Coding Challenge Outline

This notebook was created in Google Colab and includes all of the lines of code needed to install the necessary libraries and set up the environment.

## Step 1: Retrieve files from Amazon S3 bucket

Used boto3 to establish a connection to the bucket.

Explored the bucket contents to get the keys for the data files.

Retrieved each data file to be used and stored in a local variable.

Notes: If there were more files I would like to clean up this step to automate the data retrieval.

## Step 2: Convert the strings retrieved through the API into a usable JSON object

Initially coded this out for the first sample, then moved those processes into methods to be reused on the second sample.

Steps:

1. Decode the body into a JSON string
2. Split the JSON string on “/n” delimiter to get individual JSON requests
3. Loaded each string into a JSON object
4. Appended each JSON object to a list for iteration

Notes: There is a read JSON decoder Error that I would like to explore to find a way to handle those errors.

## Step 3 Pruning the JSON objects for only the requested data and aggregating.

1. Creating a list of dictionaries from the JSON objects returning only the required data (assumed the product\_id is a unique identifier and that key names were consistent throughout the document)
2. Converting the list of dictionaries into a Pyspark dataframe for querying and aggregation
3. Spark SQL for aggregation and grouping
4. Pyspark dataframe saved to JSON file for submission

Notes: With more time I would like to move more of the final process into methods to streamline the process for larger batches of data. I would also want to have an understanding of data that we would want in the future and build those fields into this process as well. I also believe it would be useful to bring in the catagory\_name and last\_updated as it adds more context to the data.

## Languages & Libraries:

python 3.8

spark 3.0.2

hadoop 2.7

boto3

openjdk-8

Findspark

os

json