Exam Guide -GCP PDE -KirkYagami 👨‍💻🕵️

Professional Data Engineer

https://cloud.google.com/learn/certification/guides/data-engineer

Section 1: Designing Data Processing Systems (~22% of the exam)

1.1 Designing for security and compliance

- Identity and Access Management (e.g., Cloud IAM and organization policies)
- Data security (encryption and key management)
- Privacy (e.g., personally identifiable information, and Cloud Data Loss Prevention API)
- Regional considerations (data sovereignty) for data access and storage
- Legal and regulatory compliance

1.2 Designing for reliability and fidelity

- Preparing and cleaning data (e.g., Dataprep, Dataflow, and Cloud Data Fusion)
- Monitoring and orchestration of data pipelines
- Disaster recovery and fault tolerance
- ACID compliance and availability decisions
- Data validation

1.3 Designing for flexibility and portability

- Mapping current and future business requirements to architecture
- Designing for data and application portability (e.g., multi-cloud and data residency requirements)
- Data staging, cataloging, and discovery (data governance)

1.4 Designing data migrations

- Analyzing current stakeholder needs, users, processes, and technologies
- Planning migration to Google Cloud (e.g., BigQuery Data Transfer Service, Database Migration Service)
- Designing migration validation strategy
- Ensuring proper data governance in project, dataset, and table architecture

Section 2: Ingesting and Processing the Data (~25% of the exam)

2.1 Planning the data pipelines

- Defining data sources and sinks

- Defining data transformation logic
- Networking fundamentals
- Data encryption

2.2 Building the pipelines

- Data cleansing
- Services (e.g., Dataflow, Apache Beam, Dataproc, Cloud Data Fusion, BigQuery, Pub/Sub, Apache Spark, Hadoop ecosystem, Apache Kafka)
- Transformations (Batch, Streaming, Language, Ad hoc data ingestion)

2.3 Deploying and operationalizing the pipelines

- 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

- Analyzing data access patterns
- Managed services (e.g., 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

- Designing the data model
- Data normalization considerations
- Mapping business requirements
- Architecture for supporting data access patterns

3.3 Using a data lake

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

3.4 Designing for a data mesh

- Building a data mesh with Google Cloud tools (Dataplex, Data Catalog, BigQuery, Cloud Storage)
- Segmenting data for distributed team usage
- 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

- Connecting to tools
- Precalculating fields
- BigQuery materialized views (view logic)
- Time data granularity decisions
- Troubleshooting queries (IAM and Cloud DLP)

4.2 Sharing data

- Defining data sharing rules
- Publishing datasets
- Publishing reports and visualizations
- Analytics Hub

4.3 Exploring and analyzing data

- Preparing data for feature engineering
- Data discovery processes

Section 5: Maintaining and Automating Data Workloads (~18% of the exam)

5.1 Optimizing resources

- Cost minimization strategies
- Resource allocation for critical data processes
- Choosing between persistent or job-based data clusters (e.g., Dataproc)

5.2 Designing automation and repeatability

- DAGs for Cloud Composer
- Scheduling repeatable jobs

5.3 Organizing workloads based on business requirements

- Pricing models (Flex, on-demand, flat rate slot pricing)
- Interactive vs. batch query jobs

5.4 Monitoring and troubleshooting processes

- Observability of data processes (e.g., Cloud Monitoring, Logging, BigQuery admin panel)
- Monitoring planned usage
- Troubleshooting errors, billing issues, and quotas
- Managing workloads (jobs, queries, compute capacity)

5.5 Maintaining awareness of failures and mitigating impact

- Fault tolerance strategies
- Running jobs in multiple regions or zones
- Preparedness for data corruption and missing data
- Data replication and failover (e.g., Cloud SQL, Redis clusters)

New 2024 GCP Professional Data Engineer Certification Exam Guide | Google 2024 | Google Professional Data Engineer Certification | Medium