Big Data Assignment README

This assignment involves processing and analyzing Amazon product metadata using various techniques, including data preprocessing, streaming data processing, and frequent itemset mining algorithms. The assignment consists of several components:

1. **Jupyter File (Amazon_Data_Processing.ipynb):**
    - This Jupyter Notebook file demonstrates how to process large Amazon product metadata in chunks using pandas and Python.
    - It showcases the process of loading data in chunks, cleaning text data, preprocessing features, and writing preprocessed data to a new JSON file.
    - The notebook also includes a bonus section demonstrating real-time batch processing of incoming data.
2. **Producer (producer.py):**
    - The producer script reads preprocessed Amazon product data from a JSON file and publishes it to a Kafka topic.
    - It utilizes the confluent_kafka library to interact with Kafka.
3. **Consumer (consumer.py):**
    - The consumer script subscribes to the Kafka topic where preprocessed data is published by the producer.
    - It consumes the data from the Kafka topic and performs any necessary processing or analysis.
4. **PCY Algorithm Consumer (consumer_pcy.py):**
    - This script implements the PCY (Park-Chen-Yu) algorithm for frequent itemset mining in a streaming context.

- It utilizes a sliding window approach to process streaming data and identify frequent itemsets.

5. **Apriori Algorithm Consumer (consumer_apriori.py):**

    - Similar to the PCY algorithm consumer, this script implements the Apriori algorithm for frequent itemset mining.

    - It also uses a sliding window approach to process streaming data and identify frequent itemsets.

Instructions for Running the Code

1. **Environment Setup:**

    - Make sure you have Python installed on your system.Install required Python packages using pip install -r requirements.txt.

2. **Running the Jupyter Notebook:**

    Open the Jupyter Notebook file (Amazon_Data_Processing.ipynb) using Jupyter Notebook or JupyterLab.Follow the instructions provided in the notebook to execute each code cell.

3. **Running the Producer and Consumers:**

    Ensure that Kafka is installed and running on your system.Start Kafka producer by running python producer.py.Start Kafka consumers by running python consumer.py, python consumer_pcy.py, and python consumer_apriori.py in separate terminals.

4. **Note:**

   Make sure to adjust Kafka broker configurations (bootstrap.servers) in the producer and consumers to match your Kafka setup.Ensure that input/output file paths and Kafka topic names are correctly configured in the scripts.

Feel free to reach out if you have any questions or need further assistance :)