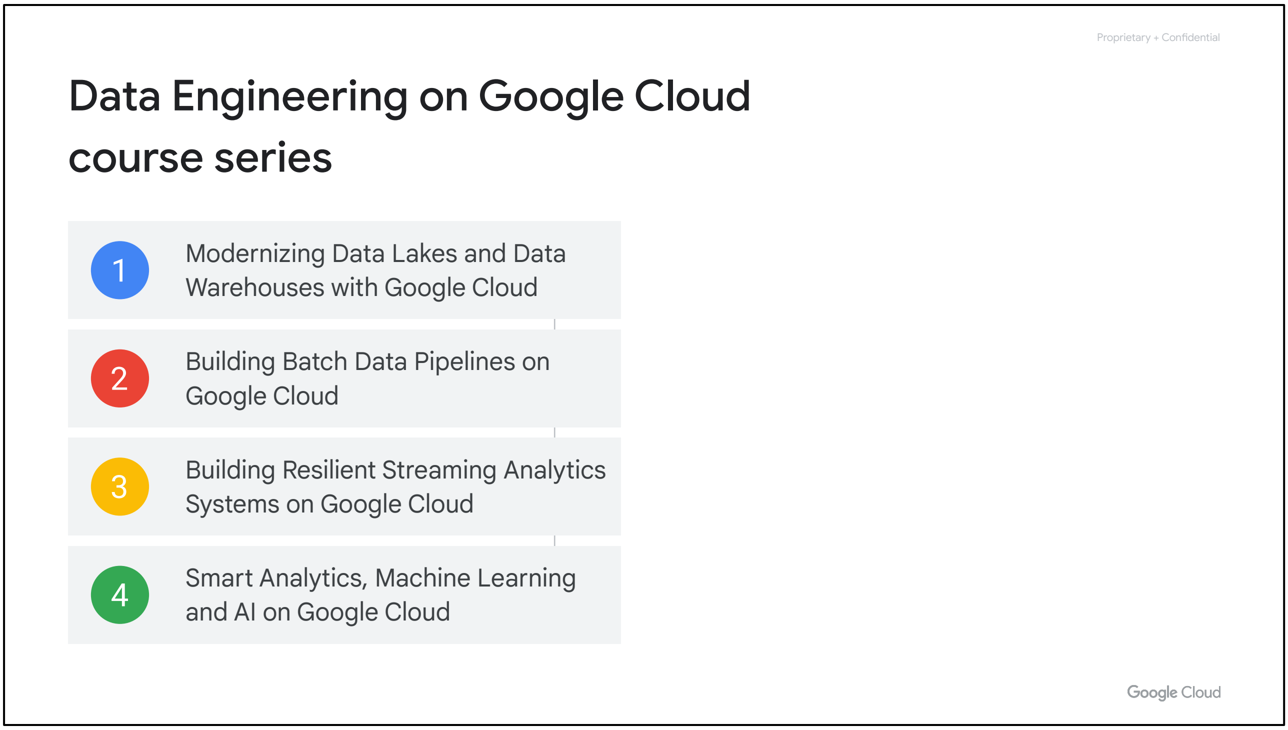
Google Cloud

Data Engineering on Google Cloud

Damon Runion
Technical Curriculum Developer

Hello and welcome to the **Data Engineering on Google Cloud** course series.

I’m Damon and I am a Technical Curriculum Developer at Google. Together with my fellow instructors, we look forward to showing you how to design data processing systems, build end-to-end data pipelines, analyze data, and implement machine learning. In addition to video lectures, you will also complete a series of hands-on labs.



Proprietary + Confidential

Data Engineering on Google Cloud course series

1. Modernizing Data Lakes and Data Warehouses with Google Cloud
2. Building Batch Data Pipelines on Google Cloud
3. Building Resilient Streaming Analytics Systems on Google Cloud
4. Smart Analytics, Machine Learning and AI on Google Cloud

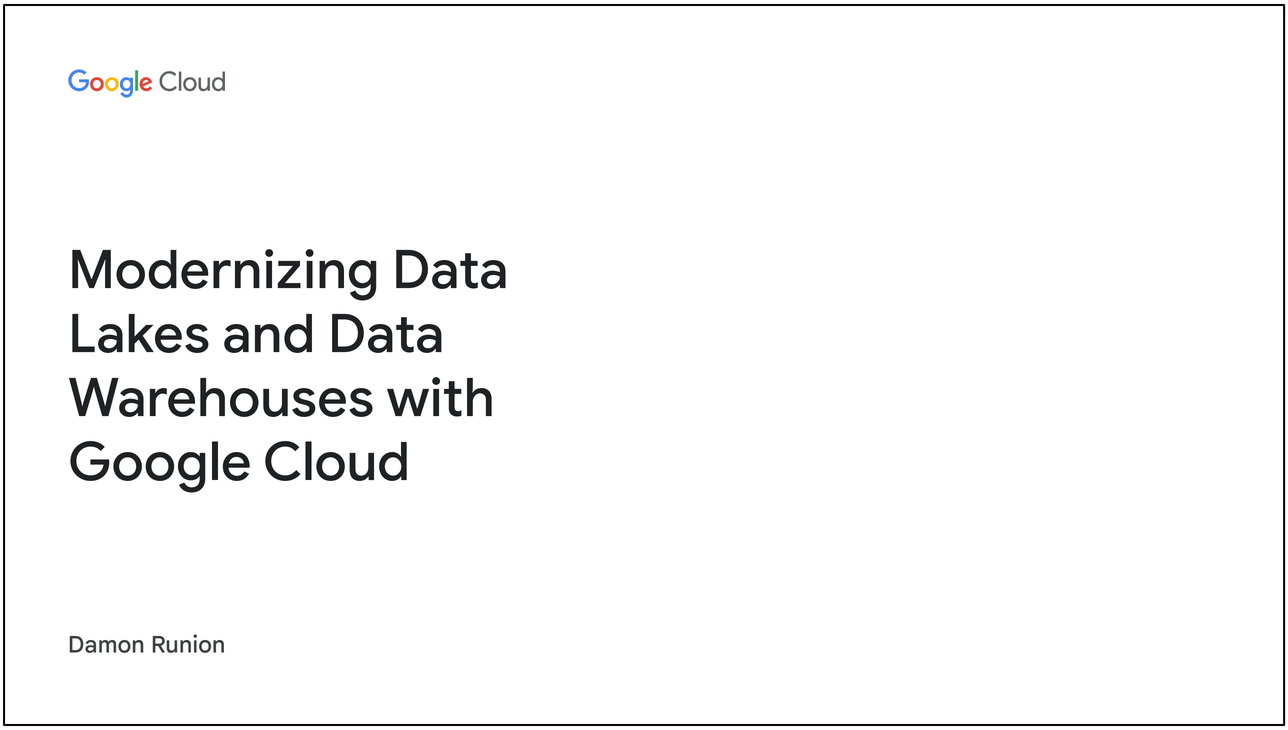
Google Cloud

As part of the Data Engineering on Google Cloud course series, we will first discuss the differences between data lakes and data warehouses, the two key components of any data pipeline. This course highlights use-cases for each type of storage and dives into the available data lake and warehouse solutions on Google Cloud in technical detail. Also, this course describes the role of a data engineer, the benefits of a successful data pipeline to business operations, and examines why data engineering should be done in a cloud environment.

Data pipelines typically fall under one of the Extract-Load, Extract-Load-Transform or Extract-Transform-Load paradigms. So the next course, Building Batch Data Pipelines, describes which paradigm should be used and when for batch data. Furthermore, it covers several technologies on Google Cloud for data transformation including BigQuery, executing Spark on Dataproc, pipeline graphs in Data Fusion and serverless data processing with Dataflow.

Processing streaming data is becoming increasingly popular as streaming enables organizations to get real-time metrics on operations. So the third course covers how to build streaming data pipelines on Google Cloud. Pub/Sub is the primary product for handling incoming streaming data. The course also covers how to apply aggregations and transformations to streaming data using Dataflow, and how to store processed records in BigQuery or Bigtable for analysis.

Incorporating machine learning into data pipelines increases the ability of organizations to extract insights from their data. The final course covers several ways for machine learning to be included in data pipelines on Google Cloud depending on the level of customization required. For little to no customization, the course covers AutoML. For more tailored machine learning capabilities, the course introduces Notebooks and BigQuery Machine Learning. Also, the final course covers how to productionize machine learning solutions using Kubeflow.

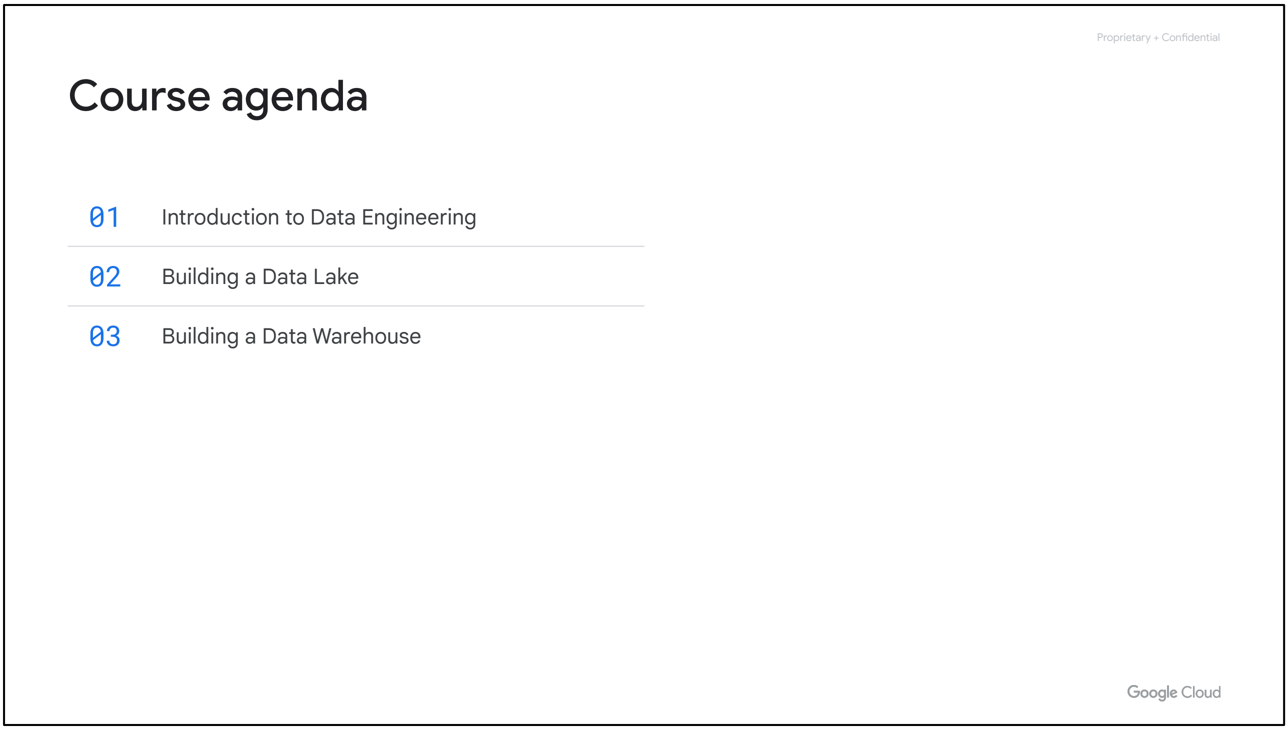


Google Cloud

Modernizing Data Lakes and Data Warehouses with Google Cloud

Damon Runion

Welcome to **Modernizing Data Lakes and Data Warehouses with Google Cloud**, the first course of the Data Engineering on Google Cloud course series.



Proprietary + Confidential

Course agenda

01 Introduction to Data Engineering

02 Building a Data Lake

03 Building a Data Warehouse

Google Cloud

We'll start off by describing the role of a data engineer. We'll talk about a data engineer's clients and what the benefits of a successful data pipeline are for your organization. Also, we'll explain why data engineering should be done in a cloud environment.

We'll concentrate on data lakes and data warehouses in this course. These are the two key components of any data pipeline. We'll describe the differences between data lakes and warehouses and highlight use-cases for each type of storage. Also, we'll go into the available data lake and data warehouse solutions on Google Cloud in some technical detail.

Finally, you'll get hands-on experience with data lakes and data warehouses on Google Cloud using QwikLabs.