Cheat Sheet for comprehensive Google Cloud Professional Data Engineer

Google Cloud Platform (GCP) Overview

- **Core Services**: Compute Engine, Kubernetes Engine (GKE), Cloud Storage, BigQuery, Cloud SQL, Pub/Sub, Dataflow, Dataproc, Bigtable, Cloud Spanner, Cloud Functions, Cloud Run.

- **Regions and Zones**:

  - **Regions**: Geographical areas (e.g., us-central1, europe-west1).

  - **Zones**: Specific data centers within regions (e.g., us-central1-a, us-central1-b).

Compute Services

Compute Engine

- **VM Instances**:

  - `gcloud compute instances create [INSTANCE_NAME]`

  - **Machine Types**: `n1-standard-1`, `n1-highmem-2`, etc.

  - **Boot Disks**: `pd-standard`, `pd-ssd`

  - **Custom Machine Types**: `--custom-cpu=4 --custom-memory=15`

- **Instance Groups**:

  - **Managed**: `gcloud compute instance-groups managed create`

  - **Unmanaged**: `gcloud compute instance-groups unmanaged create`

- **Autoscaling**:

  - `--max-num-replicas=10 --min-num-replicas=1`

  - **Metrics**: CPU utilization, load balancing capacity, etc.

Kubernetes Engine (GKE)

- **Clusters**:

  - `gcloud container clusters create [CLUSTER_NAME]`

  - **Node Pools**: `gcloud container node-pools create [POOL_NAME]`

- **Deployments**:

- `kubectl create deployment [DEPLOYMENT_NAME] --image=[IMAGE_NAME]`

- **Services**:

- `kubectl expose deployment [DEPLOYMENT_NAME] --type=LoadBalancer --port=80`

- **Autoscaling**:

  - **Cluster Autoscaler**: `--enable-autoscaling --min-nodes=1 --max-nodes=5`

  - **Horizontal Pod Autoscaler**: `kubectl autoscale deployment [DEPLOYMENT_NAME] --cpu-percent=50 --min=1 --max=10`

Storage Services

Cloud Storage

- **Buckets**:

- `gsutil mb gs://[BUCKET_NAME]/`

  - **Class**: `STANDARD`, `NEARLINE`, `COLDLINE`, `ARCHIVE`

- **Objects**:

- `gsutil cp [LOCAL_FILE] gs://[BUCKET_NAME]/`

  - **ACLs**: `gsutil acl ch -u [USER]:[PERMISSION] gs://[BUCKET_NAME]/`

- **Lifecycle Management**:

- `gsutil lifecycle set [LIFECYCLE_CONFIG_FILE] gs://[BUCKET_NAME]/`

Cloud SQL

- **Instances**:

- `gcloud sql instances create [INSTANCE_NAME]`

  - **Databases**: `gcloud sql databases create [DATABASE_NAME] --instance=[INSTANCE_NAME]`

- **Users**:

- `gcloud sql users set-password [USER_NAME] --instance=[INSTANCE_NAME] --password=[PASSWORD]`

- **Backups**:

- `gcloud sql backups create --instance=[INSTANCE_NAME]`

Data Processing Services

BigQuery

- **Datasets**:

  - `bq mk --dataset [PROJECT_ID]:[DATASET_NAME]`

- **Tables**:

  - `bq mk --table [PROJECT_ID]:[DATASET_NAME].[TABLE_NAME] [SCHEMA]`

- **Queries**:

  - `bq query --use_legacy_sql=false '[SQL_QUERY]'`

- **Loading Data**:

  - `bq load --source_format=[FORMAT] [PROJECT_ID]:[DATASET_NAME].[TABLE_NAME] [FILE_PATH] [SCHEMA]`

Dataflow

- **Templates**:

  - `gcloud dataflow jobs run [JOB_NAME] --gcs-location=[TEMPLATE_PATH] --parameters [PARAMETERS]`

- **Direct Runner**:

  - `mvn compile exec:java -Dexec.mainClass=[MAIN_CLASS]`

- **Streaming vs Batch**:

  - **Streaming**: `--streaming`

  - **Batch**: Default mode

Dataproc

- **Clusters**:

  - `gcloud dataproc clusters create [CLUSTER_NAME]`

- **Jobs**:

  - `gcloud dataproc jobs submit [JOB_TYPE] --cluster=[CLUSTER_NAME] -- [JOB_ARGS]`

- **Autoscaling**:

  - `--enable-autoscaling --max-workers=10 --min-workers=2`

Data Storage Services

Bigtable

- **Instances**:

- `cbt createinstance [INSTANCE_NAME] [DISPLAY_NAME] [CLUSTER_NAME] [ZONE] [NUM_NODES]`

- **Tables**:

- `cbt createtable [TABLE_NAME]`

- **Columns**:

- `cbt createfamily [TABLE_NAME] [FAMILY_NAME]`

- **Data Operations**:

- `cbt set [TABLE_NAME] [ROW_KEY] [FAMILY_NAME]:[COLUMN_QUALIFIER]=[VALUE]`
- `cbt read [TABLE_NAME]`

Cloud Spanner

- **Instances**:

- `gcloud spanner instances create [INSTANCE_NAME] --config=[CONFIG] --nodes=[NUM_NODES] --description=[DESCRIPTION]`

- **Databases**:

- `gcloud spanner databases create [DATABASE_NAME] --instance=[INSTANCE_NAME]`

- **Tables**:

- `gcloud spanner databases ddl update [DATABASE_NAME] --instance=[INSTANCE_NAME] --ddl='CREATE TABLE [TABLE_NAME] ([SCHEMA])'`

Serverless Services

Cloud Functions

- **Deployments**:

- `gcloud functions deploy [FUNCTION_NAME] --runtime [RUNTIME] --trigger-http`

- **Triggers**:

  - **HTTP**: `--trigger-http`

  - **Pub/Sub**: `--trigger-topic [TOPIC_NAME]`

- **Cloud Storage**: `--trigger-bucket [BUCKET_NAME]`

Cloud Run

- **Deployments**:

- `gcloud run deploy [SERVICE_NAME] --image [IMAGE_NAME] --platform managed`

- **Autoscaling**:

- `--min-instances=0 --max-instances=10`

- **Environment Variables**:

- `--set-env-vars [KEY]=[VALUE]`

Networking

VPC Networks

- **Creation**:

- `gcloud compute networks create [NETWORK_NAME] --subnet-mode=auto`

- **Subnets**:

- `gcloud compute networks subnets create [SUBNET_NAME] --network=[NETWORK_NAME] --range=[IP_RANGE]`

- **Firewalls**:

- `gcloud compute firewall-rules create [RULE_NAME] --network=[NETWORK_NAME] --allow=[PROTOCOLS]`

Cloud Load Balancing

- **Types**:

  - **HTTP(S)**: Global load balancing

  - **Network**: Regional load balancing

  - **Internal**: Internal load balancing

- **Creation**:

- `gcloud compute forwarding-rules create [RULE_NAME] --load-balancing-scheme=[SCHEME] --target-pool=[POOL_NAME]`

Monitoring and Logging

Cloud Monitoring

- **Dashboards**:

- `gcloud monitoring dashboards create --config=[CONFIG_FILE]`

- **Alerts**:

- `gcloud alpha monitoring policies create --policy-from-file=[POLICY_FILE]`

Cloud Logging

- **Logs**:

- `gcloud logging logs list`

- **Sinks**:

- `gcloud logging sinks create [SINK_NAME] [DESTINATION]`

- **Metrics**:

- `gcloud logging metrics create [METRIC_NAME] --description=[DESCRIPTION] --log-filter=[FILTER]`

Security and Identity

IAM

- **Roles**:

- `gcloud projects add-iam-policy-binding [PROJECT_ID] --member=[MEMBER] --role=[ROLE]`

- **Service Accounts**:

- `gcloud iam service-accounts create [SA_NAME]`
- `gcloud projects add-iam-policy-binding [PROJECT_ID] --member=serviceAccount:[SA_EMAIL] --role=[ROLE]`

Cloud KMS

- **Keyrings**:

- `gcloud kms keyrings create [KEYRING_NAME] --location=[LOCATION]`

- **Keys**:

- `gcloud kms keys create [KEY_NAME] --keyring=[KEYRING_NAME] --location=[LOCATION] --purpose=encryption`

- **Encryption/Decryption**:

- `gcloud kms encrypt --key=[KEY_NAME] --keyring=[KEYRING_NAME] --location=[LOCATION] --plaintext-file=[PLAINTEXT_FILE] --ciphertext-file=[CIPHERTEXT_FILE]`

Tips and Tricks

- **CLI Shortcuts**:

- `gcloud config set project [PROJECT_ID]`
- `gcloud auth login`
- `gcloud components update`

- **Cost Management**:

  - **Budgets**: `gcloud billing budgets create`

  - **Quotas**: `gcloud services quota list`

- **Troubleshooting**:

  - **Logs**: `gcloud logging read`

  - **Metrics**: `gcloud monitoring metrics list`

- **Best Practices**:

  - **Data Lifecycle**: Use appropriate storage classes.

  - **Security**: Enable VPC Service Controls.

  - **Scalability**: Use autoscaling where possible.

Examples

- **Create a VM**:

```
  gcloud compute instances create my-vm --machine-type=n1-standard-1 --
zone=us-central1-a
```

- **Deploy a Cloud Function**:

```
  gcloud functions deploy helloWorld --runtime nodejs14 --trigger-http
```

- **Query BigQuery**:

```
  bq query --use_legacy_sql=false 'SELECT * FROM `my_dataset.my_table` LIMIT 1000'
```

Additional Resources

- **Documentation**: [Google Cloud Documentation](https://cloud.google.com/docs)
- **Training**: [Google Cloud Training](https://cloud.google.com/training)
- **Community**: [Google Cloud Community](https://cloud.google.com/community)

This cheat sheet provides a comprehensive overview of essential features, commands, and best practices for a Google Cloud Professional Data Engineer. Use it as a quick reference guide for your daily tasks and projects.

By Ahmed Baheeg Khorshid

ver 1.0