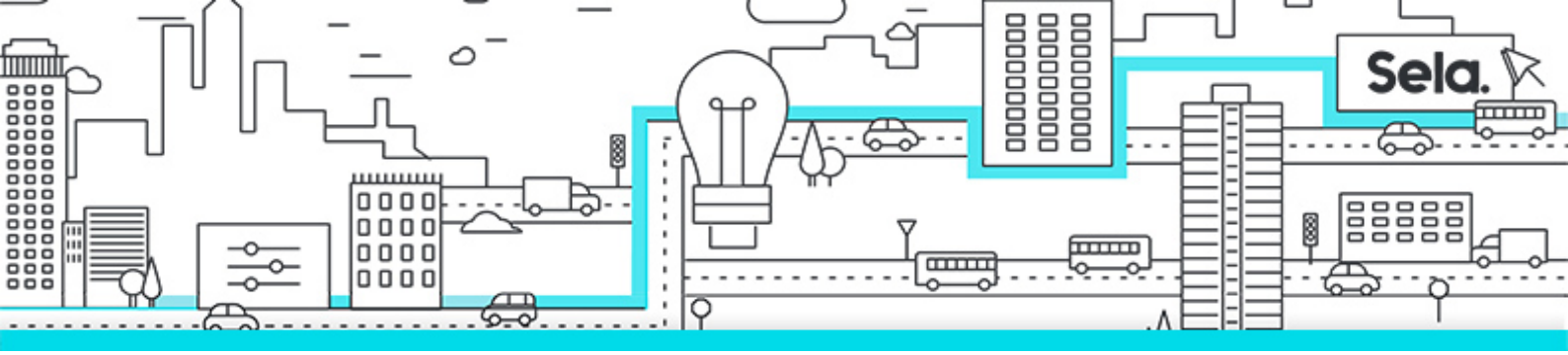
GCPDataEng

Data Engineering on Google Cloud

college@sela.co.il 03-6176666



Data Engineering on Google Cloud

GCPDataEng - Version: 1

4 days course

Description:

This course provides participants a hands-on introduction to designing and building data processing systems on Google Cloud Platform. Participants will learn how to design data processing systems, build end-to-end data pipelines, analyze data and carry out machine learning. The course covers structured, unstructured, and streaming data.

Intended Audience:

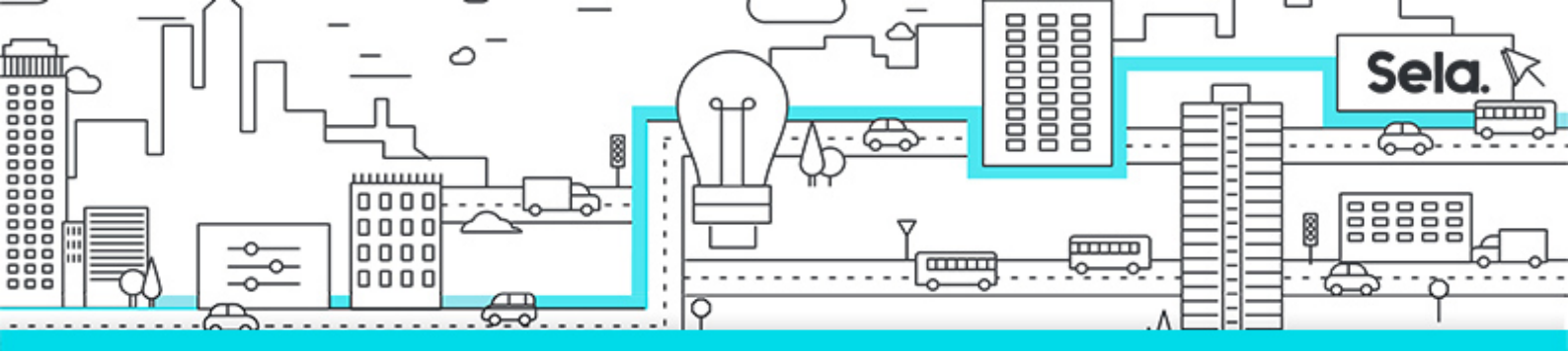
Extracting, Loading, Transforming, cleaning, and validating data
Designing pipelines and architectures for data processing
Creating and maintaining machine learning and statistical models
Querying datasets, visualizing query results and creating reports

Prerequisites:

• Completed Google Cloud Fundamentals: Big Data & Machine Learning course
OR
• have equivalent experience
• Basic proficiency with common query language such as SQL
• Experience with data modeling, extract, transform, load activities
• Developing applications using a common programming language such as Python
• Familiarity with Machine Learning and/or statistics

Objectives:

• Design and build data processing systems on Google Cloud Platform



• Process batch and streaming data by implementing autoscaling data pipelines on
• Cloud Dataflow
• Derive business insights from extremely large datasets using Google BigQuery
• Train, evaluate and predict using machine learning models using Tensorflow and
• Cloud ML
• Leverage unstructured data using Spark and ML APIs on Cloud Dataproc
• Enable instant insights from streaming data

Topics:

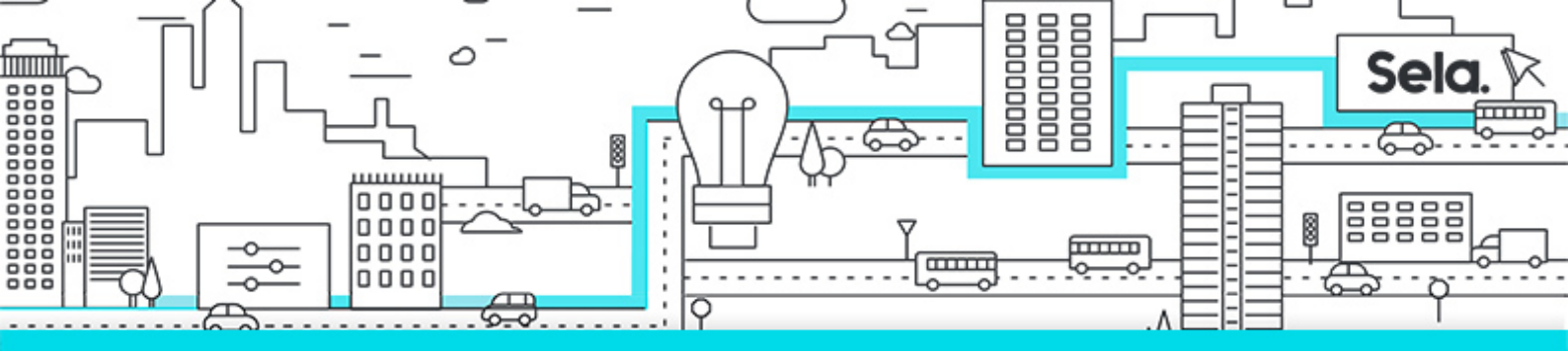
- **Module 1: Google Cloud Dataproc Overview**
  - Creating and managing clusters.
  - Leveraging custom machine types and preemptible worker nodes.
  - Scaling and deleting Clusters.

- **Module 2: Running Dataproc Jobs**
  - Running Pig and Hive jobs.
  - Separation of storage and compute.

- **Module 3: Integrating Dataproc with Google Cloud Platform**
  - Customize cluster with initialization actions.
  - BigQuery Support.

- **Module 4: Making Sense of Unstructured Data with Google's Machine**
  - Google's Machine Learning APIs.
  - Common ML Use Cases.
  - Invoking ML APIs.

- **Module 5: Serverless data analysis with BigQuery**
  - What is BigQuery.



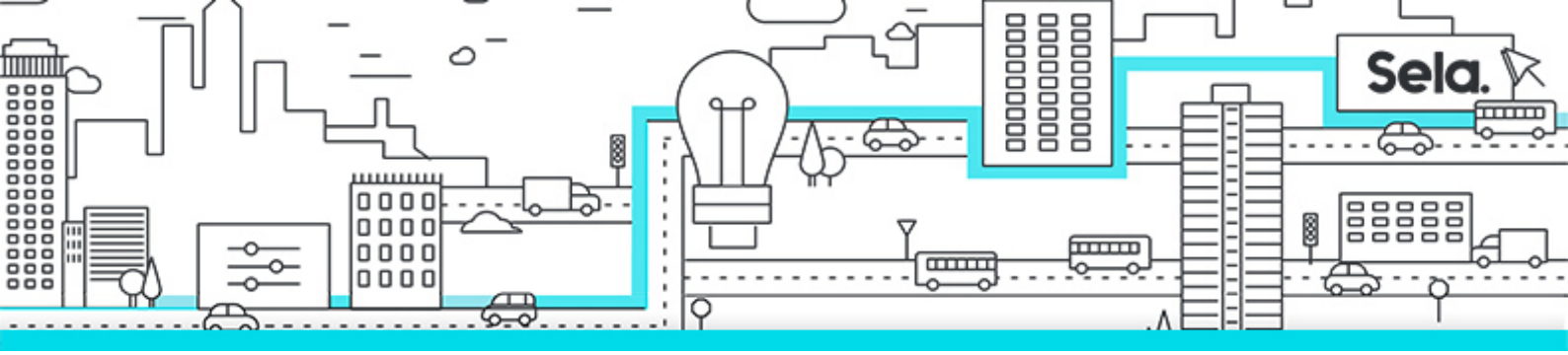
- Queries and Functions.
- Lab: Writing queries in BigQuery.
- Loading data into BigQuery.
- Exporting data from BigQuery.
- Nested and repeated fields.
- Querying multiple tables.
- Performance and pricing.

• **Module 6: Serverless, autoscaling data pipelines with Dataflow**
- The Beam programming model.
- Data pipelines in Beam Python.
- Data pipelines in Beam Java.
- Scalable Big Data processing using Beam.
- Incorporating additional data.
- Handling stream data.
- GCP Reference architecture.

• **Module 7: Getting started with Machine Learning**
- What is machine learning (ML).
- Effective ML: concepts, types.
- ML datasets: generalization.

• **Module 8: Building ML models with Tensorflow**
- Getting started with TensorFlow.
- TensorFlow graphs and loops + lab.
- Monitoring ML training.

• **Module 9: Scaling ML models with CloudML**
- Why Cloud ML?
- Packaging up a TensorFlow model.
- End-to-end training.



- **Module 10: Feature Engineering**
  - Creating good features.
  - Transforming inputs.
  - Synthetic features.
  - Preprocessing with Cloud ML.

- **Module 11: Architecture of streaming analytics pipelines**
  - Stream data processing: Challenges.
  - Handling variable data volumes.
  - Dealing with unordered/late data.

- **Module 12: Ingesting Variable Volumes**
  - What is Cloud Pub/Sub?
  - How it works: Topics and Subscriptions.

- **Module 13: Implementing streaming pipelines**
  - Challenges in stream processing.
  - Handle late data: watermarks, triggers, accumulation.

- **Module 14: Streaming analytics and dashboards**
  - Streaming analytics: from data to decisions.
  - Querying streaming data with BigQuery.
  - What is Google Data Studio?

- **Module 15: High throughput and low-latency with Bigtable**
  - What is Cloud Spanner?
  - Designing Bigtable schema.
  - Ingesting into Bigtable.