

Module 22 Challenge

Start Assignment

Due Wednesday by 11:59pm **Points** 100 **Submitting** a text entry box or a website url

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. They are quite large and can exceed the capacity of local machines. One dataset alone contains over 1.5 million rows; with over 40 datasets, data analysis can be very demanding on the average local computer. Your first goal 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.

This Challenge contains two parts. Part 1 is required. Part 2 is optional but highly recommended to strengthen your new skills.

- **Part 1:** Extract two Amazon customer review datasets, transform each dataset into four DataFrames, and load the DataFrames into an RDS instance.
- **Part 2:** Extract two Amazon customer review datasets and use either SQL or PySpark to analyze whether reviews from Amazon's Vine program are trustworthy.

Before You Begin

1. Create a new repository for this project called "Big-Data-ETL". **Do not add this homework to an existing repository.**
2. Clone the new repository to your computer.
3. Inside your local Git repository, create two folders, "part-1" and "part-2", corresponding to the two parts. If you are not planning on doing "part-2" then create the "part-1" folder only.



Files

Download the following files to help you get started:

Module 22 Challenge files  https://static.bc-edx.com/data/dl-1-1/m22/lms/starter/Starter_Code_v1.zip

Instructions

Part 1

1. Upload the `part_one_starter_code.ipynb` into Google Colab and create a duplicate of this file.
2. Explore the **Amazon Reviews**  <https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt> datasets and pick two datasets to perform ETL.
3. Rename each `part_one_starter_code.ipynb` file according to the dataset you are using. For example, if you are going to use the **Video Game reviews**  https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video_Games_v1_00.tsv.gz file, rename file, `part_one_video_games.ipynb`. Repeat the process for the duplicate file you created in Step 2.

Extract the Data

1. Read in each dataset using the correct `header` and `sep` parameters.
2. Get the number of rows in the dataset.

Transform the Data

For each dataset use the `schema.sql` file located in the Resources folder of the `Starter_Code.zip` file to create the four DataFrames as follows:

1. Create the "review_id_df" DataFrame with the appropriate columns and data types.
2. Create the "products_df" DataFrame that drops the duplicates in the "product_id" and "product_title" columns.
3. Create the "customers_df" DataFrame that groups the data on the "customer_id" by the number of times a customer reviewed a product.
4. Create the "vine_df" DataFrame that has the "review_id", "star_rating", "helpful_votes", "total_votes", and "vine" columns.

Load the Data into an RDS Instance


Export each DataFrame into the RDS instance to create four tables for each dataset.


NOTE

This process can take up to 10 minutes for each. Ensure that everything is correct before uploading.

Part 2

Recall that this part is completely optional; you can complete it as a way to challenge yourself and boost your new skills.

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)  (<https://www.amazon.com/gp/vine/help?ie=UTF8>).

 **You Won't Be Sorry That You Read This Book**
By **Linda Hendrex** **TOP 500 REVIEWER** on July 10, 2015
Format: Hardcover | **Vine Customer Review of Free Product (What's this?)** | **Verified Purchase**

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

Requirements

These requirements are for Part 1 only, as Part 2 is optional.

Extract (35 points)

- Uses PySpark to connect to and load the AWS datasets into DataFrames. (10 points)
- The correct parameters are used to read in the data into a DataFrame. (15 points)
- The first 20 rows of each DataFrame is displayed. (5 points)

- The number of rows for each DataFrame is displayed. (5 points)

Transform (45 points)

- The "review_id_df" DataFrame is created with the appropriate columns and data types. (15 points)
- The "products_df" DataFrame is created and the duplicate values are dropped. (10 points)
- The "customers_df" DataFrame is created and displays the number of times a customer reviewed a product grouped by the "customer_id". (10 points)
- The "vine_df" DataFrame is created that has the "review_id", "star_rating", "helpful_votes", "total_votes", and "vine" columns. (10 points)

Load (20 points)

- The four DataFrames for each dataset are successfully loaded into an RDS instance. (20 points)

Grading

This assignment will be evaluated against the requirements and assigned a grade according to the following table:

Grade	Points
A (+/-)	90+
B (+/-)	80–89
C (+/-)	70–79
D (+/-)	60–69
F (+/-)	< 60

Submission

- Download your Google Colab notebooks as `.ipynb` files and upload them into the "part-1" folder of your "Big-Data-ETL" Git repository.

IMPORTANT

Do not clear the outputs of your `.ipynb` files, and *do not* upload notebooks that contain your RDS password and endpoint. Delete these two items before making your notebook public!

- Ensure your repository has regular commits and a thorough `README.md` file to explain the ETL project.
- If you are doing "part-2" of this assignment, copy your SQL queries into `.sql` files and upload them into the "part-2" folder of your "Big-Data-ETL" Git repository.

IMPORTANT

Remember to closely monitor any AWS resources that you choose to use! It's crucial that you clean up and stop, or shut down any AWS resources to avoid accruing additional costs. Please refer to the `AWS_cleanup.pdf` and the `AWS_check_billing.pdf` files in the Resources folder of the `Starter_Code.zip` file. Or, you can download the


To submit your Challenge assignment, click Submit, and then provide the URL of your GitHub repository for grading.

NOTE

You are allowed to miss up to two Challenge assignments and still earn your certificate. If you complete all Challenge assignments, your lowest two grades will be dropped. If you wish to skip this assignment, click Next, and move on to the next Module.

Comments are disabled for graded submissions in BootCamp Spot. If you have questions about your feedback, please notify your instructional staff or your Student Success Manager. If you would like to resubmit your work for an additional review, you can use the Resubmit Assignment button to upload new links. You may resubmit up to three times for a total of four submissions.

References

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