Storage and Files*

Distributed Data Processing Environments

Lab Guide 4

This session aims at converting, storing, and retrieving data files in the cloud. It uses the Free Tier[1] and should not consume credits.

Steps

1. Setup a Vagrant recipe to create a VM instance in the cloud.
   - Obtain the command from: `https://duckdb.org/install/?platform=linux&environment=cli`
   - Add the command to the Vagrantfile recipe.
   - Run the recipe.
   - Log in the VM instance with SSH.
   - (At this time, it might be useful to install DuckDB also on your laptop.)
2. Create a cloud storage bucket:
   - Go to Cloud Storage > Buckets and select "Create".
   - Select a unique name for your bucket.
   - Uncheck the enforce public access prevention option under access control.
   - Choose "Create".
   - Manually upload a CSV file to the bucket.
3. Create a key to access the bucket:
   - Go to Cloud Storage > Settings, then to Interoperability tab.
   - Create a key for a service account.
   - Choose the only service account.
   - Save the displayed keys for later.
4. Configure DuckDB to access the bucket:
   - Open DuckDB with: `duckdb test.db`
   - Create a secret with data obtained in the previous step and the `CREATE SECRET` statement.[2]
5. Obtain copies of the file in different formats and locations:
   - Load a DuckDB table from the cloud bucket with: `CREATE TABLE ... AS SELECT * FROM 'gs://...';`
   - Export table to instance storage as CSV[3] and Parquet.[4]

*Avoid AI tools in step 6.

[1] `https://cloud.google.com/free/docs/free-cloud-features#free-tier`
[2] `https://duckdb.org/docs/stable/guides/network_cloud_storage/gcs_import`
[3] `https://duckdb.org/docs/stable/guides/file_formats/csv_export`
[4] `https://duckdb.org/docs/stable/guides/file_formats/parquet_export`

- Save in the cloud as Parquet.[5]

6. Test the time it takes to execute a query in each of these formats and locations.
   - How does storage location impact performance?
   - How does file format impact performance?
7. At the end of the session:
   - Destroy the VM and cleanup with: `vagrant destroy`
   - Navigate to the Compute Engine > Instances page on Cloud Console and delete the working environment, if running in the cloud.
   - Navigate to the Cloud Storage > Buckets page on Cloud Console and delete the bucket.

Learning Outcomes Deploy data in instance disks and cloud object storage. Assess the performance of data processing operations. Relate storage type and file formats with processing performance.

[5] `https://duckdb.org/docs/stable/guides/network_cloud_storage/s3_export`