Assignment

1. Create a Resource Group

2. Create the Following Resources

    2a. Databricks workspace

    2b. Storage Account with 3 blank containers named Landing, Staging and Reporting.

    2c. Download the datasets orders.csv and customers.csv from the downloadable section made available in the portal and upload it to the Landing container of the storage account.

3. Launch the Azure databricks workspace and create a single-node cluster inside the workspace which terminates on 30 mins of inactivity (Choose the Worker Node type to be Standard_DS3_v2 of configuration 14GB RAM and 4 Cores)

4. Create a Notebook in the Databricks workspace and execute the following
    4a. Mount the Azure Storage to DBFS. Mount all the three containers.

    4b. Create a Spark dataframe for orders data and another Spark dataframe for customers data

    Note : Datasets don't have the schema inbuilt, therefore enforce the schema as provided in Dataframe-Schema.txt (The schema text file will be available in the downloadable section of the current week)

    4c. Write a spark transformation on orders dataframe which adds 2 new columns order_year and order_month by extracting that from orders_date.

    4d. Write back the transformed orders and customers in a suitable format that will optimize storage and subsequent query performance

    Hint: Most of the queries that are run are trying to fetch the order details of a customer based on order_year, order_status and state (from customers data). Thus, accordingly partitioning can be used.

5. Now, again create spark dataframes for customers and orders from the data present in the staging container.

5a. Write a query on these dataframes to get details of the customer. Fetch these columns - ("order_id","order_date","order_customer_id","order_month","order_year","order_status","customer_fname","customer_lname","customer_city","customer_zipcode","customer_state"). We can combine the details of both the tables. Try to do an optimized join.

5b. Write back the results of the above query into the reporting container of the storage container (Before writing, partition it based on city, state, status and order_year)

6. Create a temporary SQL view containing from the data in the reporting container, and it should have the following details:

customer_name (concatenation of first_name and last_name separated by space),

city

order_date

status

state

order_year

order_mo

nth

of all customers.

6a. This view created has to be used for multiple operations further and thus for performance benefits, it is better to cache this view.

6b. Write a SQL query on the view created above to find names of all the customers in state=TX, status='COMPLETE' and order_year=2014.

6c. Parameterize the above query to take inputs dynamically from the widgets and display the same results.

Hint: Create 3 widgets in the Notebook - State(drop down), Status(drop down), order_year(text box)

6d. Add another widget of order_month and write a query to calculate the total number of customers based on the values selected in the above 4 widgets.