Google Cloud Data Engineer Certificate Study Guide

Exam Overview

* Storage (20%)
  + GCS, Cloud SQL, DataStore, BigTable, BigQuery
* Big Data Processing (35%)
  + BigQuery, Dataflow, Dataproc, Datalab, Pub/Sub
* Machine Learning (18%)
  + ML APIs, TensorFlow
* Case Studies (15%)
* Others (Hadoop and Security about 12%)

**Google Cloud Storage (GCS)**

* Blob storage. Content not indexed.
* Virtually unlimited storage.
* Can have domain name buckets
* Can make requesters pay (ex. requester in different project)
* Pub/Sub can have notifications based on operations to buckets/objects
* Objects are immutable
* Can set Cache-Control metadata for frequently accessed objects
* Keep in mind compliance requirements when storing data in certain regions.
* **Storage Classes**
  + **Multi-regional**
    - Serving website content, interactive workloads, mobile game/gaming applications
    - Highest availability
    - Geo-redundant: Stores data in at least 2 regions separated by at least 100 miles within the multi-regional location of the bucket.
  + **Regional**
    - Storing data used by Compute Engine
    - Better performance for data-intensive computation
  + **Nearline**
    - Accessed once a month max
    - 30 day min. storage duration
    - Ex. Data backup, disaster recovery, archival storage
  + **Coldline**
    - Accessed once a year max
    - 90 day min. storage duration
    - Ex. Data stored for legal or regulatory reasons
* Storage classes can change, but the objects (files) within them retain their storage class.
* Not ideal for high volume read/write
* A way to store data that can be commonly used by Dataproc and Bigquery
* Versioning
  + Needs to be enabled
  + Things this enables:
    - List archived versions of an object
    - Restore live version of an object from an older state
    - Permanently delete an archived version
  + Archived versions retain ACLs and does not necessarily have same permissions as live version of object.
* **Encryption**
  + **Encryption at rest (Google-Managed Encryption Keys)**
    - Default (AES-256)
    - Use TLS or HTTPS to protect data as it travels over Internet
  + **Server-side encryption:**
    - Layers on top of default encryption
    - Occurs after GCS receives data, but before written to disk
      * **Customer-supplied encryption keys**
        + Provide key for each GCS operation
        + Key purged from servers after operation is complete
        + Stores only a cryptographic hash of key for future requests
        + Transfer Service, Dataflow, and Dataproc do not support this currently
        + Key rotation

Edit .boto config file

Encryption\_key = [NEW\_KEY]

Decryption\_key1 = [OLD\_KEY]

gsutil rewrite -k gs:://[BUCKET]/[OBJECT]

* + - * **Customer-managed encryption keys**
        + Generate and manage keys using Cloud Key Management Service (KMS)
        + KMS can be independent from the project that contains buckets (separation of duties)
        + Uses service accounts to encrypt/decrypt
        + Cloud SQL exports to GCS and Dataflow do not support this currently
  + **Client-side encryption:**
    - Occurs before data sent to GCS
    - GCS performs default encryption on it as well.
* No native directory support
  + Forward slashes have no special meaning
  + Performance of a native filesystem is not present.
* **Storage Transfer Service**
  + Transfers data from an online data source (Amazon S3, HTTP/HTTPS location, GCS bucket) to a data sink (always GCS bucket).
  + Use cases:
    - Backup data to GCS from other storage providers
    - Move data from one GCS bucket to another (enables availability to different groups of users or applications)
    - Periodically move data as part of a processing pipeline or analytical workflow
  + Schedule one-time transfer operations or recurring ones
  + Delete existing objects in the destination bucket if they don’t have a corresponding object in source
  + Delete source objects after transferring them
  + Schedule periodic synchronization from data source to data sink with advanced filters based on file creation data, file-name filters, and the times of day you prefer to import data.
  + **Transfer Service vs. Gsutil**
    - On premise data source : gsutil
    - Another cloud storage provider data source : Transfer Service

**Cloud SQL**

**DataStore**

**BigTable**

**BigQuery**

**Dataflow**

**Dataproc**

**Datalab**

**Pub/Sub**

**ML APIs**

**TensorFlow**

**Hadoop**

**Security**