

Unit 3: Data Pipelines

Foundations of Data Warehousing (CS 378)

Oct 31, 2025

Happy Halloween

Data Pipeline Fundamentals

- **Definition:** A system 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).

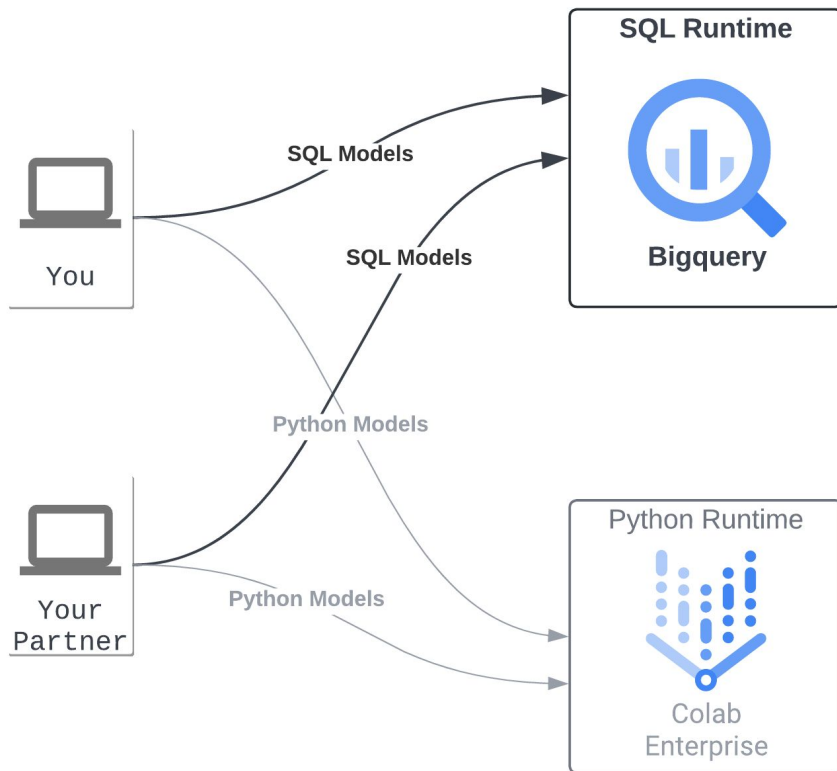
Streaming vs. Batch Pipelines

	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 & 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	Tradeoffs
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 Dev Setup



Call-to-Action - next 7 days:
Set up your dbt dev environment by
following the steps in our [guide](#).

Note: dbt Python models are translated to [BigQuery Dataframes](#) and run on [Colab Enterprise](#) via the dbt bigquery adaptor.

DBT Dev Workflow

