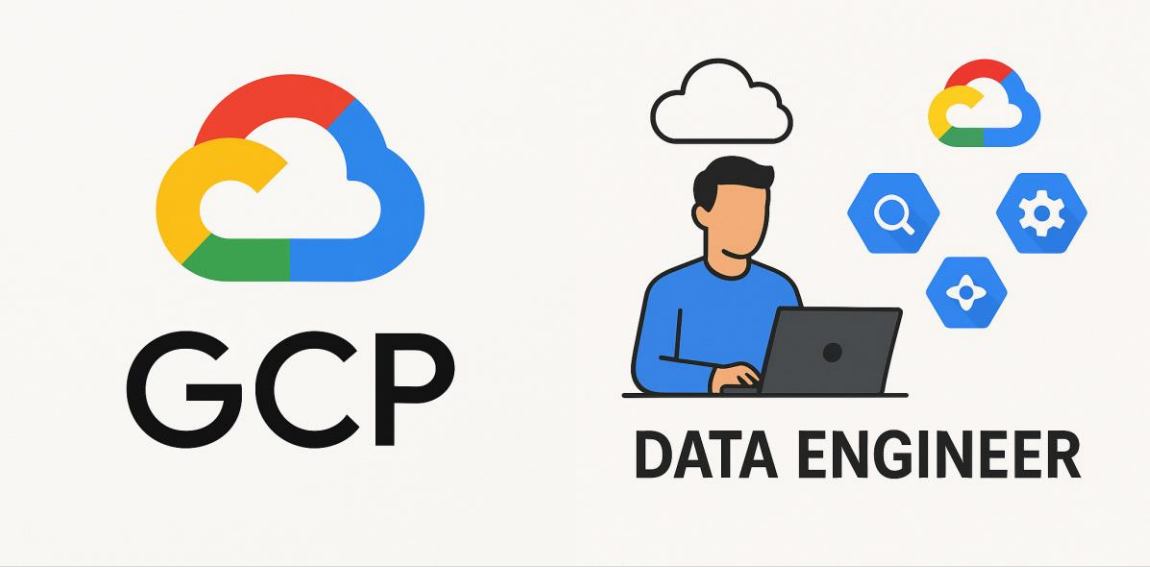
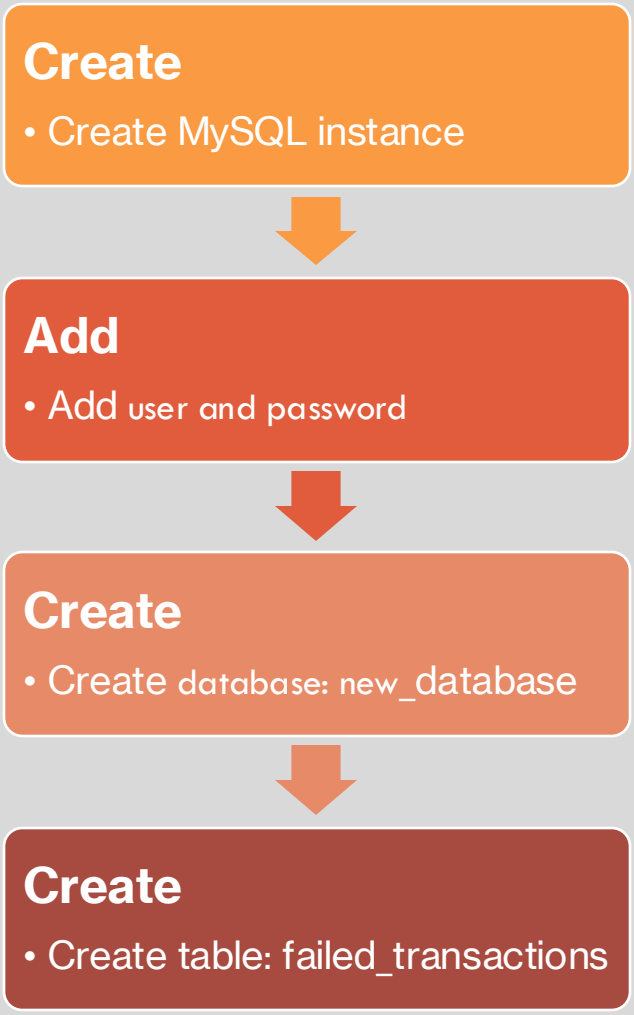


PROJECT OVERVIEW

FAILED BANKING TRANSACTION ANALYSIS ON GCP USING DATAPROC, CLOUD SQL, AND BIGQUERY.

Setting up MySQL instance

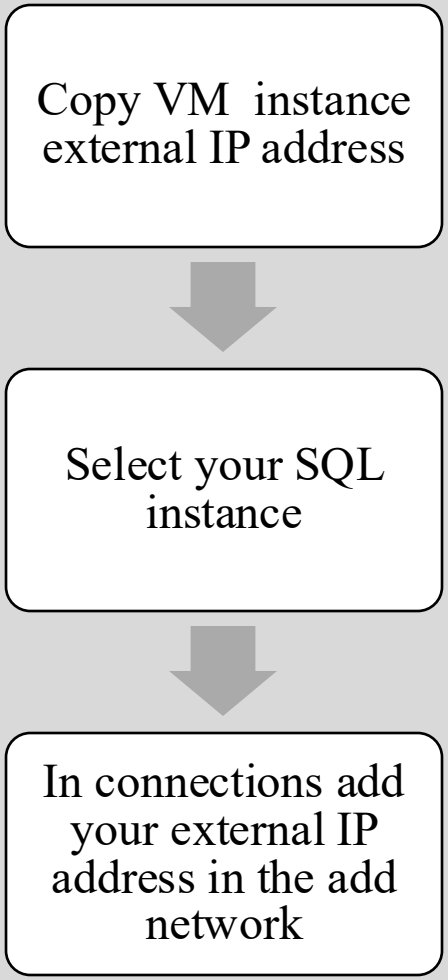


For Creating a Dataproc Cluster

Open Cloud Shell or Console:

gcloud dataproc clusters create training-cluster --region=us-central1 --zone=us-central1-a --single-node --master-machine-type=n1-standard-2 --image-version=2.0-debian10 --enable-component-gateway --optional-components=jupyter

Connect Dataproc to MySQL



INGESTING THE DATA INTO THE GCS BUCKET

Create a bucket in the google cloud storage:

gsutil mb gs://kamalkumar-bucket/

Upload the CSV files to the bucket:

gsutil cp "E:/revature/p1bankproject/new_dataset/*.csv" gs://kamalkumar-bucket/transactions/

Upload Python scripts to the bucket:

gsutil cp "E:/revature/p1banking/cleancode.py" gs://kamalkumar-sbucket/scripts/

gsutil cp "E:/revature/p1banking/failedtrans.py" gs://kamalkumar-bucket/scripts/

RUN PYSPARK JOB ON DATAPROC CLUSTER

Submitting a PySpark job for cleaning the dataset using a Python script

Command:

gcloud dataproc jobs submit pyspark gs://kamalkumar-bucket/scripts/cleancode.py --cluster=training-cluster --region=us-central1 --jars=gs://kamalkumar-bucket/jars/mysql-connector-j-8.0.33.jar --properties=spark.driver.memory=4g,spark.executor.memory=4g

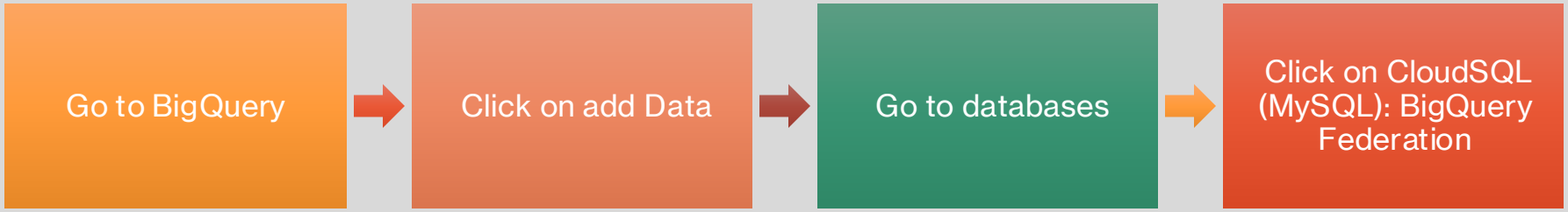
Submitting a PySpark job for processing failed transactions using the failedtrans.py script.

Command:

gcloud dataproc jobs submit pyspark gs://kamalkumar-bucket/scripts/failedtrans.py --cluster=training-cluster --region=us-central1 --jars=gs://kamalkumar-bucket/jars/mysql-connector-j-8.0.33.jar -properties=spark.driver.memory=4g,spark.executor.memory=4g

VERIFY DATA IN CLOUD SQL

- Verify failed_transactions table and data insertion



Federated Connection: BigQuery to MySQL

CONFIGURE IAM PERMISSIONS

- Copy service account from external connection
- Add it to IAM with 'Cloud SQL Client' role

BigQuery Analysis and Visualization

Query federated tables

Export results to CSV ,json or bigquery tables etc

Visualize data using bigquery charts or looker studio

CONCLUSION:

- Successfully integrated multiple GCP services for end-to-end data analysis
- Automated the data pipeline from ingestion to visualization
- Designed a scalable, secure, and efficient architecture
- Delivered actionable insights to support informed decision-making