Data Lakes and Data Warehouses Course Summary

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

Module summary

- Data engineers build data pipelines
- The customers of a data engineer are all the people who make decisions with data

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 a business to use data to make faster and better decisions to improve their business. While the role of a data engineer is not new, being able to build data pipelines entirely in the cloud is relatively new.

Module summary

- 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

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.

Module summary

- Difference between a data lake and data warehouse
- Google Cloud Storage as a data lake solution
- Getting your data to Google Cloud

Google Cloud

We introduced data lakes and data warehouses and the key differences between the two. At a high level, a data lake is a place to store unprocessed data. 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 GCP in some technical depth. We also presented other GCP solutions for low-latency requirements, transactional workloads, and structured data.

Module summary

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

Google Cloud

Finally, 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 EL, ELT, and ETL and how these relate to data lakes and warehouses. Finally, we presented some reference architectures on GCP for streaming and batch data pipelines. The hope is that these reference architectures serve as a starting point for your data pipeline.