Factored Datathon 2023

Data Sources

In preparation for the upcoming Datathon, we have two primary sources of data to work with, and it's essential to note that these are the only data sources available for this event.

The first source involves two **batch-format** tables containing valuable information and insights related to reviews for products that customers buy on Amazon. These batch tables are distributed in partitions and stored using JSON format files hosted in an Azure Data Lake Storage instance. Tables Included:

Amazon Product Reviews: This is a large crawl of product reviews from Amazon. This dataset contains 82.83 million unique reviews from around 20 million users.

Amazon Metadata: Contains the description and metadata for all products included in the dataset.

The second source involves Amazon Product Reviews, available in streaming mode, offering real-time data updates. The streaming topic continuously receives new data as it becomes available, enabling us to stay up-to-date with the latest developments.

The challenge for the Datathon is to effectively combine the data from both sources and use them as sources for analytics and machine learning purposes. Please take into consideration the objectives that are defined for the Data Engineering section to be sure that the solution that your team is proposing is compliant.

Integrating the batch and streaming tables allows you to understand review data better, leveraging static and dynamic data advantages. This combined approach will enable you to extract valuable patterns, trends, and correlations, enhancing our ability to develop innovative and impactful solutions during the hackathon.

Batch Processing Acces:

Data Stored in Azure Data Lake Storage (ADLS):

- Storage Account: safactoreddatathon
- Container: source-files
- Authentication Method: SAS Token
- SAS Token: sp=r&st=2023-07-21T22:27:46Z&se=2023-08-19T06:27:46Z&sv=2022-11-02&sr=c&sig=VF6y7LwGSmTHpKbOwGhy6DKUxn5HYZTK4wuvA22Q%2FWI%3D
- SAS URL:

  https://safactoreddatathon.blob.core.windows.net/source-files?sp=r&st=2023-07-21T22:27:46Z&se=2023-08-19T06:27:46Z&sv=2022-11-02&sr=c&sig=VF6y7LwGSmTHpKbOwGhy6DKUxn5HYZTK4wuvA22Q%2FWI%3D

- Objects:
  - source-fies/amazon_metadata
  - source-files/amazon_reviews

Documentation to extract ADLS information using SAS Tokens:

- https://docs.databricks.com/storage/azure-storage.html#language-SAS%C2%A0tokens
- https://pypi.org/project/azure-storage-file-datalake/

Streaming Process Access:

Data Stored in Azure Event Hub:

- Event Hub Namespace: factored-datathon
- Event Hub (Topic): factored_datathon_amazon_review
- Event Hub Listen Policy Key: sJJnyi8GGTBAa55jY89kacoT6hXAzWx2B+AEhCPEKYE=
- Event Hub Listen Policy Connection String:

  Endpoint=sb://factored-datathon.servicebus.windows.net/;SharedAccessKeyName=datathon_listener;SharedAccessKey=sJJnyi8GGTBAa55jY89kacoT6hXAzWx2B+AEhCPEKYE=;EntityPath=factored_datathon_amazon_review

*Note: Payload for topic factored_datathon_amazon_review have the same structure as source-files/amazon_reviews table for batch processing.

Documentation Reference to Read Event Hub Data:

- https://learn.microsoft.com/en-us/azure/event-hubs/event-hubs-python-get-started-send?tabs=passwordless%2Croles-azure-portal