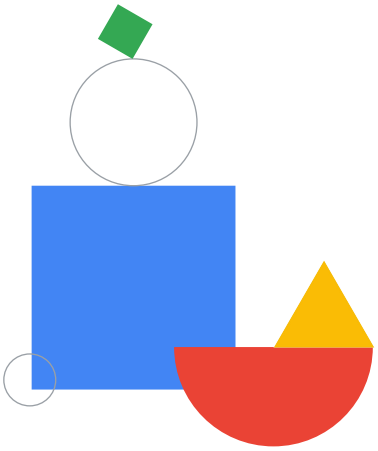
Google Cloud

Modernizing Data Lakes and Data Warehouses with Google Cloud



Summary

Let’s review some keys concepts we covered on data lakes and data warehouses.

Proprietary + Confidential

Summary

- Data engineers build data pipelines.
- The customers of a data engineer are all the people who make decisions with data.
- The three primary advantages of doing data engineering in the cloud are:
  - Ability to separate compute and storage
  - Serverless products
  - Not having to manage infrastructure

Google Cloud

- The primary role of a data engineer is to build data pipelines.
- The ultimate purpose of a data pipeline is to enable stakeholders in an organization to use data to make faster and better decisions.
- While the role of a data engineer is not new, being able to build data pipelines entirely in the cloud is relatively new. We argue that doing data engineering in the cloud is advantageous because you can separate compute from storage, and you don't have to worry about managing infrastructure and even software. This allows you to spend more time on what matters; getting insights from data.

Proprietary + Confidential

Summary

- Difference between a data lake and data warehouse.
- Google Cloud Storage as a data lake solution.
- BigQuery as a data warehouse solution.
- Differences between ETL, ELT and EL.
- Google Cloud reference architectures for ETL, ELT and EL.

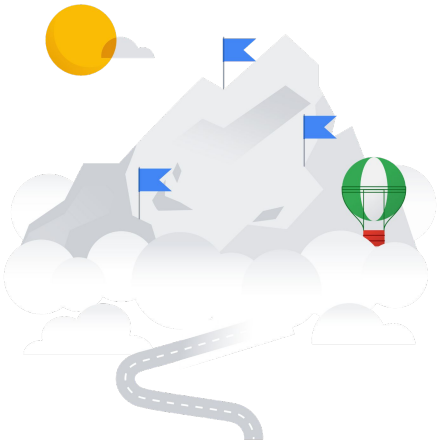
Google Cloud

- We introduced data lakes and data warehouses and discussed the key differences between the two. At a high level, a data lake is a place to store unprocessed data. While a data warehouse is a place to store transformed data that you ultimately want to use for analytics, machine learning, and dashboards.
- Next, we discussed Cloud Storage as the data lake solution on Google Cloud in some technical depth. We also presented other Google Cloud solutions for low-latency requirements, transactional workloads, and structured data.
- We introduced BigQuery as the data warehouse solution on Google Cloud. We discussed partitioning and clustering in BigQuery as techniques for improving query performance.
- Also, we talked about E-L, E-L-T, and E-T-L and how these relate to data lakes and warehouses.
- Finally, we presented some reference architectures on Google Cloud for streaming and batch data pipelines. The hope is that these reference architectures serve as a starting point for your data pipeline.

Proprietary + Confidential

Data Engineering on Google Cloud learning path

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

We've completed **Modernizing Data Lakes and Data Warehouses with Google Cloud**.

Building Batch Data Pipelines on Google Cloud is the second part of the Data Engineering on Google Cloud course, which we will cover on day 2.