

**Google Professional Data Engineer Training Curriculum**

**STRUCTURE**

**www.cromacampus.com | helpdesk@cromacampus.com | +91-120-4155255**

**Google Professional Data Engineer Training Curriculum**

*“Become a GCP certified Professional Data Engineer and bag your dream job in no time”*

**Course Objectives:**

* Prepare yourself for the GCP Professional Data engineer certification exam and clear your certification exam in the first attempt
* Add an attractive credential in your resume that is really appreciated by Companies.
* Improve your overall Data engineering, data processing skills, and explore more job prospects with better salary packages.
* Boost your social media profiles especially LinkedIn by adding this certification and become one of the top persons to be chosen by industries.

**Course Description:**

A Google Certified Professional - Data Engineer enables data-driven decision making by collecting, transforming, and visualizing data. With our Professional Data Engineer Certification Training Program, you will learn how to design, build, maintain, and troubleshoot data processing systems with a particular emphasis on the security, reliability, fault-tolerance, scalability, fidelity, and efficiency of such systems.

You will also learn to analyse data to gain insight into business outcomes, builds statistical models to support decision-making, and creates machine learning models to automate and simplify key business processes.

Becoming a Google Certified Professional – Data Engineer means, you demonstrate the ability to the following:

* Build and maintain data structures and databases
* Design data processing systems
* Analyze data and enable machine learning
* Model business processes for analysis and optimization
* Design for reliability
* Visualize data and advocate policy
* Design for security and compliance

**Necessary Details about the Certification:**

* Format: Multiple choice and multiple select
* Length: 2 hours
* Registration fee: USD $200
* Prerequisites: None

**Certification Structure:**

* Design data processing systems (20-25%)
* Build and operationalize data processing systems (20-25%)
* Operationalize machine learning models (20-25%)
* Ensure solution quality (20-25%)

**Course Content:**

**Module 1: Data Processing and Data Storage Fundamentals in GCP**

* Data processing Fundamentals
  + Data Processing Concepts
  + Data Processing Pipelines
* Data Storage Fundamentals
  + About GCP
  + Data Storage in GCP
  + Working with Data
  + Cloud Storage
  + Data Transfer Services
  + Cloud Fire Store
  + Cloud Spanner
  + Cloud Memory Store
  + Different Memory options
* Selecting the best memory storage
  + Compare storage options
  + Mapping storage systems to business requirements
  + Data modeling
  + Trade-offs involving latency, throughput, transactions
  + Distributed systems
  + Schema design

**Module 2: Designing Data pipelines**

* Data publishing and visualization
* Online (interactive) vs. batch predictions
* Batch and streaming data (e.g., Cloud Dataflow, Cloud Dataproc, Apache Spark and Hadoop ecosystem, Cloud Pub/Sub, Apache Kafka)
* Big Data Ecosystem
  + MapReduce
  + Hadoop & HDFS
  + Apache Pig
  + Apache Spark
  + Apache Kafka
* Real-time Messaging with Pub/Sub
  + Pub/sub basics
  + pub/Sub Terminologies
  + Advanced Pub/Sub Concepts
  + Working with Pub/Sub
* Cloud Data Flow Pipelining
  + Introduction to Data flow
  + Pipeline Lifecycle
  + Dataflow pipeline concepts
  + Advanced Dataflow concepts
  + Dataflow security and access
  + Using Dataflow
* Cloud Dataproc
  + Dataproc Basics
  + Working with Dataproc
  + Advanced Dataproc
* NoSQL Data with Cloud Big Table
  + Big Table Concepts
  + Big Table Architecture
  + Big Table Data Model
  + Big Table Schema Design
  + Big Table Advanced Concepts
* Data Analytics using BigQuery
  + BigQuery Basics
  + Using BigQuery
  + Partitioning and Clustering
  + Best Practices
  + Securing BigQuery
  + BigQuery Monitoring and Logging
  + Machine Learning with BigQuery ML
  + Working with BigQuery
  + Advanced BigQuery Concepts
* Data Exploration with Cloud Datalab
  + Datalab Concepts
  + Working with Datalab
* Visualization with Cloud Data Studio
  + Reporting & Business intelligence
  + Data Distribution
  + Introduction to Cloud Data Studio
  + Charts and Filters
* Job automation and orchestration (e.g., Cloud Composer)
  + Orchestration with Cloud Composer
  + Cloud Composer Overview
  + Cloud Composer Architecture
  + Working with Cloud Composer
  + Advanced Cloud Composer Concepts

**Module 3: Design a Data Processing Solution**

* Steps for Designing
  + Choice of infrastructure
  + System availability and fault tolerance
  + Use of distributed systems
  + Capacity planning
  + Hybrid cloud and edge computing
  + Architecture options (e.g., message brokers, message queues, middleware, service-oriented architecture, serverless functions)
  + At least once, in-order, and exactly once, etc., event processing
* Migrating data warehousing and data processing
  + Awareness of current state and how to migrate a design to a future state
  + Migrating from on-premises to cloud (Data Transfer Service, Transfer Appliance, Cloud Networking)
  + Validating a migration

**Module 4: Building and operationalizing data processing systems**

* Building and operationalizing Storage Solutions
  + Cloud Managed Services
  + Effectives Use of Managed Services
  + Storage Cost and performance
  + Lifecycle Management of Data
* Building and operationalizing Pipelines
  + Data cleansing
  + Batch and streaming
  + Transformation
  + Data acquisition and import
  + Integrating with new data sources
* Building and operationalizing processing infrastructure
  + Provisioning resources
  + Monitoring pipelines
  + Adjusting pipelines
  + Testing and quality control

**Module 5: Operationalizing machine learning models**

* Introduction to Machine Learning
  + Machine Learning Introduction
  + Machine Learning Basics
  + Machine Learning Types and Models
  + Overfitting
  + Hyperparameters
  + Feature Engineering
* Machine Learning with TesnorFlow
  + Deep Learning with TensorFlow
  + Introduction to Artificial Neural Networks
  + Neural Network Architectures
  + Building a Neural Network
* Leveraging pre-built ML models as a service. Considerations include:
  + ML APIs (e.g., Vision API, Speech API)
  + Customizing ML APIs (e.g., AutoML Vision, Auto ML text)
  + Conversational experiences (e.g., Dialogflow)
* Deploying an ML pipeline
  + Ingesting appropriate data
  + Retraining of machine learning models (Cloud Machine Learning Engine, BigQuery ML, Kubeflow, Spark ML)
  + Continuous evaluation
* Choosing the appropriate training and serving infrastructure
  + Distributed vs. single machine
  + Use of edge compute
  + Hardware accelerators (e.g., GPU, TPU)
* Measuring, monitoring, and troubleshooting machine learning models
  + Machine learning terminology (e.g., features, labels, models, regression, classification, recommendation, supervised and unsupervised learning, evaluation metrics)
  + Impact of dependencies of machine learning models
  + Common sources of error (e.g., assumptions about data)

**Module 6: Ensuring Solution Quality**

* Designing for security and compliance
  + Identity and access management (e.g., Cloud IAM)
  + Data security (encryption, key management)
  + Ensuring privacy (e.g., Data Loss Prevention API)
  + Legal compliance (e.g., Health Insurance Portability and Accountability Act (HIPAA), Children's Online Privacy Protection Act (COPPA), FedRAMP, General Data Protection Regulation (GDPR))
* Ensuring scalability and efficiency
  + Building and running test suites
  + Pipeline monitoring (e.g., Stack Driver)
  + Assessing, troubleshooting, and improving data representations and data processing infrastructure
  + Resizing and autoscaling resources
* Ensuring reliability and fidelity
  + Performing data preparation and quality control (e.g., Cloud Dataprep)
  + Verification and monitoring
  + Planning, executing, and stress testing data recovery (fault tolerance, rerunning failed jobs, performing retrospective re-analysis)
  + Choosing between ACID, idempotent, eventually consistent requirements
* Ensuring flexibility and portability
  + Mapping to current and future business requirements
  + Designing for data and application portability (e.g., multi-cloud, data residency requirements)
  + Data staging, catalog, and discovery

**Module 7: Placement Guide**

* What is an Interview?
* Tips to clear an Interview
* Common Interview questions and answers
* GCP Interview Questions and Answers
* Resume Building Guide
* Attempt for GCP Professional Data Engineer Certification Exam
* Earn Credentials and Start applying for Jobs