Homework-2: SparkSQL and SQL on PostgreSQL

Systems and Toolchains for AI Engineers

Deadline: September 12th, 11:59PM ET

In this homework, you will be working with SQL and SparkSQL, specifically with the variant `postgres`. You will do your work on a `postgres` database and cut-and-paste your answers into a Jupyter Notebook or SQL files (files you create that end with .sql extension).

Submit the HW in one Jupyter notebook. If you need to include scripts outside Jupyter for a given question, note it in your notebook. If you need to include output or screenshots, you may include them in your notebook or attach them as separate image files using the naming convention qx.png (where x is the question number). If you have more than one image for a given question, name the files as qx_1.png, qx_2.png, etc.

A sample structure of your notebook is shown below

Homework-2 (Sample Example Output to Show Doc Organization)

Q1

Refer to q1.sql and q1.png

Q2

```
from pyspark.sql.functions import monotonically_increasing_id


### Some code here
## For output, refer to q2.png
```

If the output has many rows, it is sufficient to take a screenshot of the first five, unless otherwise instructed.

You will use GitHub classroom to create your GitHub repository. The URL for this assignment's GitHub classroom is: https://classroom.github.com/a/SMbcXdGA

Don't create a public repository in your own GitHub account. Use the URL above to create the GitHub repository for this assignment. It includes a starter ReadMe file.

On Canvas, only submit a URL to your GitHub repository.

Not following the submission guidelines may result in grading penalties.

Question 1

(10 points)

Create a table in Postgres DB to store the Google news records from RSS Feed. The table should include the following columns:

- Title
- Link
- PubDate
- Description
- Source

Here is a sample record that will be inserted in your table.


```
<title>Local case highlights technology's use in solving crime - WKBN.com</title>
<link>https://news.google.com/rss/articles/CBMipwFBVV95cUxPOE9PcTRHTVM5dDBJQXNvSEZpUkF4dENjelhJU29FR294SGgxamxBVkRpdkRfWjM2WVZaQU9JellIQkNpSGRyaDh2UUlVd2NlU0pLSmktYWJ2bnBic1FVZVNtQWNPcUotRkY0bjRtYVNhd3BNRkRVRFJHWHpSSoc=5</link>
<guid isPermaLink="false">CBMipwFBVV95cUxPOE9PcTRHTVM5dDBJQXNvSEZpUkF4dENjelhJU29FR294SGgxamxBVkRpdkRfWjM2WVZaQU9JellIQkNpSGRyaDh2UUlVd2NlU0pLSmktYWJ2bnBic1FVZVNtQWNPcUotRkY0bjRtYVNhd3BNRkRVRFJHWHpSSkh0S2tHS2hBdzFzblhINGVT</guid>
<pubDate>Mon, 02 Sep 2024 21:55:45 GMT</pubDate>
<description><a href="https://news.google.com/rss/articles/CBMipwFBVV95cUxPOE9PcTRHTVM5dDBJQXNvSEZpUkF4dENjelhJU29FR294SGgxamxBVkRpdkRfWjM2WVZaQU9JellIQkNpSGRyaDh2UUlVd2NlU0pLSmktYWJ2bnBic1FVZVNtQWNPcUotRkY0bjRtYVNhd3BNRkRVRFJHWHpSSoc=5" target="_blank">Local case highlights technology's use in solving crime</a>  <font color="#6f6f6f">WKBN.com</font></description>
<source url="https://www.wkbn.com">WKBN.com</source>
```


Design your table with appropriate unique identifier and choose the appropriate column data types. Your table must be defined in its unique schema with the name "news".

For this question, submit the following:

- .sql file with the script that you used to create the table.
- Screenshot of the column data types when you use the proper PostgreSQL command to display the table description.
    - Your screenshot must include your name. Open Text editor application and write your name then take a screenshot of the entire screen including the requested information above. Don't photoshop your name on the image.

Question 2

(10 points)

Populate your table in q1 with the data in this Google news feed URL:
https://news.google.com/rss/search?q=technology&hl=en-US&gl=US&ceid=US:en

There could be large amounts of data in this URL (or a similar URL). So, choose an appropriate technique to populate the table that enables scalability.

For this question, submit the following:

- The code you used to populate the table.
- Screenshot of fetching random 5 records from your PostgreSQL DB table. Use PgAdmin to fetch the records.
    - Your screenshot must include your name. Open Text editor application and write your name then take a screenshot of the entire screen including the requested information above. Don't photoshop your name on the image.

Question 3

(15 points)

Write a function to find all the news that were published in the last 24 hours. Make sure 1) your function is scalable to execute on large amounts of data and 2) your function is dynamic to execute on any day.

For this question, submit the following:

- Your code.
- Screenshot of the output of the function. For this requirement, you may elect to display the output in your notebook or attach it as a separate image.
    - If you attach the output as a separate image, your screenshot must include your name. Open Text editor application and write your name then take a screenshot of the entire screen including the requested information above. Don't photoshop your name on the image.

Question 4

(15 points)

You decided to extend your table to include non-technology related news as well.
Update your table to add a new column called "category" and pre-populate it with the value of "technology" for existing records. Populate the table with the following feeds:

- Business: https://news.google.com/rss/search?q=business&hl=en-US&gl=US&ceid=US:en
- Sports: https://news.google.com/rss/search?q=sports&hl=en-US&gl=US&ceid=US:en

For this question, submit the following:

- Your SQL script in a separate SQL file.
- Your code to populate the tables.
- Query all the distinct categories in your notebook. Print the output or attach it as a separate image.

- If you attach the output as a separate image, your screenshot must include your name. Open Text editor application and write your name then take a screenshot of the entire screen including the requested information above. Don't photoshop your name on the image.

Question 5

(15 points)

Delete all the records that include the term "NFL" in their title. Use a scalable approach that can delete large amounts of records in a large DB table and can be executed smoothly in real-time.

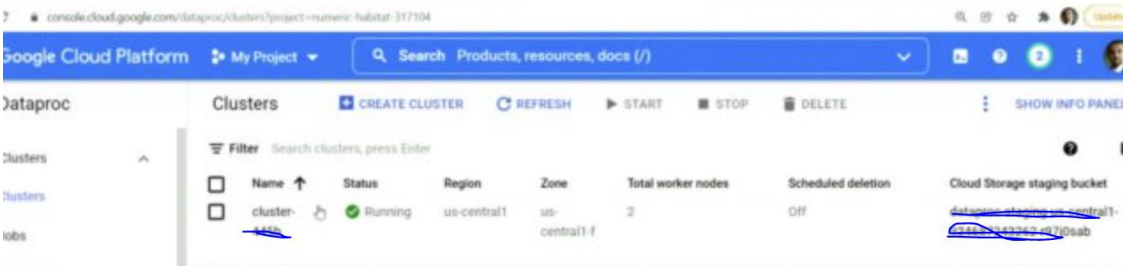
For this question, submit the following:

- Your code.
- Screenshot of the output of a query to select all the records that include the term "NFL" after executing your code. Print the output or attach it as a separate image.
    - If you attach the output as a separate image, your screenshot must include your name. Open Text editor application and write your name then take a screenshot of the entire screen including the requested information above. Don't photoshop your name on the image.

Question 6

(15 points)

Create your first cluster on Google Cloud and Submit a screenshot of your cluster while it's "Running". Name your cluster after your Andrew ID (and add any other characters).
Your screenshot should look like the following (except for the cluster name):



Don't forget to disable your Cloud billing.