

Sela.

DataEngGC

Data Engineering

college@sela.co.il

03-6176666





Data Engineering

DataEngGC - Version: 1

 4 days Course

Description:

Get hands-on experience with designing and building data processing systems on Google Cloud. This course uses lectures, demos, and hands-on labs to show you how to design data processing systems, build end-to-end data pipelines, analyze data, and implement machine learning. This course covers structured, unstructured, and streaming data.

Intended audience:

This class is intended for developers who are responsible for: Extracting, loading, transforming, cleaning, and validating data.
Designing pipelines and architectures for data processing.
Integrating analytics and machine learning capabilities into data pipelines.
Querying datasets, visualizing query results, and creating reports.

Prerequisites:

Objectives:

Design and build data processing systems on Google Cloud.
Process batch and streaming data by implementing autoscaling data pipelines on Dataflow.
Derive business insights from extremely large datasets using BigQuery.
Leverage unstructured data using Spark and ML APIs on Dataproc.
Enable instant insights from streaming data.



Understand ML APIs and BigQuery ML, and learn to use AutoML to create powerful models without coding

Topics:

Introduction to Data Engineering

- Explore the role of a data engineer
- Analyze data engineering challenges
- Introduction to BigQuery
- Data lakes and data warehouses
- Transactional databases versus data warehouses
- Partner effectively with other data teams
- Manage data access and governance
- Build production-ready pipelines
- Review Google Cloud customer case study

Building a Data Lake

- Introduction to data lakes
- Data storage and ETL options on Google Cloud
- Building a data lake using Cloud Storage
- Securing Cloud Storage
- Storing all sorts of data types
- Cloud SQL as a relational data lake

Building a Data Warehouse

- The modern data warehouse
- Introduction to BigQuery



- Getting started with BigQuery
- Loading data
- Exploring schemas
- Schema design
- Nested and repeated fields
- Optimizing with partitioning and clustering

Introduction to Building Batch Data Pipelines

- EL, ELT, ETL
- Quality considerations
- How to carry out operations in BigQuery
- Shortcomings
- ETL to solve data quality issues

Executing Spark on Dataproc

- The Hadoop ecosystem
- Run Hadoop on Dataproc
- Cloud Storage instead of HDFS
- Optimize Dataproc

Serverless Data Processing with Dataflow

- Introduction to Dataflow
- Why customers value Dataflow
- Dataflow pipelines
- Aggregating with GroupByKey and Combine
- Side inputs and windows
- Dataflow templates



- Dataflow SQL

Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

- Building batch data pipelines visually with Cloud Data Fusion
- Components
- UI overview
- Building a pipeline
- Exploring data using Wrangler
- Orchestrating work between Google Cloud services with Cloud Composer
- Apache Airflow environment
- DAGs and operators
- Workflow scheduling
- Monitoring and logging

Introduction to Processing Streaming Data

- Process Streaming Data

Serverless Messaging with Pub/Sub

- Introduction to Pub/Sub
- Pub/Sub push versus pull
- Publishing with Pub/Sub code

Dataflow Streaming Features

- Streaming data challenges
- Dataflow windowing



High-Throughput BigQuery and Bigtable Streaming Features

- Streaming into BigQuery and visualizing results
- High-throughput streaming with Cloud Bigtable
- Optimizing Cloud Bigtable performance

Advanced BigQuery Functionality and Performance

- Analytic window functions
- Use With clauses
- GIS functions
- Performance considerations

Introduction to Analytics and AI

- What is AI?
- From ad-hoc data analysis to data-driven decisions
- Options for ML models on Google Cloud

Prebuilt ML Model APIs for Unstructured Data

- Unstructured data is hard
- ML APIs for enriching data

Big Data Analytics with Notebooks

- What's a notebook?
- BigQuery magic and ties to Pandas



Production ML Pipelines

- Ways to do ML on Google Cloud
- Vertex AI Pipelines
- AI Hub

Custom Model Building with SQL in BigQuery ML

- BigQuery ML for quick model building
- Supported models

Custom Model Building with AutoML

- Why AutoML?
- AutoML Vision
- AutoML NLP
- AutoML tables