analysis is doing.
3. A listing of your modified PySpark program.
4. The first page of preview rows from your BigQuery output table.

Note: Please share your experiences on Piazza. In particular, if you figure out how to solve a particular problem on your own, please share what you learned.