Google Cloud

Professional Data Engineer

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 (~22% of the exam)

1.1 Designing for security and compliance. Considerations include:

- Identity and Access Management (e.g., Cloud IAM and organization policies)
- Data security (encryption and key management)
- Privacy (e.g., strategies to handle personally identifiable information)
- Regional considerations (data sovereignty) for data access and storage
- Legal and regulatory compliance
- Designing the project, dataset, and table architecture to ensure proper data governance
- Multi-environment use cases (development vs. production)

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)
- Monitoring and orchestration of data pipelines
- Disaster recovery and fault tolerance
- Making decisions related to ACID (atomicity, consistency, isolation, and durability) compliance and availability
- Data validation

1.3 Designing for flexibility and portability. Considerations include:

- Mapping current and future business requirements to the architecture

- Designing for data and application portability (e.g., multi-cloud and data residency requirements)
- Data staging, cataloging, profiling, and discovery (data governance)

1.4 Designing data migrations. Considerations include:

- Analyzing current stakeholder needs, users, processes, and technologies, and creating a plan to get to desired state
- Planning migration and validation to Google Cloud (e.g., BigQuery Data Transfer Service, Database Migration Service, Transfer Appliance, Google Cloud networking, Datastream)

Section 2: Ingesting and processing the data (~25% of the exam)

2.1 Planning the data pipelines. Considerations include:

- Defining data sources and sinks
- Defining data transformation and orchestration logic
- Networking fundamentals
- Data encryption

2.2 Building the pipelines. Considerations include:

- Data cleansing
- Identifying the services (e.g., Dataflow, Apache Beam, Dataproc, Cloud Data Fusion, BigQuery, Pub/Sub, Apache Spark, Hadoop ecosystem, and Apache Kafka)
- Transformations
    - Batch
    - Streaming (e.g., windowing, late arriving data)
    - Processing logic
    - AI data enrichment
- Data acquisition and import
- Integrating with new data sources

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 (~20% of the exam)

3.1 Selecting storage systems. Considerations include:

- Analyzing data access patterns
- Choosing managed services (e.g., BigQuery, BigLake, AlloyDB, Bigtable, Spanner, Cloud SQL, Cloud Storage, Firestore, Memorystore)
- Planning for storage costs and performance
- Lifecycle management of data

3.2 Planning for using a data warehouse. Considerations include:

- Designing the data model
- Deciding the degree of data normalization
- Mapping business requirements
- Defining architecture to support data access patterns

3.3 Using a data lake. Considerations include:

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

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 (~15% of the exam)

4.1 Preparing data for visualization. Considerations include:

- Connecting to tools
- Precalculating fields
- BigQuery features for business intelligence (e.g., BI Engine, materialized views)
- Troubleshooting poor performing queries

- 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:

- Defining rules to share data
- Publishing datasets
- Publishing reports and visualizations
- BigQuery sharing (Analytics Hub)

Section 5: Maintaining and automating data workloads (~18% of the exam)

5.1 Optimizing resources. Considerations include:

- Minimizing costs per required business need for data
- Ensuring that enough resources are available for business-critical data processes
- Deciding between persistent or job-based data clusters (e.g., Dataproc)

5.2 Designing automation and repeatability. Considerations include:

- Creating directed acyclic graphs (DAGs) for Cloud Composer
- 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)
- Interactive or batch query jobs

5.4 Monitoring and troubleshooting processes. Considerations include:

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

- Troubleshooting error messages, billing issues, and quotas
- Manage workloads, such as jobs, queries, and compute capacity (reservations)

5.5 Maintaining awareness of failures and mitigating impact. Considerations include:

- Designing system for fault tolerance and managing restarts
- Running jobs in multiple regions or zones
- Preparing for data corruption and missing data
- Data replication and failover (e.g., Cloud SQL, Redis clusters)