Unit 3: Data Pipelines

Elements of Data Integration (CS 333E)

Mar 27, 2026

Today’s agenda

Midterm Presentations

- chris-varun
- CS333E-sree-bai
- Griffin_Lucas_Repository
- sharkcat

Unit 2 Resubmissions

- Projects 5 - 7

Unit 3 Overview

- Data pipeline intro
- dbt setup

Reading Quiz

- Chapter 4 (Data Eng text)

Data Pipeline Fundamentals

- **Definition:** A program that transforms data in a well-defined way, with a distinct **input** and **output**. The transformation is performed in multiple, ordered **stages**.

- **Execution:** Can be **periodic** (e.g., hourly) or **continuous** (real-time).

- **Architecture:**
  - Most are a **chained series**, where the output of one stage becomes the input for the next.
  - Intermediate transformations are often **persisted** (stored to disk or another non-volatile system).

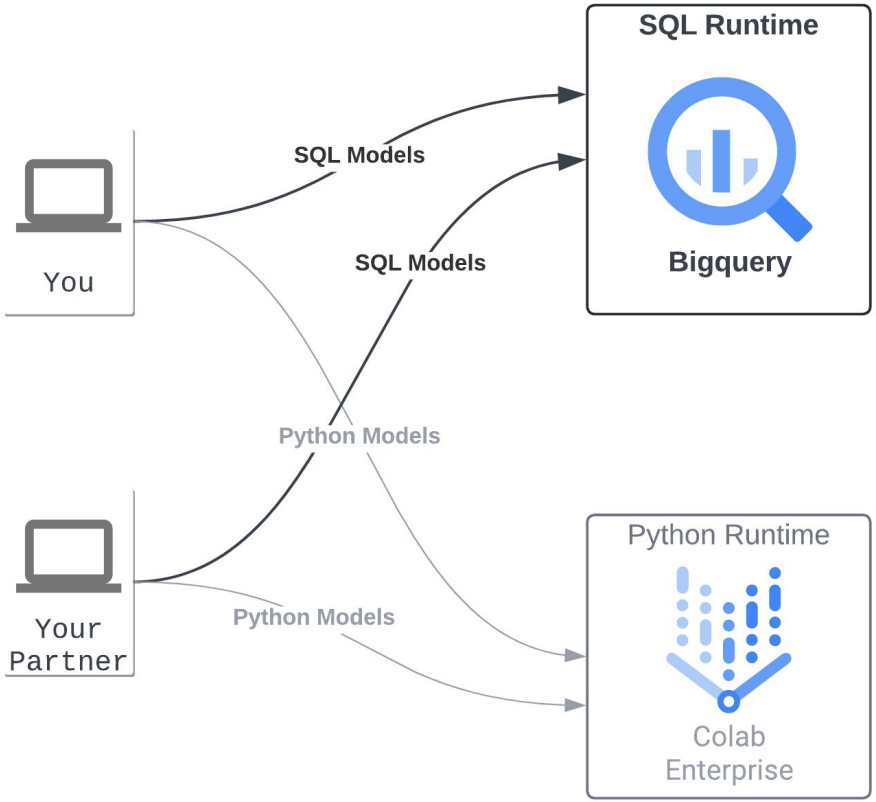
Streaming vs. Batch Pipelines

	Streaming Pipelines	Batch Pipelines
Execution Model	**Always running**, waiting to process work as it arrives.	**Spun up periodically** (on a schedule) or when sufficient work is available.
Data Handling	Consumes a **continuous stream** of data, processing each chunk as it's received and continually adjusting results.	Processes a large, finite batch of data at a specific point in time.
Primary Goal	**Low latency** and **reliability**. Designed for very fast, real-time analysis.	**High throughput** & **resource utilization**. Designed to optimize cluster usage (e.g., >80%).

Major Pipeline Frameworks

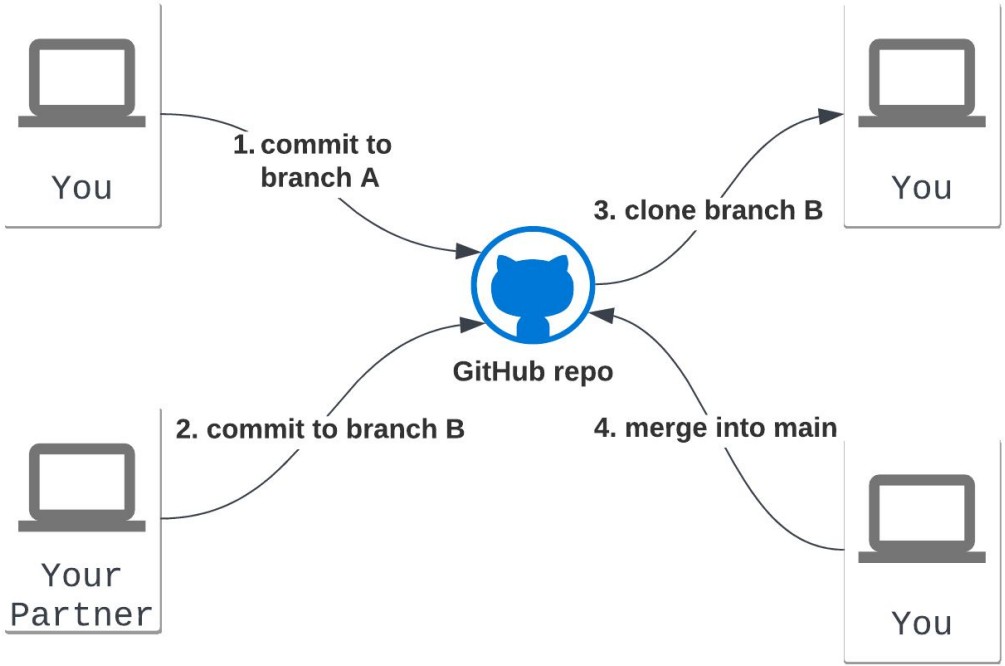
	Primary Role	**Data Model**	**Trade Offs**
Spark	Unified processing	Batch & micro-batch streaming	+ Large-scale batch ETL - Streaming is micro-batch
Flink	Stream Processing	True event-at-a-time streaming	+ SOTA low-latency streaming (ms) - Steep learning curve
Airflow	Orchestration	Batch-oriented (DAGs)	+ Industry standard for scheduling - Scheduler can be a bottleneck
dbt	In-Warehouse Transformation (ELT)	Batch (SQL-based)	+ Orchestrates SQL queries inside DW - Not a full data pipeline, only handles the T

dbt setup



Note: dbt Python models are translated into BigQuery Dataframes and run on Colab Enterprise via the dbt bigquery adaptor.

Development workflow



Let’s get started: setup guide