📦 Data Pipelines

Definition: A data pipeline is a series of processes that move and transform data from source to destination.

Why it matters: It automates data flow, improves efficiency, and ensures data is ready for analysis.

🔄 ETL Process (Extract, Transform, Load)

1. Extract

- **Purpose**: Pull data from various sources.
- **Techniques**: Batch ingestion (periodic loads) vs. Streaming ingestion (real-time updates).

2. Transform

- **Purpose**: Clean, enrich, and reshape data.
- **Examples**: Removing duplicates, formatting values, calculating new fields.

3. Load

- **Purpose**: Store transformed data in a destination (e.g., BigQuery).
- **Includes**: Validation, monitoring, and ensuring schema compatibility.

🧪 Validation Techniques

- **Duplicate Validation**: Ensures uniqueness of records.
- **Format Validation**: Checks if data follows expected patterns.
- **Type Validation**: Verifies correct data types (e.g., INT64, STRING).
- **Null Validation**: Identifies missing or incomplete data.

🌐 Geographic Data Transformation

- **ST_GEOGPOINT**: Converts latitude and longitude into a geographic point.

- **ST_DISTANCE**: Calculates distance between two geographic points (e.g., customer and distribution center).

🧰 Stored Procedures

- **Definition**: A reusable block of SQL statements stored in BigQuery.
- **Benefits**: Simplifies updates, supports schema changes, and enables automation.

📅 Scheduled Queries

- **Purpose**: Automate recurring data updates.
- **Setup**: Can be configured directly in the BigQuery Query Editor.

🎯 Business Application

- Use pipelines to solve real-world problems like:
  - Optimizing delivery routes.
  - Improving customer satisfaction.
  - Supporting logistics decisions with data.