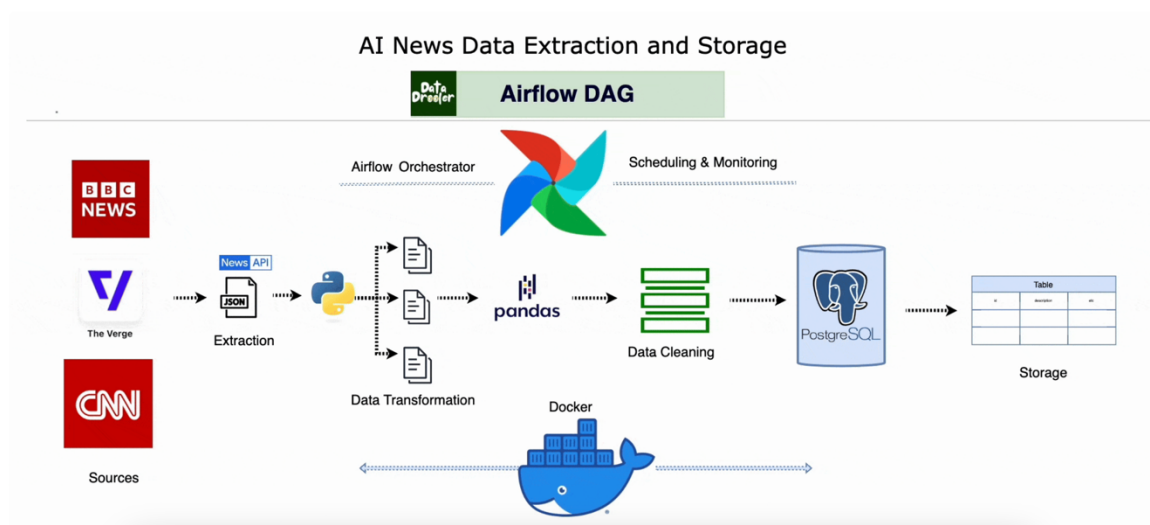


# Week 1

## Intro + Building DAGs using Airflow

 by Sunjana Ramana

### What We Will Build Today



# Agenda for Today

## P1: Data Engineering Overview

- What do Data Engineers do?
- RoadMap
- Data Engineer's Jargon
- How to get Started
- ETL & ELT Pipeline
- Rating TechStack

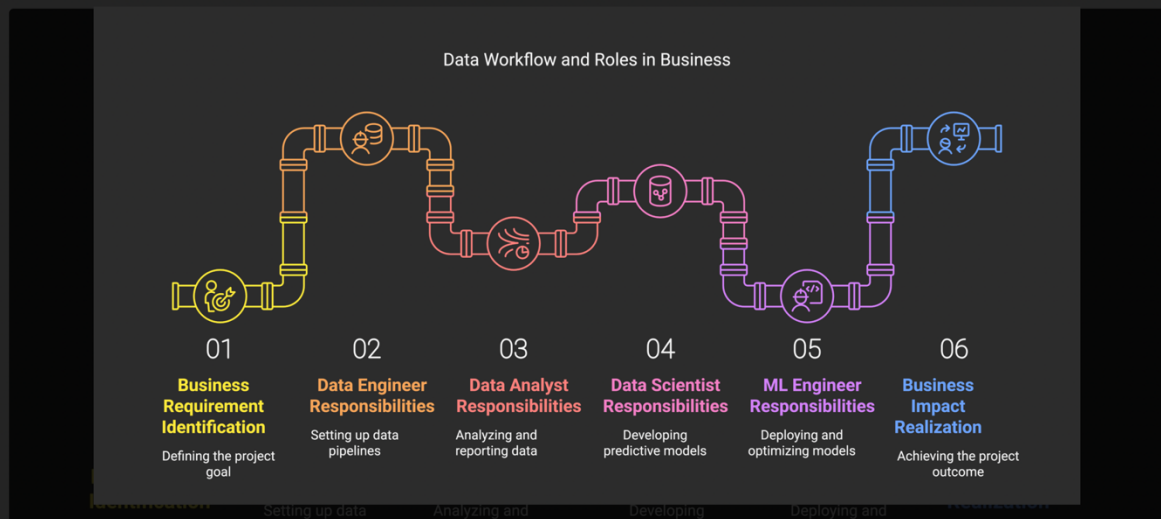
## P2: Building DAGs on Airflow - Overview

- Intro to Airflow
- Docker
- Postgres and pgAdmin
- What is DAG

## P3: Building our first DAG - Implementation

- Extracting data from NewsAPI
- Transforming data using Pandas
- Loading Data into Postgresql using SQL
- Scheduling, Monitoring and Logs
- Sending emails to email using SendgridAPI

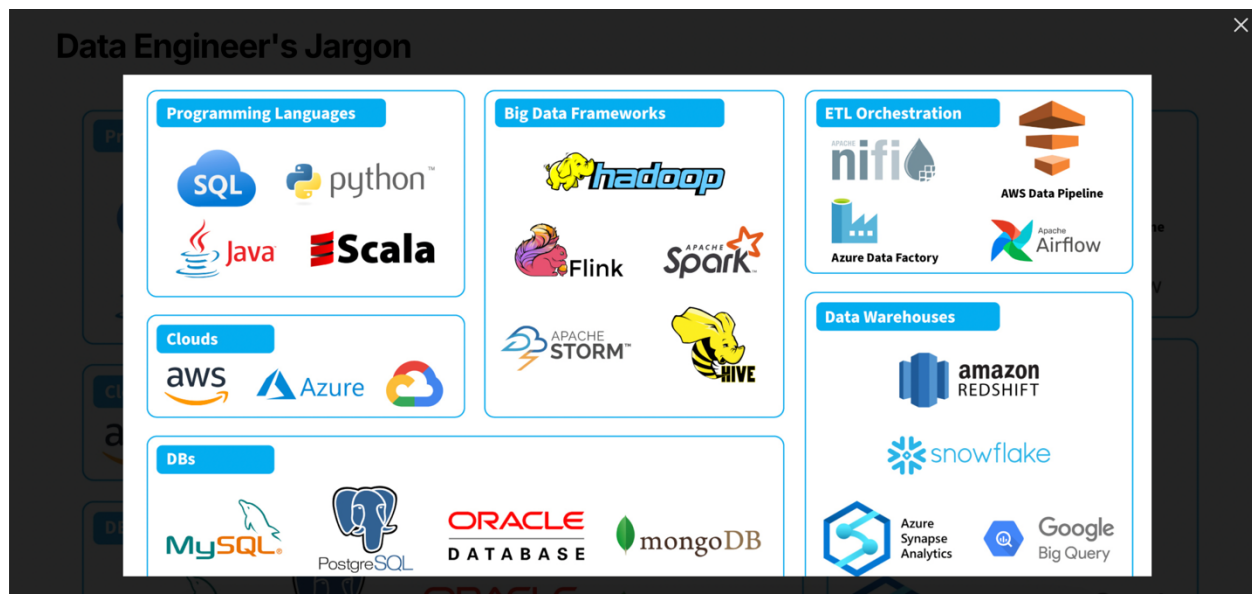
## What do Data Engineer's do?































Feature/Aspect	Data Warehouse	Database	Data Lake
Purpose	Analytical processing (OLAP)	Transactional processing (OLTP)	Store raw, unstructured, and structured data
Data Type	Structured data (tables)	Structured (SQL) and unstructured (NoSQL)	Structured, semi-structured, unstructured
Use Case	Business intelligence, reporting, analytics	Operational systems, transactions, apps	Big data, machine learning, real-time analytics
Storage	Optimized for fast queries and reporting	Optimized for read-write operations	Stores raw data at scale
Data Processing	ETL (Extract, Transform, Load)	Real-time transactional updates	Schema-on-read, raw data processing
Tools	Snowflake, BigQuery, Redshift, Teradata	MySQL, PostgreSQL, MongoDB, SQL Server	S3, Azure Data Lake, Hadoop, Databricks

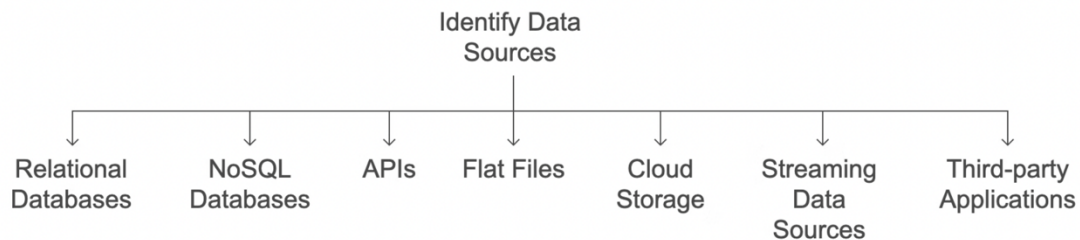
Feature	SQL (Relational)	NoSQL (Non-Relational)
<b>Data Structure</b>	Tables (Rows & Columns)	JSON, Key-Value, Graph, Column
<b>Schema</b>	Fixed & Predefined	Flexible & Schema-less
<b>Scalability</b>	Vertical (Scale Up)	Horizontal (Scale Out)
<b>Transactions</b>	ACID-Compliant	BASE (Eventual Consistency)
<b>Best for</b>	Structured Data	Unstructured & Big Data
<b>Example Use Cases</b>	Banking, ERP, CRM	Social Media, IoT, Real-time Apps



How to get Started?		
low Effort, high Usage	low Effort, high Usage	      
high Effort, high Usage	high Effort, high Usage	       
		  
low effort, low Usage	low effort, low Usage	 
high Effort, low Usage	high Effort, low Usage	  
dead tools	dead tools	  

## Building Data Pipelines ( ETL & ELT)

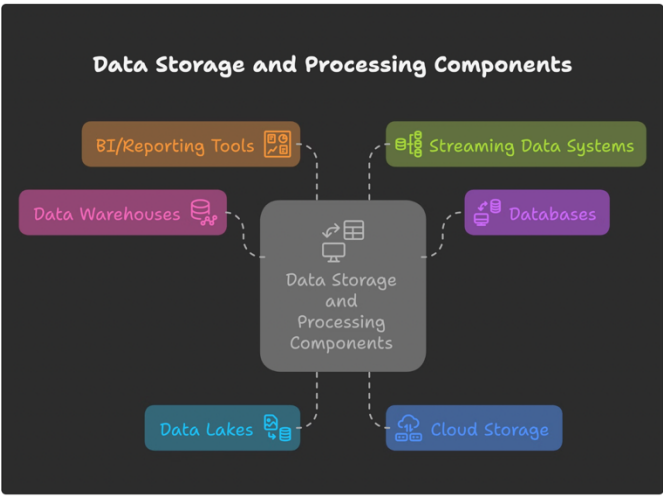
### Data Source Extraction in ETL Processes



Data Transformation Process



Data Storage and Processing Components



Feature/ Aspect	Data Warehouse	Database	Data Lake
Purpose	Analytical processing (OLAP)	Transactional processing (OLTP)	Store raw, unstructured, and structured data
Data Type	Structured data (tables)	Structured (SQL) and unstructured (NoSQL)	Structured, semi-structured, unstructured
Use Case	Business intelligence, reporting, analytics	Operational systems, transactions, apps	Big data, machine learning, real-time analytics
Storage	Optimized for fast queries and reporting	Optimized for read-write operations	Stores raw data at scale
Data Processing	ETL (Extract, Transform, Load)	Real-time transactional updates	Schema-on-read, raw data processing
Tools	Snowflake, BigQuery, Redshift, Teradata	MySQL, PostgreSQL, MongoDB, SQL	S3, Azure Data Lake, Hadoop, Databricks

# Intro to Apache Airflow



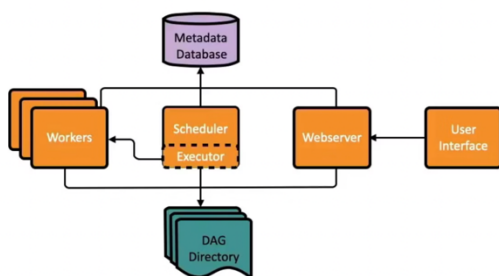
**Launched in 2014:** Airflow was created by Airbnb as an internal tool to manage and schedule workflows, primarily for data pipelines.

**Open Sourced in 2015:** After recognizing its potential, Airbnb released Airflow as an open-source project under the Apache Software Foundation.

**Top-Level Apache Project (2019):** Airflow became a top-level Apache project, marking its success and widespread adoption in the tech community.

**Used by Leading Companies:** Apache Airflow is used by major tech companies like **Airbnb, Spotify, Uber, Google, Slack, Netflix, Pinterest, Yahoo, Zendesk**, and **Turo**, among others, for orchestrating and managing complex data pipelines and workflows.

## Airflow Architecture



# Docker



- **Containerization Technology:** Docker is a platform for developing, shipping, and running applications in lightweight, portable containers that package an app and its dependencies together.
- **Simplifies Development and Deployment:** It ensures consistency across various environments (development, staging, production) by using container images, which work the same way everywhere.

Feature	Docker	Virtual Machines (VMs)
Architecture	Containers share the host OS kernel	Each VM has its own full OS with its own kernel
Resource Efficiency	More efficient, fewer resources used	More resource-intensive (each VM needs a full OS)
Performance	Faster performance, lower overhead	Slower due to full OS and virtualization overhead
Portability	Highly portable, runs anywhere	Less portable, depends on hypervisor and OS
Isolation	Process and file system isolation, shares host kernel	Strong isolation with separate OS and kernel for each VM

What are containers, images and volumes

Feature	Image 🖼️	Container 🐳	Volume 📁
What is it?	A template for containers	A running instance of an image	A storage method for persistent data
Mutable?	❌ No (read-only)	✅ Yes (changes happen here)	✅ Yes (stores data permanently)
Use case	Defines the environment	Runs applications	Stores data shared between containers

## PostgreSQL and pgAdmin



### PostgreSQL:

1. **Open-source RDBMS:** PostgreSQL is a powerful, open-source relational database management system (RDBMS).
2. **ACID Compliant:** It ensures reliability with support for ACID (Atomicity, Consistency, Isolation, Durability) properties.
3. **SQL and NoSQL:** Supports both SQL and NoSQL queries, offering flexibility in data handling.

### pgAdmin:

- **Open-source GUI:** pgAdmin is a web-based graphical user interface (GUI) for managing PostgreSQL databases.
- **Database Management:** It allows users to manage database objects like tables, schemas, and queries.
- **Cross-Platform:** pgAdmin works across multiple platforms (Windows, macOS, Linux).