accenture

Learning and Knowledge Management

Cloud Dataflow

To Create a Dataflow Pipeline

- **Definition:** The Apache Beam SDK is an open source programming model for data pipelines. In Google Cloud, you can define a pipeline with an Apache Beam program and then use Dataflow to run your pipeline.

- **Purpose**: To set up the Python development environment for Dataflow (using the Apache Beam SDK for Python) and run an example Dataflow pipeline.

- Follow the link to complete the Lab
  - Open the url – https://console.cloud.google.com
  - Login with your Accenture credentials (e-mail id)
  - Follow the below steps to complete your lab

Steps:

A. To create a Cloud Storage Bucket

1. On the Navigation menu, click Cloud Storage.
2. Click Create bucket.
3. In the Create bucket dialog, specify the following attributes:
   a. Name – example-bucket-current-date
   b. Location type – Multi-region and select US
4. Click Create

B. Install the Apache Beam SDK for Python

1. To use a supported version of Python, trigger the below command –
   a. **docker run -it -e DEVSHELL_PROJECT_ID=$DEVSHELL_PROJECT_ID python:3.9 /bin/bash**
   b. Once the container is running, install the latest version of Apache Beam SDK - **pip install 'apache-beam[gcp]'==2.42.0**
   c. Ignore the warning messages, and run the given example(wordcount.py) locally by triggering the below command –
      a. **python -m apache_beam.examples.wordcount --output OUTPUT_FILE**

    d. List the files using **ls** command in the local directory and fetch the OUTPUT_FILE name
    e. Using the **cat** command, check the contents of the generated output file

C. To run an example Dataflow Pipeline remotely

1. Set the Bucket Environment variable –
    a. BUCKET=gs://<bucket name provided earlier>
2. Run the wordcount.py example remotely –

    a.
    ```
    python -m apache_beam.examples.wordcount --project
    $DEVSHELL_PROJECT_ID --runner DataflowRunner --staging_location
    $BUCKET/staging --temp_location $BUCKET/temp --output
    $BUCKET/results/output --region "filled in at lab start"
    ```

3. Check if the workers have started successfully.

D. Check if the Dataflow Job succeeded

1. Open the Navigation menu and click Dataflow from the list of services.
2. You should see your wordcount job with a **status** of **Running at first.**
3. Click on the name to watch the process.
4. When all the boxes are checked off, you can continue watching the logs in Cloud Shell.
5. The process is complete when the **status** is **Succeeded**.