Google Cloud

Associate Data Practitioner

Certification exam guide

An Associate Data Practitioner secures and manages data on Google Cloud. This individual has experience using Google Cloud data services for tasks like data ingestion, transformation, pipeline management, analysis, machine learning, and visualization. Candidates have a basic understanding of cloud computing concepts like infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS).

Section 1: Data Preparation and Ingestion (~30% of the exam)

1.1 Prepare and process data. Considerations include:

- Differentiate between different data manipulation methodologies (e.g., ETL, ELT, ETLT)
- Choose the appropriate data transfer tool (e.g., Storage Transfer Service, Transfer Appliance)
- Assess data quality
- Conduct data cleaning (e.g., Cloud Data Fusion, BigQuery, SQL, Dataflow)

1.2 Extract and load data into appropriate Google Cloud storage systems. Considerations include:

- Distinguish the format of the data (e.g., CSV, JSON, Apache Parquet, Apache Avro, structured database tables)
- Choose the appropriate extraction tool (e.g., Dataflow, BigQuery Data Transfer Service, Database Migration Service, Cloud Data Fusion)
- Select the appropriate storage solution (e.g., Cloud Storage, BigQuery, Cloud SQL, Firestore, Bigtable, Spanner, AlloyDB)
    - Choose the appropriate data storage location type (e.g., regional, dual-regional, multi-regional, zonal)
    - Classify use cases into having structured, unstructured, or semi-structured data requirements
- Load data into Google Cloud storage systems using the appropriate tool (e.g., gcloud and BQ CLI, Storage Transfer Service, BigQuery Data Transfer Service, client libraries)

Section 2: Data Analysis and Presentation (~27% of the exam)

2.1 Identify data trends, patterns, and insights by using BigQuery and Jupyter notebooks.

Considerations include:

- Define and execute SQL queries in BigQuery to generate reports and extract key insights
- Use Jupyter notebooks to analyze and visualize data (e.g., Colab Enterprise)
- Analyze data to answer business questions

2.2 Visualize data and create dashboards in Looker given business requirements.
Considerations include:

- Create, modify, and share dashboards to answer business questions
- Compare Looker and Looker Studio for different analytics use cases
- Manipulate simple LookML parameters to modify a data model

2.3 Define, train, evaluate, and use ML models. Considerations include:

- Identify ML use cases for developing models by using BigQuery ML and AutoML
- Use pretrained Google large language models (LLMs) using remote connection in BigQuery
- Plan a standard ML project (e.g., data collection, model training, model evaluation, prediction)
- Execute SQL to create, train, and evaluate models using BigQuery ML
- Perform inference using BigQuery ML models
- Organize models in Model Registry

Section 3: Data Pipeline Orchestration (~18% of the exam)

3.1 Design and implement simple data pipelines. Considerations include:

- Select a data transformation tool (e.g., Dataproc, Dataflow, Cloud Data Fusion, Cloud Composer, Dataform) based on business requirements
- Evaluate use cases for ELT and ETL
- Choose products required to implement basic transformation pipelines

3.2 Schedule, automate, and monitor basic data processing tasks. Considerations include:

- Create and manage scheduled queries (e.g., BigQuery, Cloud Scheduler, Cloud Composer)
- Monitor Dataflow pipeline progress using the Dataflow job UI
- Review and analyze logs in Cloud Logging and Cloud Monitoring
- Select a data orchestration solution (e.g., Cloud Composer, scheduled queries, Dataproc Workflow Templates, Workflows) based on business requirements
- Identify use cases for event-driven data ingestion from Pub/Sub to BigQuery

- Use Eventarc triggers in event-driven pipelines (Dataform, Dataflow, Cloud Functions, Cloud Run, Cloud Composer)

Section 4: Data Management (~25% of the exam)

4.1 Configure access control and governance. Considerations include:

- Establish the principles of least privileged access by using Identity and Access Management (IAM)
    - Differentiate between basic roles, predefined roles, and permissions for data services (e.g., BigQuery, Cloud Storage)
- Compare methods of access control for Cloud Storage (e.g., public or private access, uniform access)
- Determine when to share data using Analytics Hub

4.2 Configure lifecycle management. Considerations include:

- Determine the appropriate Cloud Storage classes based on the frequency of data access and retention requirements
- Configure rules to delete objects after a specified period to automatically remove unnecessary data and reduce storage expenses (e.g., BigQuery, Cloud Storage)
- Evaluate Google Cloud services for archiving data given business requirements

4.3 Identify high availability and disaster recovery strategies for data in Cloud Storage and Cloud SQL. Considerations include:

- Compare backup and recovery solutions offered as Google-managed services
- Determine when to use replication
- Distinguish between primary and secondary data storage location type (e.g., regions, dual-regions, multi-regions, zones) for data redundancy

4.2 Apply security measures and ensure compliance with data privacy regulations. Considerations include:

- Identify use cases for customer-managed encryption keys (CMEK), customer-supplied encryption keys (CSEK), and Google-managed encryption keys (GMEK)
- Understand the role of Cloud Key Management Service (Cloud KMS) to manage encryption keys
- Identify the difference between encryption in transit and encryption at rest