Assignment

All these activities have to be performed on Azure Cloud.

1. Perform the following pre-work

   a. Create a Storage Account in the azure portal
   b. Create and launch a Databricks Workspace in the azure portal
   c. Create a Single Node Cluster (Ideally Standard F4) in the databricks workspace.

2. Choose any 2 datasets of your choice and upload the datasets into the Storage Account created in the previous step. The files have to uploaded into a container named “week15inputdatasets”.

3. Once the Cluster is deployed, create a Notebook and execute the following

   a. Create a mount point /mnt/week15assignmentdb to access the files in the container - week15inputdatasets
   b. Create Dataframes by reading the data present in storage account through the mount point created in the previous step

4. Create a Database and create delta tables on the data stored in the Storage account

   a. Create Spark tables in Parquet format
   b. Create Spark tables in Delta format
   c. How is the Parquet format structuring of files different from that of Delta format? Give a detailed explanation with appropriate examples and diagrams.
   d. Check on which of the tables, the following query - describe history <table-name> gets executed successfully and why?

5. Create a delta table in a single step while writing the data to the dataframe using saveAsTable option

6. Insert the data into the Delta tables using the following 3 approaches

a. Insert
b. Append
c. Copy

7. Depict how Schema mis-match is handled in Delta format. Explain by considering a usecase on the data present in the storage account.

8. How does Delta format support Schema Evolution? Explain by considering a usecase on the data present in the storage account.

9. Depict the internal working of update and delete operation by updating and deleting records of the data present in the storage account.

10. Apply NOT NULL and CHECK constraints on the data and demonstrate the behaviour when data violating the constraints are inserted into the delta table.

11. There have been several changes being made to the table. Say you are required to present the original data without any changes, restore the table to its first version.

12. Make sure to delete the resources that you have created.

Note:

- There are no restrictions with the Datasets that are used for the assignment. You can feel free to choose a dataset of your choice and explore (you could pick datasets from kaggle / use sample datasets provided by Databricks / download the datasets from external lab)
- You would be executing the complete assignment in your Azure Databricks account.

Process to Submit the Assignment -

You need to create a Google Document consisting of answers to all the above questions. Name the Google Document as **yourname_week15_assignment**
Please upload your solution by filling the following form - https://forms.gle/fr31BhgVvpeSU8Pk6

Top 5 answers will be selected and they will be compiled into a solution document and added to the Learning portal.