Prepared by Pavani Korada

# *Databricks*

Databricks is a cloud platform used to work with big data and AI.

## Get started with Data Engineering

# *About Databricks:*

### UseCases

Used to:

- Handle Big Data Easily
- Faster AI/ML Development
- Unified Platform
- Real-Time Analytics

### In which domains Databricks Is used

- Healthcare
- Banking & Finance
- Telecom
- Marketing & Advertising
- Retail & E-commerce

### Which type of companies use databricks??

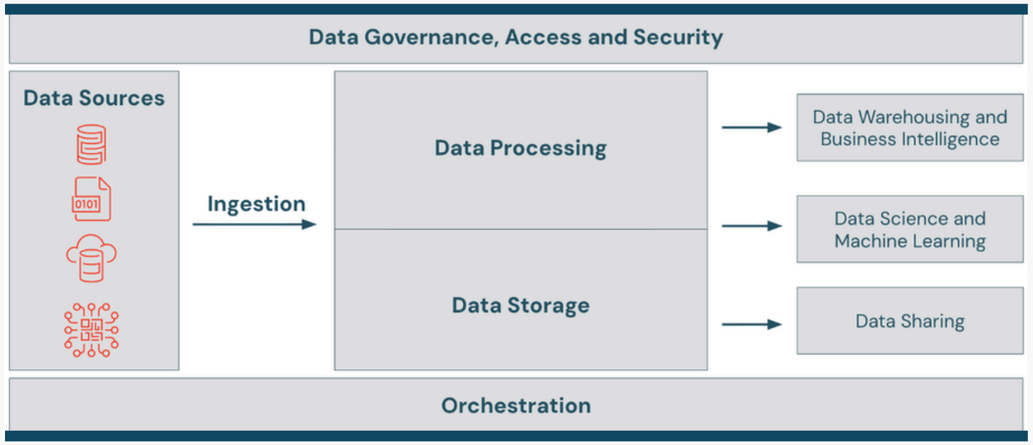
- Technology & Digital-Native Firms
- Retail & Consumer Goods
- Financial Services & Banking
- Manufacturing & Automotive
- Healthcare & Life Sciences

# Data Engineering Basics

| | | |
|---|---|---|
| Transforming raw data to clean, reliable data | Ensuring the quality and integrity of data | Design, build, and maintain data pipelines |

## Data Engineer Responsibilities

- Transform raw data into clean, structured, reliable data
- Perform data extraction, cleansing, and transformation (ETL)
- Ensure data quality, accuracy, and integrity
- Design, build, automate, and maintain data pipelines




## Data Engineering Architecture

- Data Sources – Databases, cloud storage, logs, files
- Data Ingestion – Load data into data lakes/warehouses
- Data Processing – Clean and transform data
- Data Availability – Provide data for ML and analytics
- Orchestration & Governance – Automate workflows, manage security and access

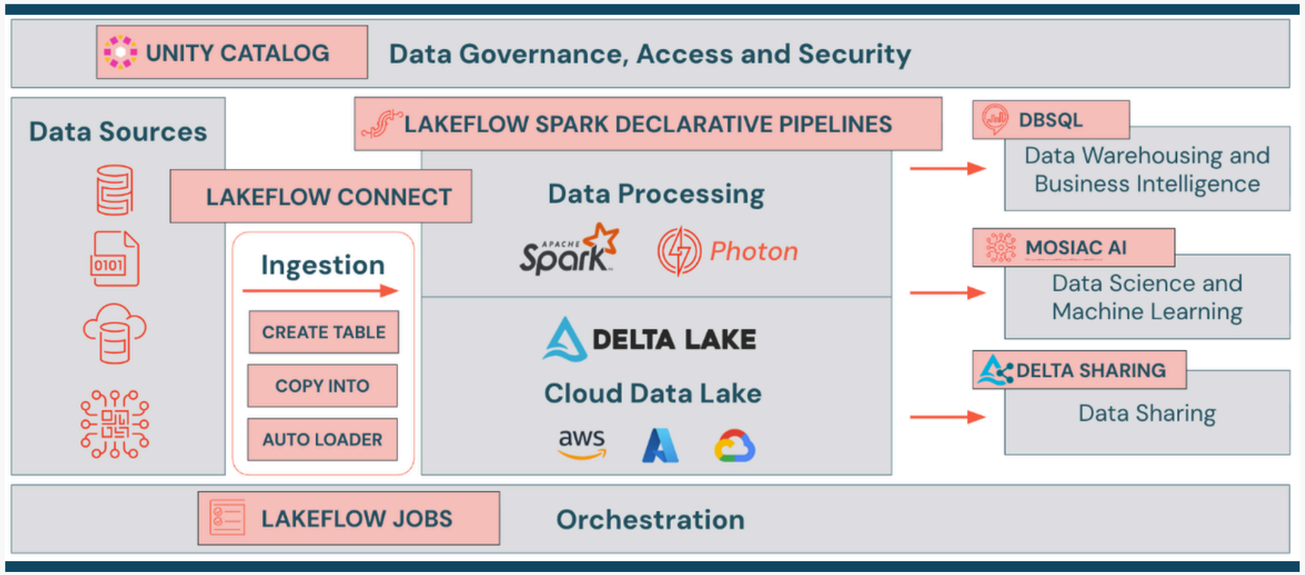
## Common Challenges:

**Complex data ingestion methods**

**Managing key data engineering principles**

---

To address these challenges, many organizations are turning to unified platforms like Databricks.




## Databricks Data Intelligence Platform (Solution)

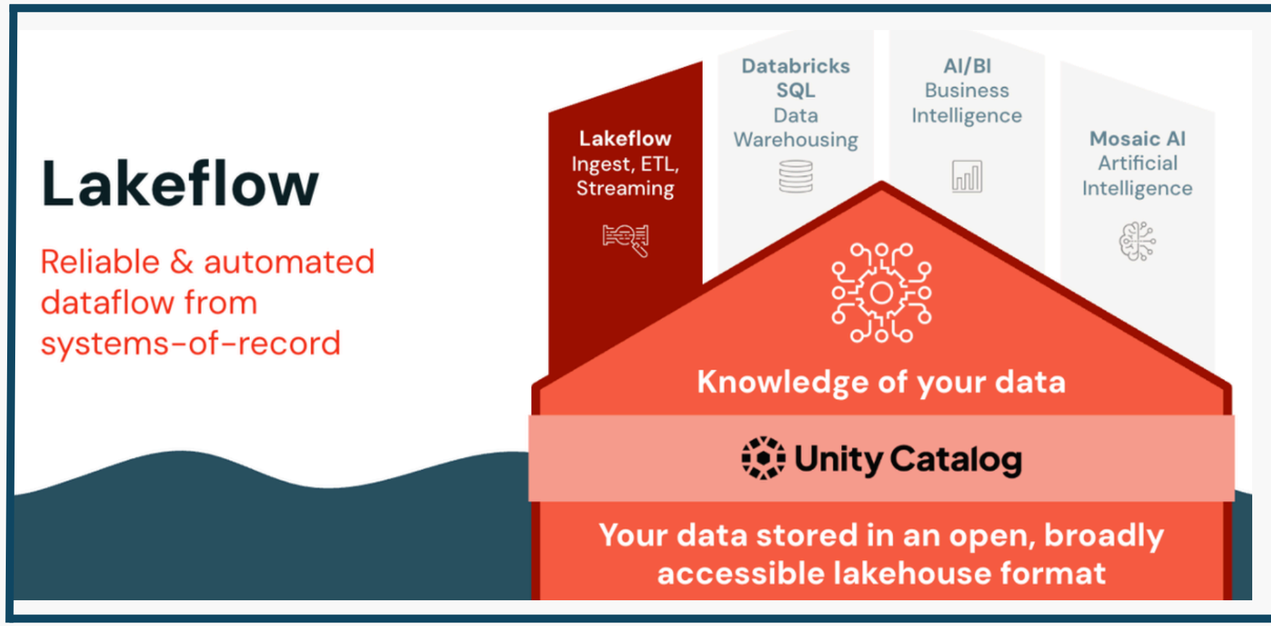
- Unified platform for ingestion, processing, storage & analytics
- Delta Lake for reliable lakehouse architecture
- Lakeflow for automated ETL & orchestration
- DBSQL (BI), Mosaic AI (ML), Delta Sharing (data sharing)
- Unity Catalog for governance, security & access control

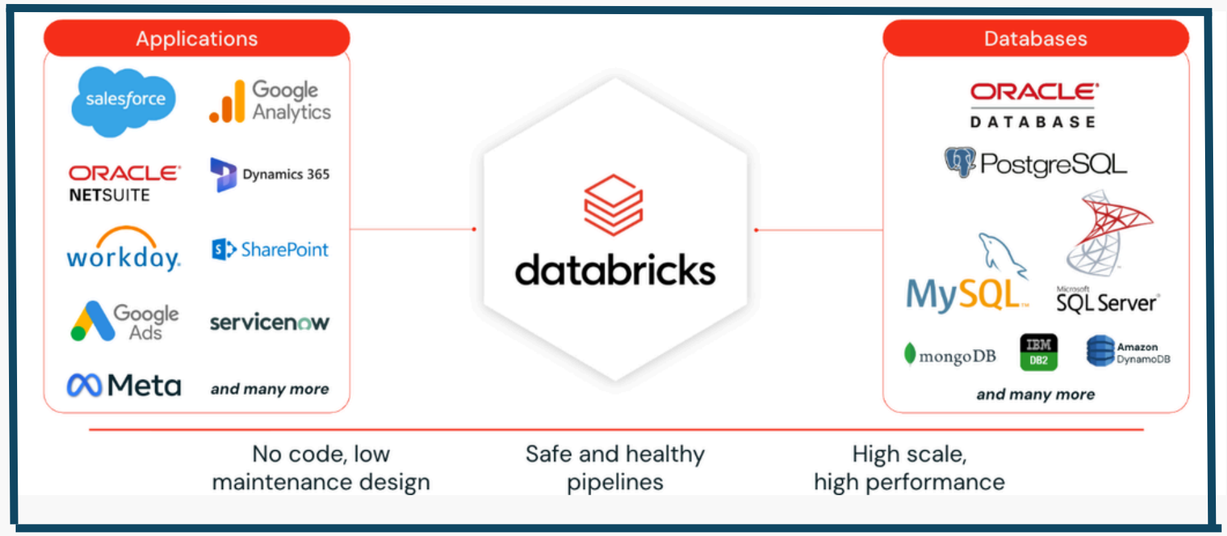
# *Intro to Lakeflow*

Lakeflow is a unified set of tools in Databricks that manages data from ingestion to delivery in a reliable, automated way.

**Lakeflow consists of three powerful components:**

- Lakeflow Connect
- Lakeflow Spark
- Lakeflow Jobs




## Lakeflow Connect (Ingestion)

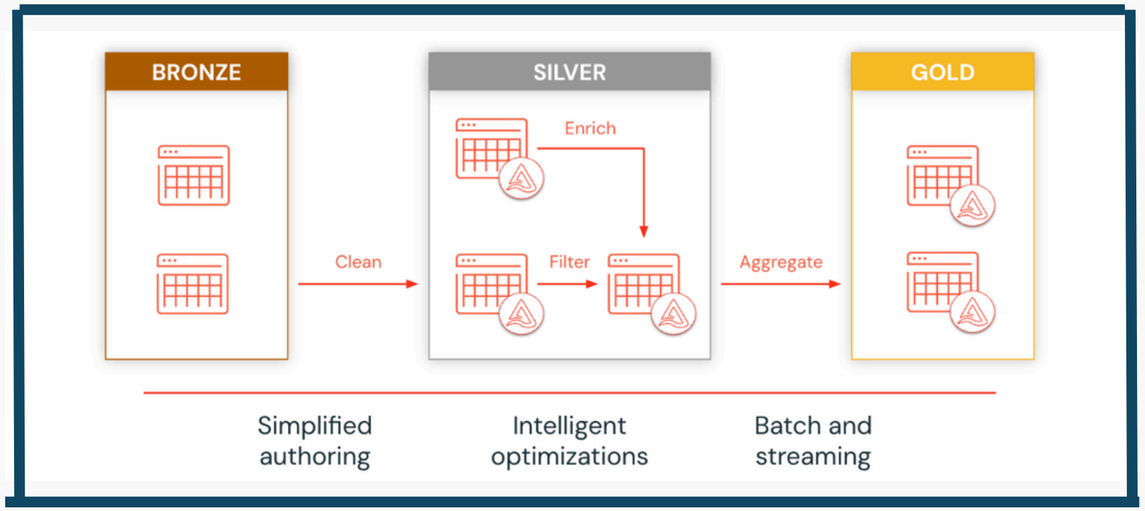
- No-code & low-code data ingestion
- Managed and standard connectors
- Ingest from applications, databases, and third-party systems
- Supports batch and streaming data

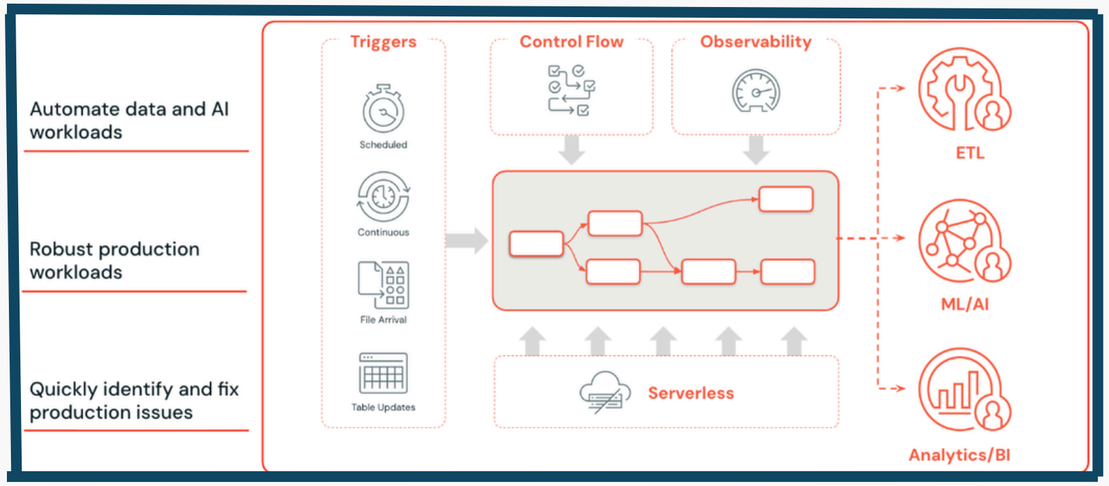
# Lakeflow Spark Declarative Pipelines (ETL & Transformation):

- Build automated ETL pipelines
- Follows Medallion Architecture:

1. **Bronze** – Raw data
2. **Silver** – Cleaned & validated data
3. **Gold** – Aggregated, business-ready datasets

- Ensures data quality and reliable transformations




# Lakeflow Jobs (Orchestration):

- Automates workflows with task dependencies
- Supports triggers (scheduled, continuous, file-based, etc.)
- Enables ETL, ML/AI, and Analytics workloads
- Serverless and production-ready

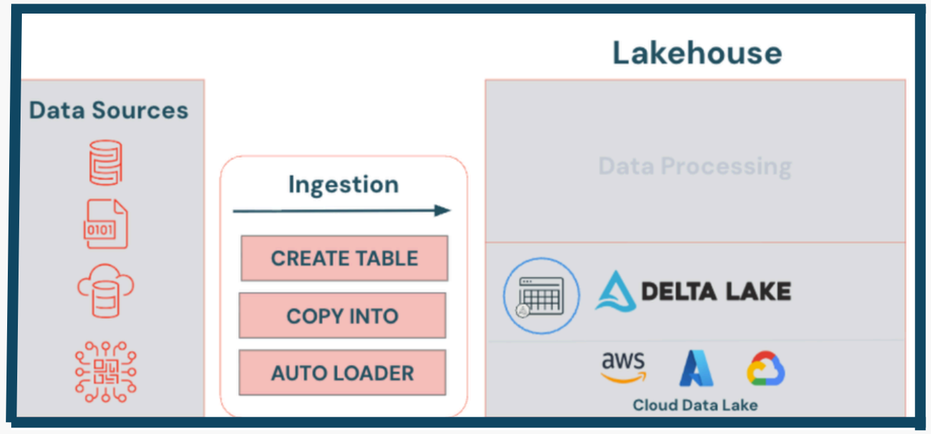
# *Ingestion with Lakeflow Connect*

## Upload via UI

- Upload CSV, JSON, Avro, Parquet, Text files directly
- Best for quick analysis or small datasets
- Not suitable for large-scale or streaming workloads

## COPY INTO

- Loads files from cloud storage into Delta tables
- Idempotent – skips already ingested files
- Supports multiple formats and schema changes
- Good for regular incremental loads




**syntax:**

```
CREATE TABLE mydeltatable
USING DELTA    -- Optional
AS
your query
```

**syntax:**

```
COPY INTO mydeltatable
FROM 'your-path'
FILE_FORMAT = 'format'
FILE_OPTIONS = ('format-options')
```

## CREATE TABLE AS (CTAS)

- Creates a Delta table from a SELECT query
- Delta format is default (USING DELTA)
- Ideal for full batch loads (replacing table each time)

## Auto Loader

- Automatically ingests new files as they arrive
- Highly scalable for large datasets
- Handles schema evolution
- “Rescues” unexpected or mismatched data

*Thank you*