adding homework
Mohamed Hamza authored 4 weeks ago

32f8cf69

README.md 5.71 KiB

Big Data Homework - "Alexa, can you handle big data?"

Before You Begin

1. Create a new repository for this project called `big-data-challenge`. **Do not add this homework to an existing repository**.
2. Clone the new repository to your computer.
3. Inside your local git repository, create a directory for the level of challenge Challenge you choose. Use folder names corresponding to the challenges: **level-1** or **level-2**.
4. Download a Google Colab Notebook as a `ipynb` file and add it to this folder. This will be the main script to run for analysis. Be sure to also add any SQL queries you used to a `.sql` file and add it to your repo.
5. Push the above changes to GitHub or GitLab.

Background

In this assignment you will put your ETL skills to the test. Many of Amazon's shoppers depend on product reviews to make a purchase. Amazon makes these datasets publicly available. However, they are quite large and can exceed the capacity of local machines to handle. One dataset alone contains over 1.5 million rows; with over 40 datasets, this can be quite taxing on the average local computer. Your first goal for this assignment will be to perform the ETL process completely in the cloud and upload a DataFrame to an RDS instance. The second goal will be to use PySpark or SQL to perform a statistical analysis of selected data.

There are two levels to this homework assignment. The second level is optional but highly recommended.

1. Create DataFrames to match production-ready tables from two big Amazon customer review datasets.
2. Analyze whether reviews from Amazon's Vine program are trustworthy.

Instructions

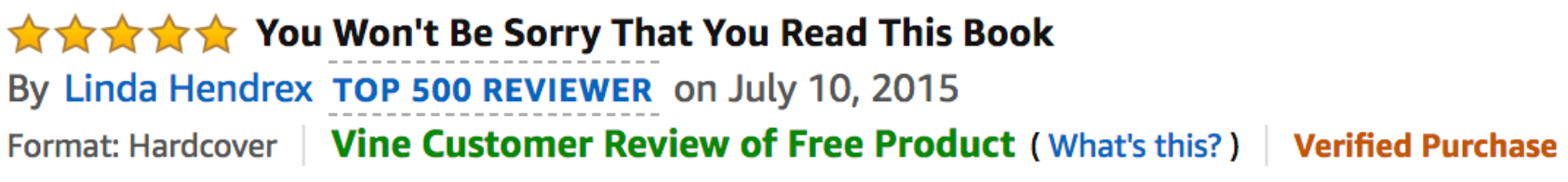
Level 1

- Use the furnished schema to create tables in your RDS database.
- Create two separate Google Colab notebooks and **extract** any two datasets from the list at review dataset, one into each notebook.

  Note: It is possible to ETL both data sources in a single notebook, but due to the large data sizes, it will be easier to work with these S3 data sources in two separate Colab notebooks.
- Be sure to handle the header correctly. If you read the file without the header parameter, you may find that the column headers are included in the table rows.
- For each notebook (one dataset per notebook), complete the following:
  - Count the number of records (rows) in the dataset.
  - **Transform** the dataset to fit the tables in the schema file. Be sure the DataFrames match in data type and in column name.
  - **Load** the DataFrames that correspond to tables into an RDS instance. **Note:** This process can take up to 10 minutes for each. Be sure that everything is correct before uploading.

Level 2

In Amazon's Vine program, reviewers receive free products in exchange for reviews.



Amazon has several policies to reduce the bias of its Vine reviews: https://www.amazon.com/gp/vine/help?ie=UTF8.

But are Vine reviews truly trustworthy? Your task is to investigate whether Vine reviews are free of bias. Use either PySpark or—for an extra challenge—SQL to analyze the data.

- If you choose to use SQL, first use Spark on Colab to extract and transform the data and load it into a SQL table on your RDS account. Perform your analysis with SQL queries on RDS.
- While there are no hard requirements for the analysis, consider steps you can take to reduce noisy data, e.g., filtering for reviews that meet a certain number of helpful votes, total votes, or both.
- Submit a summary of your findings and analysis.

Hints and Considerations

- Be sure that to start each notebook with following code in the first cell and update the Spark version.

```
import os
# Find the latest version of spark 3.0  from http://www-us.apache.org/dist/spark/ and enter as the spark version
# For example:
# spark_version = 'spark-3.0.1'
spark_version = 'spark-3.<spark version>'
os.environ['SPARK_VERSION']=spark_version

# Install Spark and Java
!apt-get update
!apt-get install openjdk-11-jdk-headless -qq > /dev/null
!wget -q http://www-us.apache.org/dist/spark/$SPARK_VERSION/$SPARK_VERSION-bin-hadoop2.7.tgz
!tar xf $SPARK_VERSION-bin-hadoop2.7.tgz
!pip install -q findspark

# Set Environment Variables
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["SPARK_HOME"] = f"/content/{spark_version}-bin-hadoop2.7"

# Start a SparkSession
import findspark
findspark.init()
```

- For connection to Postgres run the following code in the next cell.

```
!wget https://jdbc.postgresql.org/download/postgresql-42.2.9.jar
```

Submission

- **Important** be sure to clean up all your AWS resources. Consult the AWS cleanup guide and AWS check billing guide as reference.
- Download you Google Colab notebooks as `.ipynb` files and upload those to GitHub.
- Copy your SQL queries into `.sql` files and upload to GitHub.
- **Important:** Do not upload notebooks that contain your RDS password and endpoint. Be sure to delete them before making your notebook public!
- Ensure your repository has regular commits and a thorough README.md file

Rubric

Unit 22 Rubric - Big Data Homework - "Alexa, can you handle big data?"

References

Amazon customer Reviews Dataset. (n.d.). Retrieved April 08, 2021, from: https://s3.amazonaws.com/amazon-reviews-pds/readme.html

---