DATA PIPELINE & ORCHESTRATION TOOLS

Group-5
Present By – Muskan & Ronak

Data Pipeline Introduction

A **data pipeline** is a sequence of steps that move data from **source to destination**, ensuring it is **cleaned, transformed, and stored** for analysis.

Extract

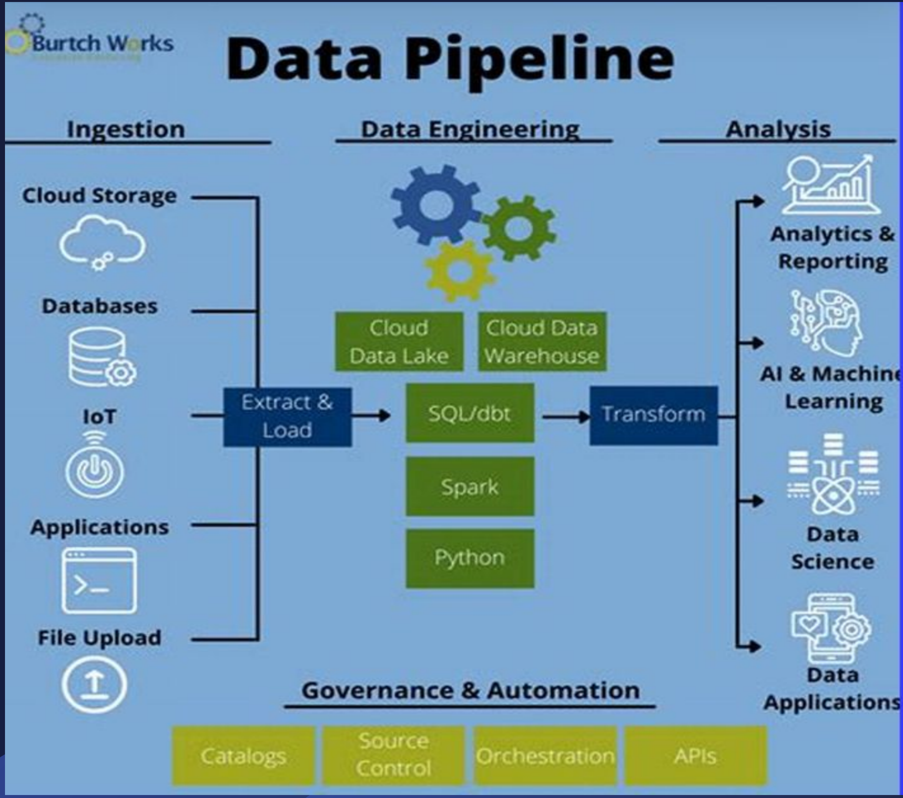
1

2

3

Load

Transform



Why Companies Need Data Pipelines

01 Data Volume & Velocity

02 Reliability & Error Handling

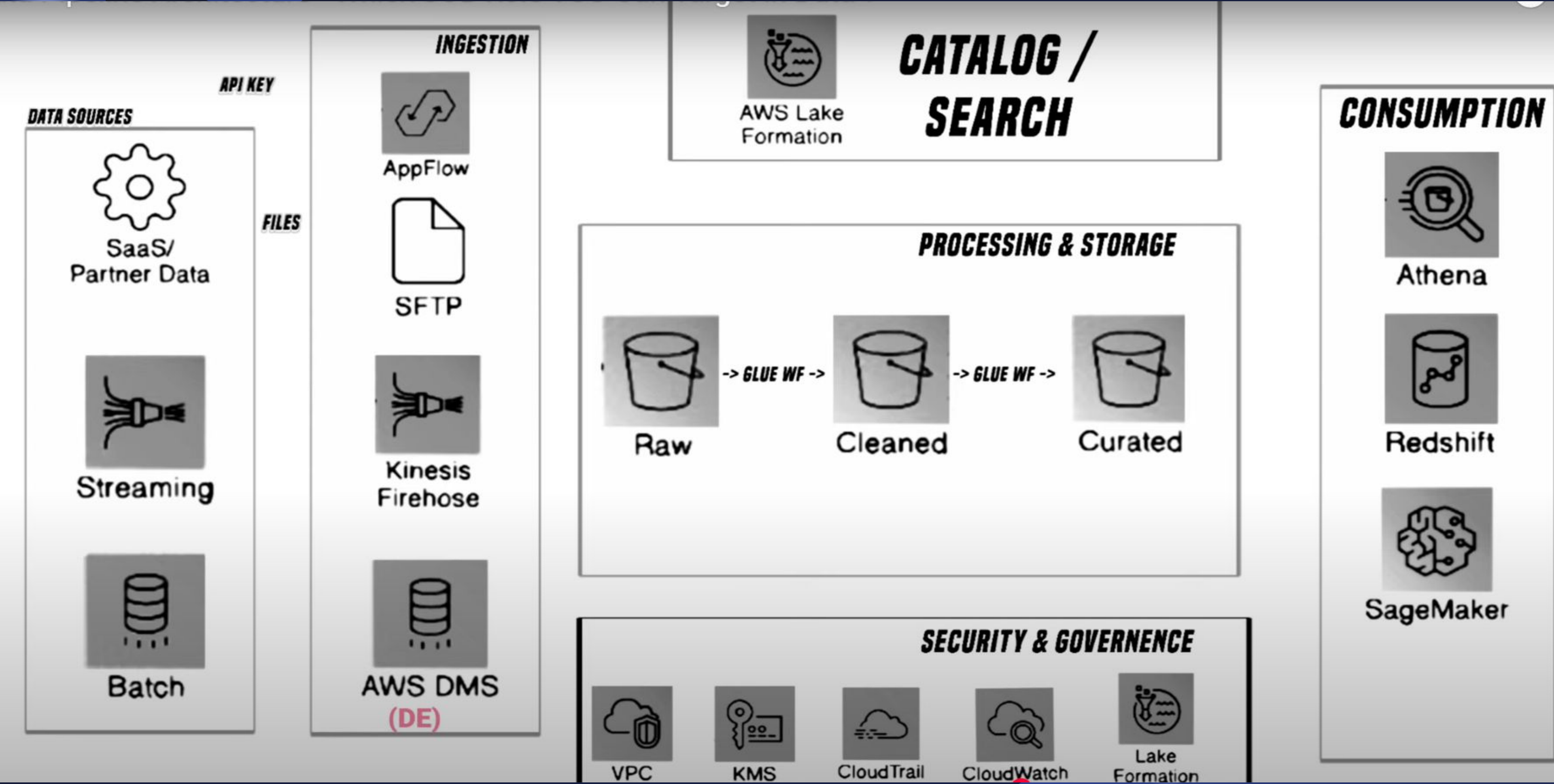
03 Scalability

04 Cost Optimization

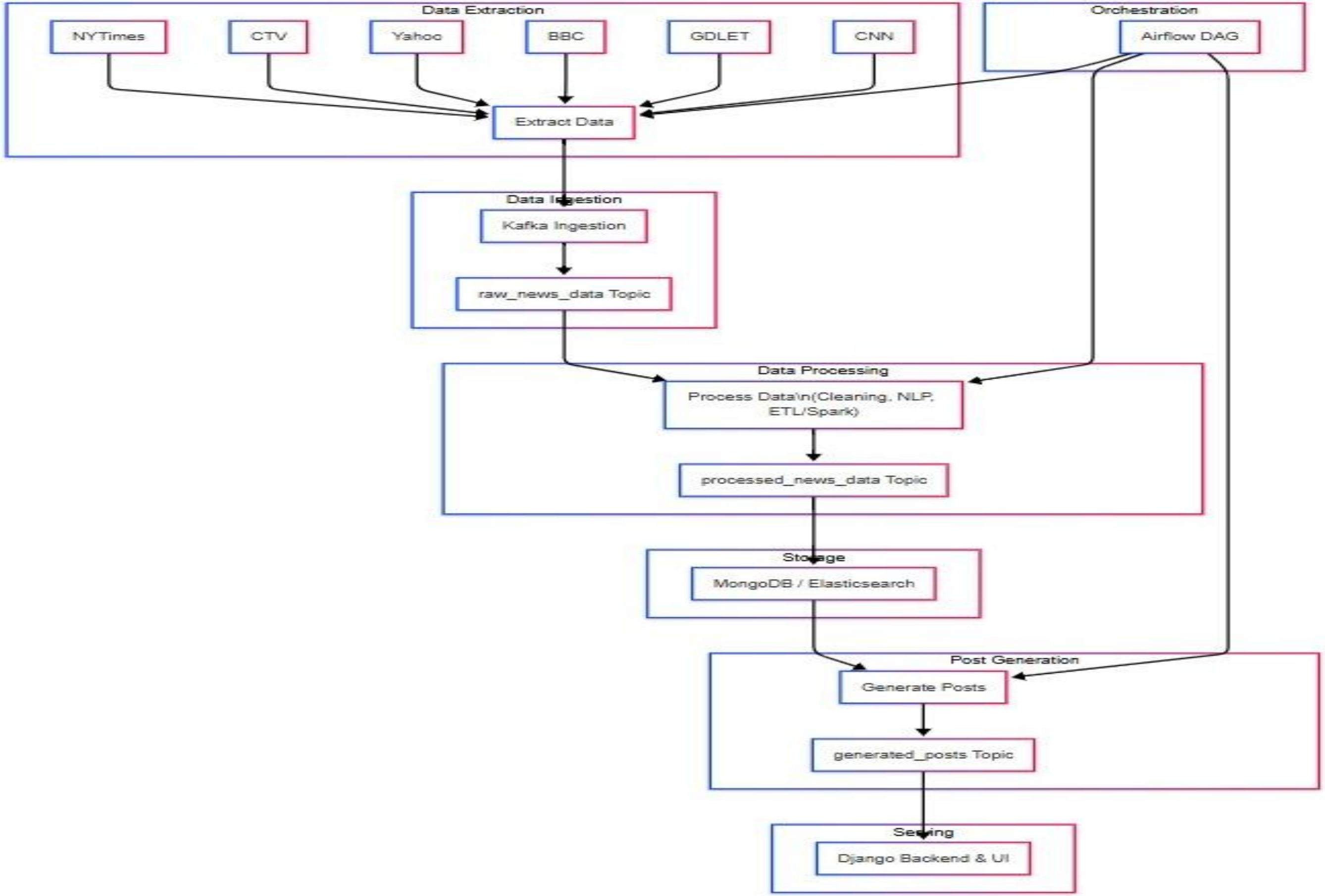
Data Pipeline Tools & Technologies

Pipeline Type	Tools	Best Use Case	Real-World Example
Batch Processing	Apache Spark, AWS Glue, Airflow	Historical analytics, ETL workflows	**Netflix** uses batch Spark jobs to analyze user watch history overnight
Streaming (Real-Time)	Apache Kafka, Flink, AWS Kinesis	Fraud detection, live recommendations	**Uber** processes real-time ride requests & traffic data using Kafka + Flink
Hybrid Processing	Spark Structured Streaming, Kafka Streams	Personalized user experiences	**Amazon** combines batch ETL for inventory and real-time streaming for order tracking

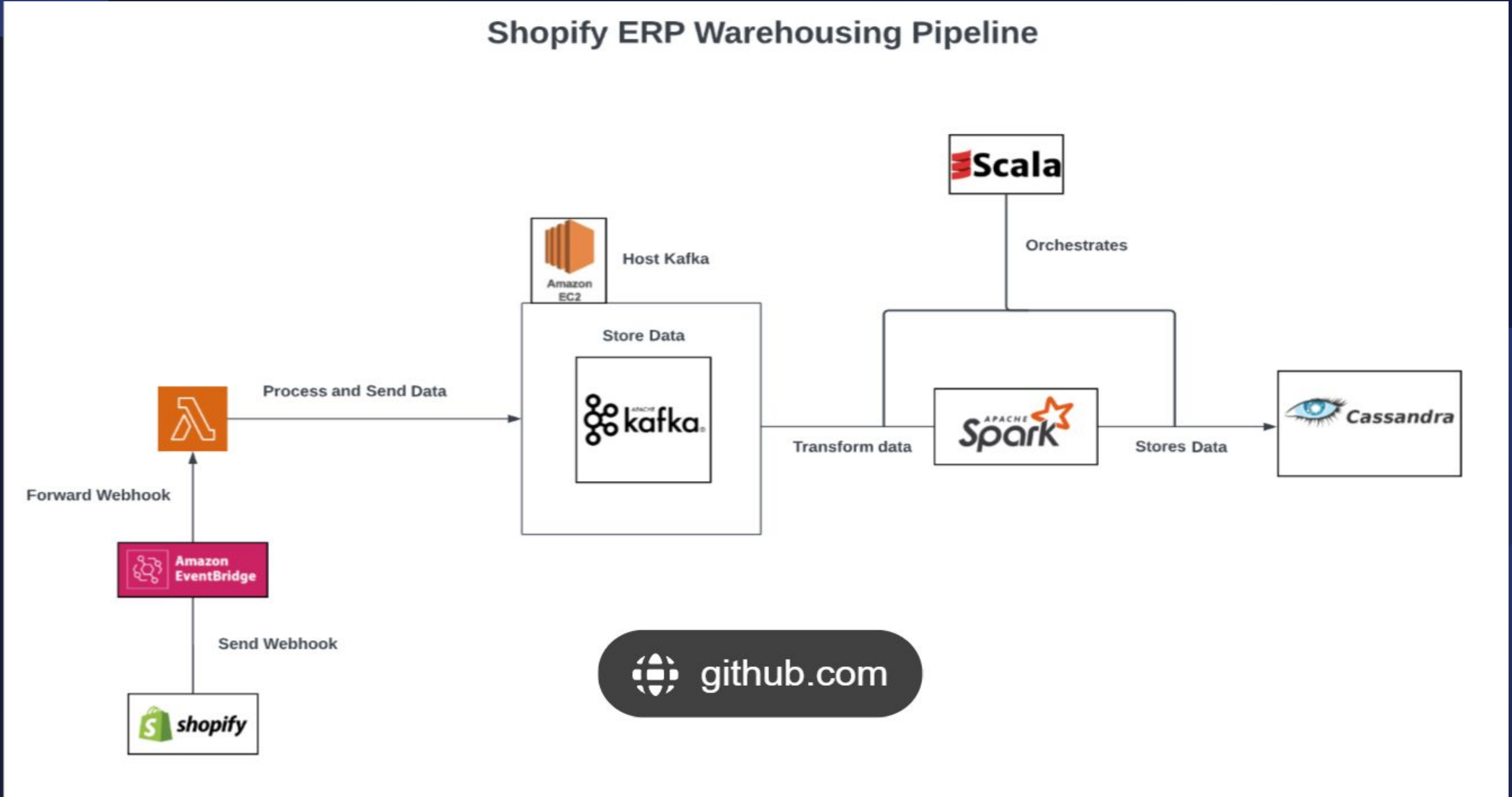
AWS Data Pipeline Architecture



Apache Kafka



Apache Spark



Shopify ERP Warehousing Pipeline

Orchestration Introduction

Orchestration is the **automated coordination, scheduling, and management** of tasks, workflows, and data pipelines across multiple systems.

🔹 **Why is Orchestration Important?**

✅ **Automation**
✅ **Scalability**
✅ **Dependency Management**
✅ **Error Handling & Recovery**

🔹 **Types of Orchestration**

1️⃣ **Workflow Orchestration** – Automates ETL, ML workflows (e.g., Airflow, Prefect).
2️⃣ **Container Orchestration** – Manages Dockerized apps (e.g., Kubernetes).
3️⃣ **Cloud Orchestration** – Automates cloud services (e.g., AWS Step Functions).
4️⃣ **Data Orchestration** – Manages large-scale data movement (e.g., Dagster).

Apache Airflow

🔹 **What is it?**

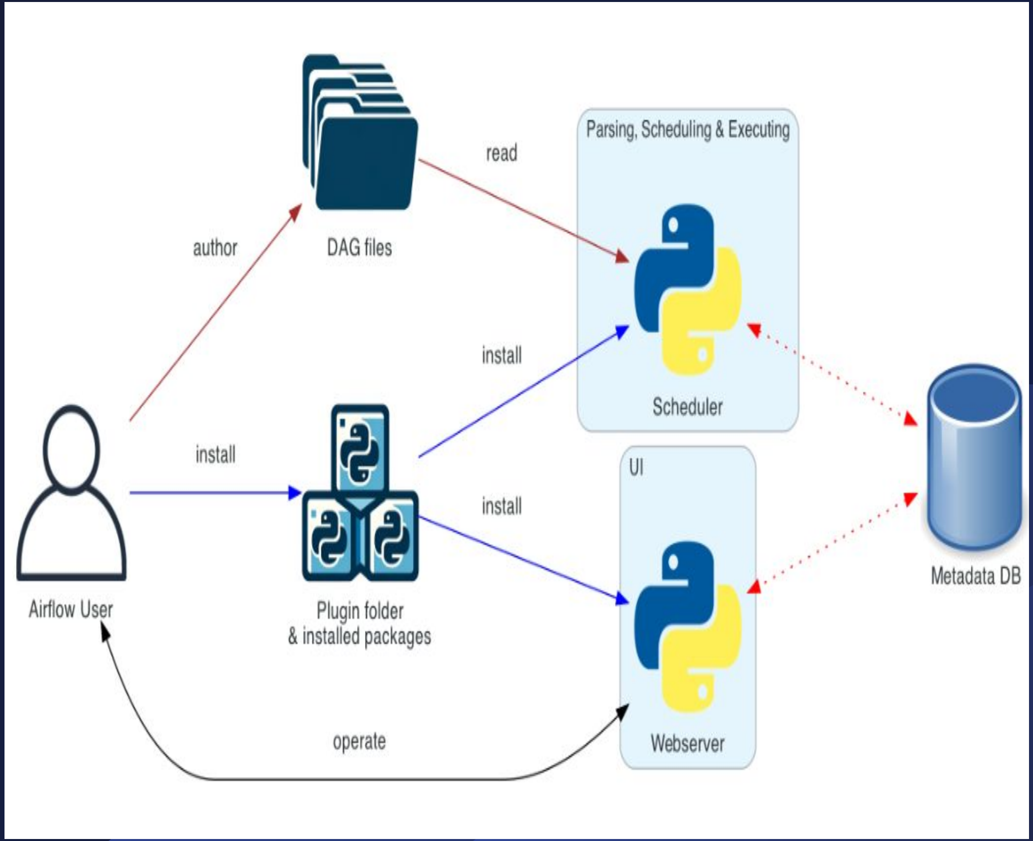
An **open-source workflow orchestration tool** that automates **ETL, data pipelines, ML workflows, and cloud automation** using **Directed Acyclic Graphs (DAGs).**

🔹 **Why Use It?**

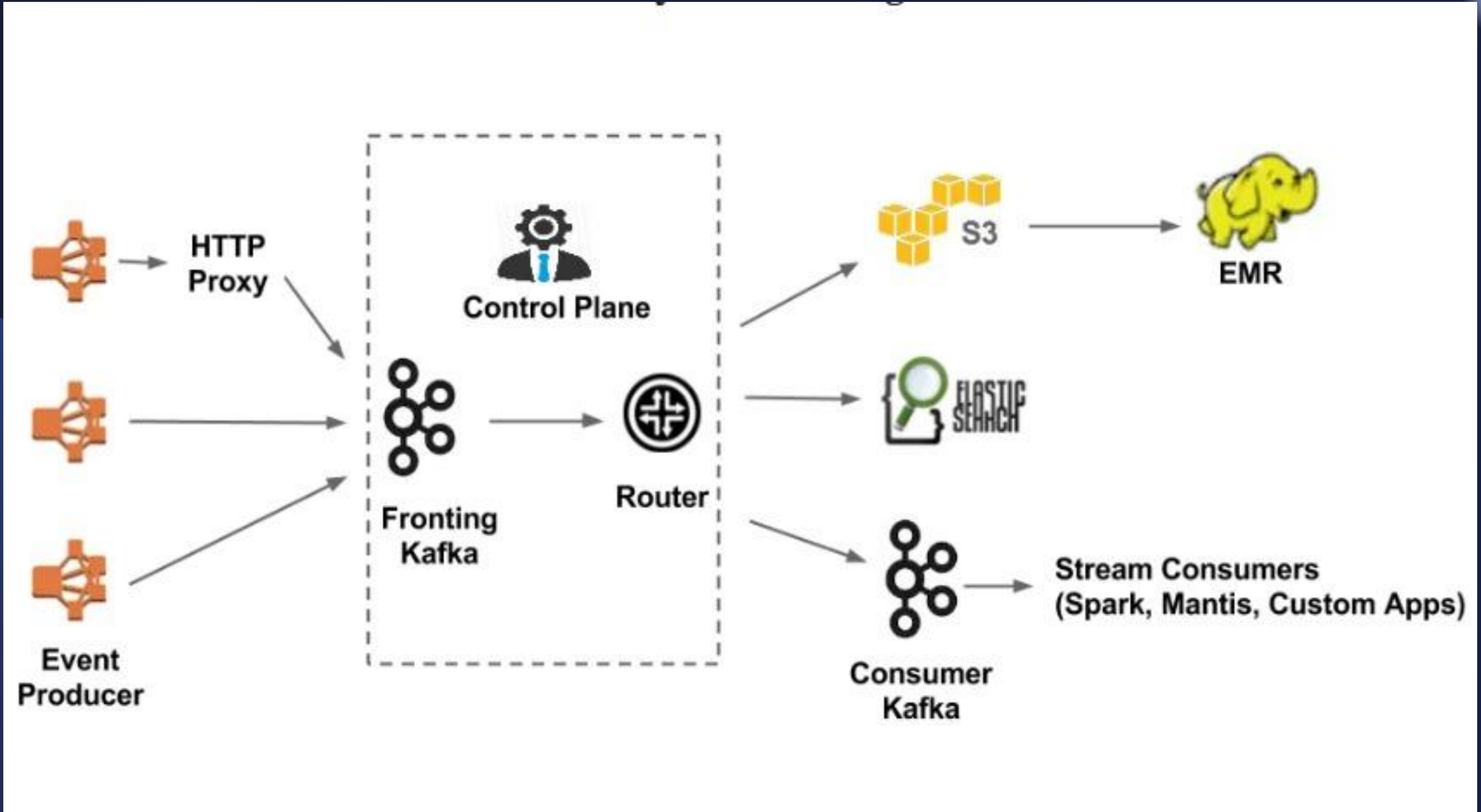
✅ **Scalable & Flexible**
✅ **Monitoring & Logging**
✅ **Integration Ready**

🔹 **Best Use Cases**

✔ **ETL & Data Pipelines**
✔ **Machine Learning Automation**
✔ **Cloud Automation**
✔ **Big Data Processing**



Netflix Real-Time Data Streaming Architecture



Docker – Containerization Platform

🔹 **What is Docker?**

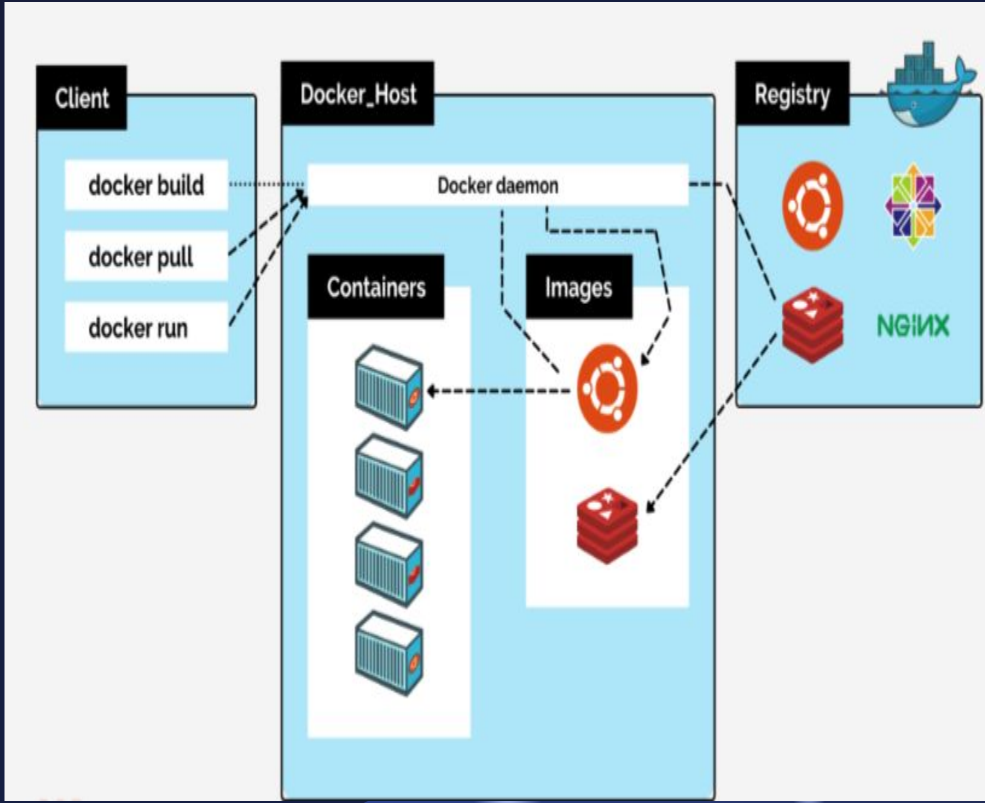
💡 **Think of Docker like a "shipping container" for apps.** It lets developers **package, deploy, and run applications consistently** across different environments.

🔹 **Why Use Docker?**

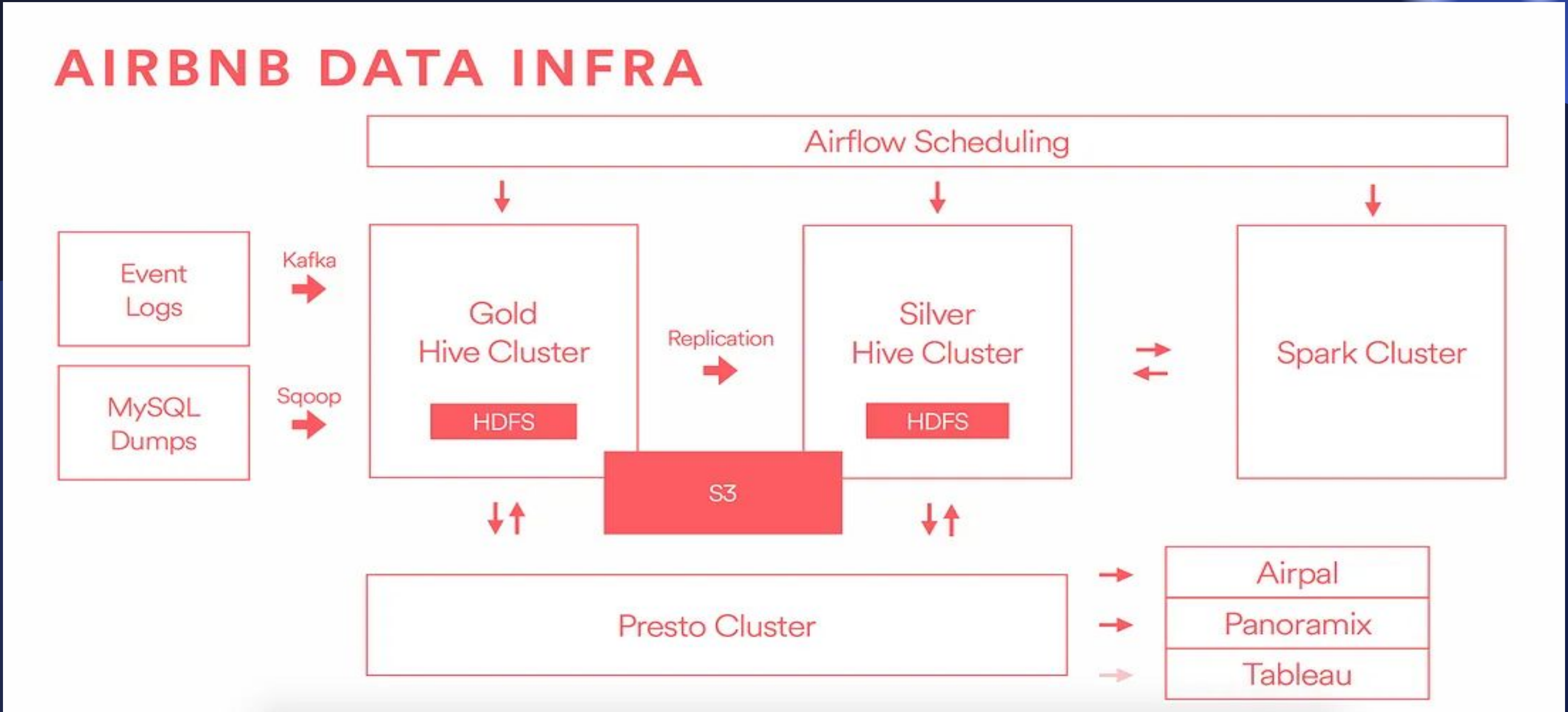
✅ **Portability**
✅ **Lightweight & Fast**
✅ **Scalability**
✅ **DevOps & CI/CD Integration**
✅ **Dependency Management**

🔹 **Best Use Cases**

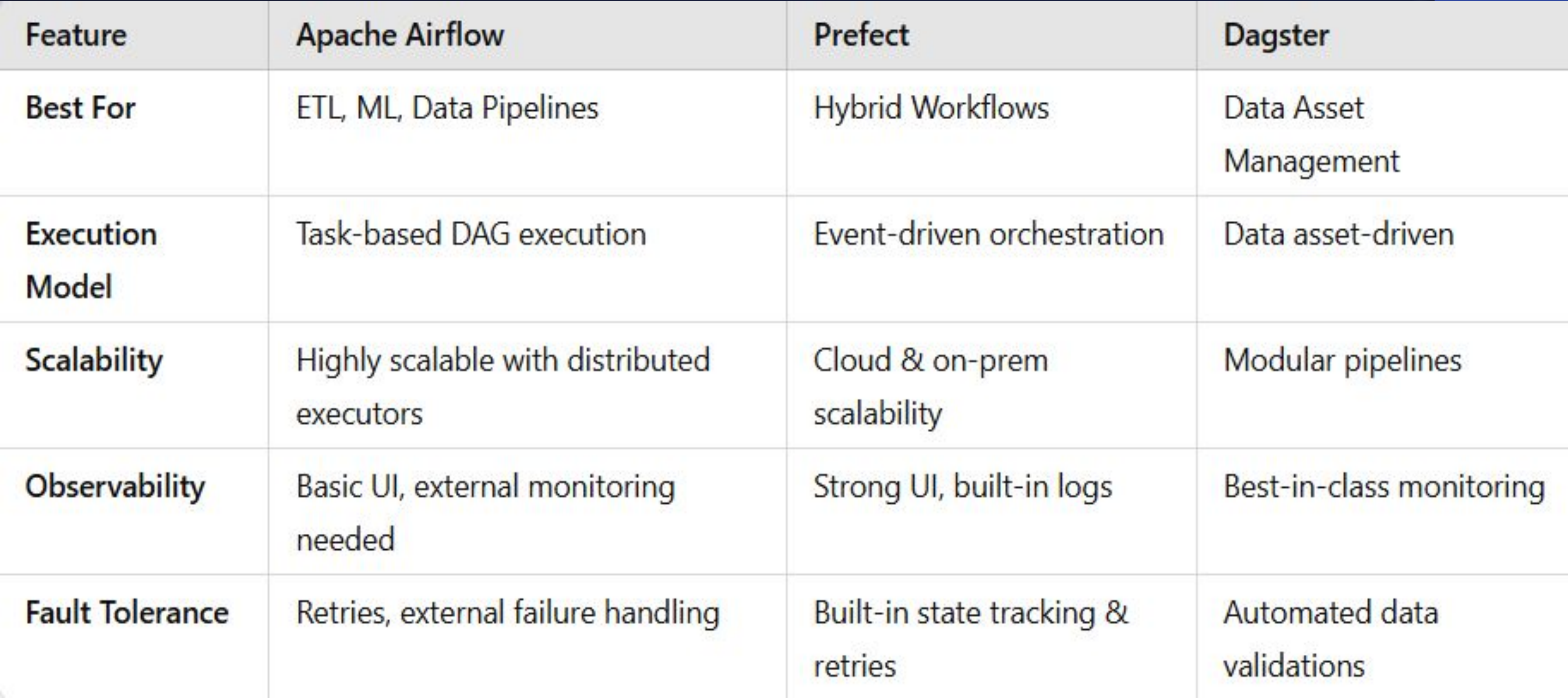
✔ **Microservices Architecture**
✔ **CI/CD Pipelines**
✔ **Cloud-Native Applications**
✔ **Big Data & AI Workloads**
✔ **Cross-Platform Development**



Airbnb Data Processing & Analytics Pipeline



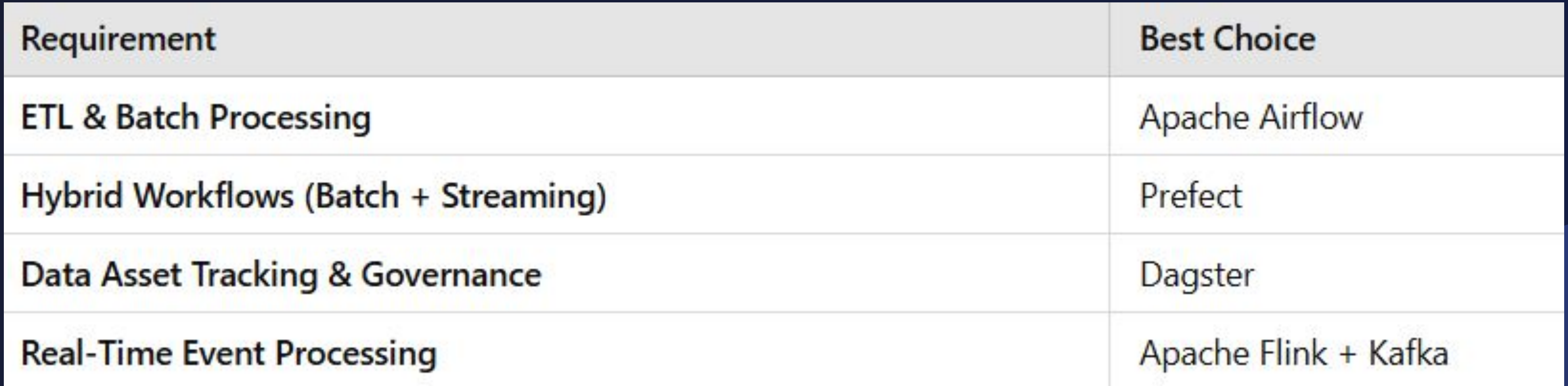
Orchestration Tools Comparison



Feature	Apache Airflow	Prefect	Dagster
Best For	ETL, ML, Data Pipelines	Hybrid Workflows	Data Asset Management
Execution Model	Task-based DAG execution	Event-driven orchestration	Data asset-driven
Scalability	Highly scalable with distributed executors	Cloud & on-prem scalability	Modular pipelines
Observability	Basic UI, external monitoring needed	Strong UI, built-in logs	Best-in-class monitoring
Fault Tolerance	Retries, external failure handling	Built-in state tracking & retries	Automated data validations

How to Choose the Right Orchestration Tool?

📌 Decision Framework for Selecting the Right Tool



Requirement	Best Choice
ETL & Batch Processing	Apache Airflow
Hybrid Workflows (Batch + Streaming)	Prefect
Data Asset Tracking & Governance	Dagster
Real-Time Event Processing	Apache Flink + Kafka

✅ Checklist for Decision Making:

- Do you need **real-time data streaming**? → **Use Kafka or Flink**
- Do you need **complex scheduling & task orchestration**? → **Use Airflow**
- Do you need **data lineage tracking**? → **Use Dagster**
- Do you need **cloud-friendly automation**? → **Use Prefect**

Future Trends in Data Pipelines & Orchestration

AI-Driven

Automating pipeline optimizations using ML.

Serverless

Adoption of AWS Glue, Google Dataflow, Snowflake.

Hybrid/Multi-Cloud

Tools like Dagster & Prefect enable cross-cloud workflows.

Data Mesh

Rise of domain-driven data architectures.
Example-Enterprises use hybrid cloud pipelines for batch & streaming data across AWS, Azure, and GCP.

Conclusion - Why Data Pipelines & Orchestration Matter

Data Pipeline

1. Data pipelines automate & scale data processing, making analytics and AI workflows more efficient.
2. Choosing batch vs. streaming pipelines depends on latency requirements & data volume.

Orchestration

1. Orchestration tools (Airflow, Prefect, Dagster) optimize workflow execution, improve reliability, and reduce errors.
2. Companies like Netflix, Uber, and Airbnb leverage hybrid models for better efficiency.
3. Future of orchestration is AI-driven automation, serverless architectures, and multi-cloud processing.

Summarize

"Data pipelines are the backbone of modern data-driven businesses. Choosing the right approach & orchestration tool can transform how your organization handles data. Are you ready to optimize your workflows?"

Thanks!_

Do you have any questions?

Email Us. Will try to get back to you in 24 hrs.

talk2group5@group5.com

group5.com