Google

Student Resource Guide

Data Engineering on Google Cloud Platform

Day 0 - Big Data Machine Learning Fundamentals

(Note: Day 0 content may or may not be part of your curriculum for the week. Check with your instructor)

PDF - Introduction to GCP
PDF - Product Recommendations using Cloud SQL and Spark
PDF - Predict Visitor Purchases with a Classification Model in BigQuery ML
PDF - Real Time Dashboards with Pub/Sub and Cloud Dataflow
PDF - Deriving Insights from Unstructured Data using Machine Learning

Day 1 - Modernizing Data Lakes and Data Warehouses with GCP

DOWNLOAD	
PDF	**M1 - Introduction to Data Engineering**
	Explore the role of a data engineer
	Analyze data engineering challenges
	Intro to BigQuery
	Data Lakes and Data Warehouses
	Demo: Federated Queries with BigQuery
	Transactional Databases vs Data Warehouses
	Website Demo: Finding PII in your dataset with DLP API
	Partner effectively with other data teams
	Manage data access and governance

	Build production-ready pipelines
	Review GCP customer case study
	Lab: Analyzing Data with BigQuery
PDF	**M2 - Building a Data Lake**
	Introduction to Data Lakes
	Data Storage and ETL options on GCP
	Building a Data Lake using Cloud Storage
	Securing Cloud Storage
	Storing All Sorts of Data Types
	Video Demo: Running federated queries on Parquet and ORC files in BigQuery
	Cloud SQL as a relational Data Lake
	Lab: Loading Taxi Data into Cloud SQL
PDF	**M3 - Building a Data Warehouse**
	The modern data warehouse
	Intro to BigQuery
	Demo: Query TB+ of data in seconds
	Getting Started
	Loading Data
	Video Demo: Querying Cloud SQL from BigQuery
	Lab: Loading Data with Console and CLI
	Exploring Schemas
	Demo: Exploring BigQuery Public Datasets with SQL using INFORMATION_SCHEMA
	Schema Design
	Nested and Repeated Fields
	Demo: Nested and repeated fields in BigQuery
	Lab: ARRAYs and STRUCTs
	Optimizing with Partitioning and Clustering
	Demo: Partitioned and Clustered Tables in BigQuery

	Preview: Transforming Batch and Streaming Data
PDF	**M4 - Summary**
	Summary

Day 2 - Batch Processing of Data with Spark and Hadoop on GCP

PDF	**M1 - Introduction to Building Batch Data Pipelines**
	EL, ELT, ETL
	Quality considerations
	How to carry out operations in BigQuery
	Demo: ELT to improve data quality in BigQuery
	Shortcomings
	ETL to solve data quality issues
PDF	**M2 - Executing Spark on Cloud Dataproc**
	The Hadoop ecosystem
	Running Hadoop on Cloud Dataproc
	GCS instead of HDFS
	Optimizing Dataproc
	Lab: Running Apache Spark jobs on Cloud Dataproc
PDF	**M3 - Serverless Data Processing with Cloud Dataflow**
	Cloud Dataflow
	Why customers value Dataflow
	Dataflow Pipelines
	Lab: A Simple Dataflow Pipeline (Python/Java)
	Lab: MapReduce in Dataflow (Python/Java)
	Lab: Side Inputs (Python/Java)
	Dataflow Templates

	Dataflow SQL
PDF	**M4 - 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
	Lab: Building and executing a pipeline graph in Cloud Data Fusion
	Orchestrating work between GCP services with Cloud Composer
	- Apache Airflow Environment
	- DAGs and Operators
	- Workflow Scheduling
	- Monitoring and Logging
	Lab: An Introduction to Cloud Composer
PDF	**M5 - Summary**
	Summary

Day 3 - Building Resilient Streaming Analytics Systems on GCP

PDF	**M1 - Introduction to Processing Streaming Data**
	Processing Streaming Data
PDF	**M2 - Serverless Messaging with Cloud Pub/Sub**
	Cloud Pub/Sub
	Lab: Publish Streaming Data into Pub/Sub
PDF	**M3 - Cloud Dataflow Streaming Features**
	Cloud Dataflow Streaming Features

	Lab: Streaming Data Pipelines
PDF	**M4 - High-Throughput BigQuery and Bigtable Streaming Features**
	BigQuery Streaming Features
	Lab: Streaming Analytics and Dashboards
	Cloud Bigtable
	Lab: Streaming Data Pipelines into Bigtable
PDF	**M5 - Advanced BigQuery Functionality and Performance**
	Analytic Window Functions
	Using With Clauses
	GIS Functions
	Demo: Mapping Fastest Growing Zip Codes with BigQuery GeoViz
	Performance Considerations
	Lab: Optimizing your BigQuery Queries for Performance
	Optional Lab: Creating Date-Partitioned Tables in BigQuery
PDF	**M6 - Summary**
	Summary

Day 4 - Smart Analytics, Machine Learning and AI on GCP

PDF	**M1 - Introduction to Analytics and AI**
	What is AI?
	From Ad-hoc Data Analysis to Data Driven Decisions
	Options for ML models on GCP
PDF	**M2 - Prebuilt ML model APIs for Unstructured Data**
	Unstructured Data is Hard
	ML APIs for Enriching Data

	Lab: Using the Natural Language API to Classify Unstructured Text
PDF	**M3 - Big Data Analytics with Cloud AI Platform Notebooks**
	What's a Notebook
	BigQuery Magic and Ties to Pandas
	Lab: BigQuery in Jupyter Labs on AI Platform
PDF	**M4 - Production ML Pipelines with Kubeflow**
	Ways to do ML on GCP
	Kubeflow
	AI Hub
	Lab: Running ML Pipelines on Kubeflow
PDF	**M5 - Custom Model building with SQL in BigQuery ML**
	BigQuery ML for Quick Model Building
	Demo: Train a model with BigQuery ML to predict NYC taxi fares
	Supported Models
	Lab Option 1: Predict Bike Trip Duration with a Regression Model in BQML
	Lab Option 2: Movie Recommendations in BigQuery ML
PDF	**M6 - Custom Model building with Cloud AutoML**
	Why Auto ML?
	Auto ML Vision
	Auto ML NLP
	Auto ML Tables
PDF	**M7 - Summary**
	Summary