Google Cloud

Professional Data Engineer Renewal

Certification exam guide

A Professional Data Engineer empowers data-driven decisions by collecting, transforming, storing, and delivering data for diverse applications. A Professional Data Engineer designs and builds robust data infrastructure, optimizing for performance and security. This individual evaluates and selects solutions to meet business and regulatory needs, and administers data platforms effectively. A Professional Data Engineer leverages the latest technologies for data processing, cleaning, enrichment, and query generation and translation. A Professional Data Engineer understands the intricacies of data storage and processing, and is adept at designing, building, deploying, monitoring, maintaining, optimizing, and securing complex workloads.

Section 1: Designing data processing systems (~25% of the exam)

1.1 Designing for security and compliance. Considerations include:

- Regional considerations (data sovereignty) for data access and storage
- Legal and regulatory compliance

1.2 Designing for reliability and fidelity. Considerations include:

- Preparing and cleaning data (e.g., Dataform, Dataflow, and Cloud Data Fusion, prompting LLMs for query generation)
- Data validation

1.3 Designing for flexibility and portability. Considerations include:

- Data staging, cataloging, profiling, and discovery (data governance)

Section 2: Ingesting and processing the data (~10% of the exam)

2.1 Planning the data pipelines. Considerations include:

- Defining data transformation and orchestration logic

2.2 Building the pipelines. Considerations include:

- Transformations
    - Processing logic
    - AI data enrichment

2.3 Deploying and operationalizing the pipelines. Considerations include:

- Job automation and orchestration (e.g., Cloud Composer and Workflows)
- CI/CD (Continuous Integration and Continuous Deployment)

Section 3: Storing the data (~25% of the exam)

3.1 Selecting storage systems. Considerations include:

- Choosing managed services (e.g., BigQuery, BigLake, AlloyDB, Bigtable, Spanner, Cloud SQL, Cloud Storage, Firestore, Memorystore)

3.3 Using a data lake. Considerations include:

- Managing the lake (configuring data discovery, access, and cost controls)

3.4 Designing for a data platform. Considerations include:

- Building a data platform based on requirements by using Google Cloud tools (e.g., Dataplex, Dataplex Catalog, BigQuery, Cloud Storage)
- Building a federated governance model for distributed data systems

Section 4: Preparing and using data for analysis (~25% of the exam)

4.1 Preparing data for visualization. Considerations include:

- Security, data masking, Identity and Access Management (IAM), and Cloud Data Loss Prevention (Cloud DLP)

4.2 Preparing data for AI and ML. Considerations include:

- Preparing data for feature engineering, training and serving machine learning models (e.g., BigQueryML)

- Preparing unstructured data for embeddings and retrieval-augmented generation (RAG)

4.3 Sharing data. Considerations include:

- BigQuery sharing (Analytics Hub)

Section 5: Maintaining and automating data workloads (~15% of the exam)

5.2 Designing automation and repeatability. Considerations include:

- Scheduling and orchestrating jobs in a repeatable way

5.3 Organizing workloads based on business requirements. Considerations include:

- Capacity management (e.g., BigQuery Editions and reservations)

5.4 Monitoring and troubleshooting processes. Considerations include:

- Observability of data processes (e.g., Cloud Monitoring, Cloud Logging, BigQuery admin panel)