Batch ETL Pipeline Implementation in GCP Using Dataflow, Cloud Storage, and Big Query

TASK-4
MALLEESWARI D

INTRODUCTION:

First, I read and understood the concepts of ETL, batch and streaming processes, and data parsing. Based on this foundational knowledge, I proceeded to implement the ETL process in a real-time GCP environment.

OBJECTIVE:

To build a batch ETL pipeline in Google Cloud Platform (GCP) that reads a CSV file from Cloud Storage, processes it using Dataflow (Apache Beam), and writes the transformed data into a BigQuery dataset named 'lake'.

STEP-BY-STEP IMPLEMENTATION GUIDE:

1. Setup GCP Project:

• Select or create a project (e.g., 'careful-synapse-461006-b5').

2. Upload Dataset to Cloud Storage:

• Dataset used: 'Medicaldataset.csv'
• Bucket name: 'malles_bucket'
• Upload file to the bucket using Cloud Console or gsutil:
gsutil cp Medicaldataset.csv gs://malles_bucket/

3. Create BigQuery Dataset:

• Open Cloud Shell.
• Run:
bq mk lake
• This creates the destination dataset for structured/clean data.

4. Develop ETL Script (Python - Apache Beam):

• Use Apache Beam with Dataflow runner.
• ETL steps:
- Read CSV from Cloud Storage

- Parse and clean each row
- Write to BigQuery table 'lake.medicalrecord'

5. Launch Dataflow Job:

• Submit the job via Cloud Shell using the Python script:
python batch_etl.py --input gs://malles_bucket/Medicaldataset.csv --output lake.medicalrecord

6. Monitor ETL Process in Dataflow Console:

• Go to Dataflow Console: https://console.cloud.google.com/dataflow
• View job status: Succeeded / Failed
• Analyze job graph to see stages (Read, Parse, Write)
• Check logs for errors

7. Verify Data in BigQuery:

• Go to BigQuery Console: https://console.cloud.google.com/bigquery
• Expand project → Dataset 'lake'
• Click table 'medicalrecord' → View 'Preview' tab
• Run SQL to check data:
SELECT * FROM lake.medicalrecord LIMIT 10;

8. Summary Explanation for Mentor:

"I implemented a batch ETL process that reads a raw medical CSV file from Cloud Storage, processes it using Dataflow, and writes the structured data to a BigQuery dataset named 'lake'. This clean dataset is now ready for analysis or reporting."

About 'lake' Dataset:

The 'lake' dataset in BigQuery acts as a centralized and clean zone for storing ETL output. It's a foundational layer in the data pipeline, ready for downstream analytics or dashboards.