Assignment

All these activities have to be performed on Azure Cloud.

Important Note : In case your Azure Account Free Trial has ended, then you can use Databricks community edition for working on the assignment questions.

1. Perform the following pre-work

   a. Create a Storage account and create a mount point to access the files that will be saved in the storage account.
   b. Create and launch a Databricks Workspace in the azure portal
   c. Create a Single Node Cluster (Ideally Standard F4) in the databricks workspace.
   d. Choose any sample datasets of your choice present in the DBFS root - /databricks-datasets and add it to the storage account created in the above.

2. Create Spark tables

   a. In Parquet format

   b. In Delta Format

Present the differences between the two tables with relevant screenshots and explanations.

3. Illustrate the benefits of Delta Caching with the help of an example Dataset considered.

   a. Enable delta caching by setting the property
   b. Enable delta caching manually through command.

4. Demonstrate the commonly occurring Small File Problem

   a. Create a Database
   b. Load any sample csv file from the DBFS file system.

   c. Create a Delta Table with a large number of files (~500 files) for each partition using the repartition option.

5. Demonstrate the Compaction / Bin-Packing technique of overcoming the small file problem explained previously.

6. Explain with an example how data-skipping is achieved in Partitioning and Bucketing.

7. Use the Z-Ordering approach to achieve Data-skipping optimization on the dataset considered in the previous steps. (Provide screenshots of executed query along with explanations for query performance with Z-Ordering and without Z-Ordering)

8. When is the VACUUM command used, its benefits and drawbacks? Demonstrate by applying it to the example dataset considered. (You can perform insert, update and delete operations to showcase the increasing transaction logs and how it would impact the resources over time.)

9. Never forget to delete all the resources created for practice.

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 use sample datasets provided by Databricks in the DBFS root **/databricks-datasets**)
- 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_week16_assignment**
Please upload your solution by filling the following form - https://forms.gle/zvEAceHumfydCxdKA

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