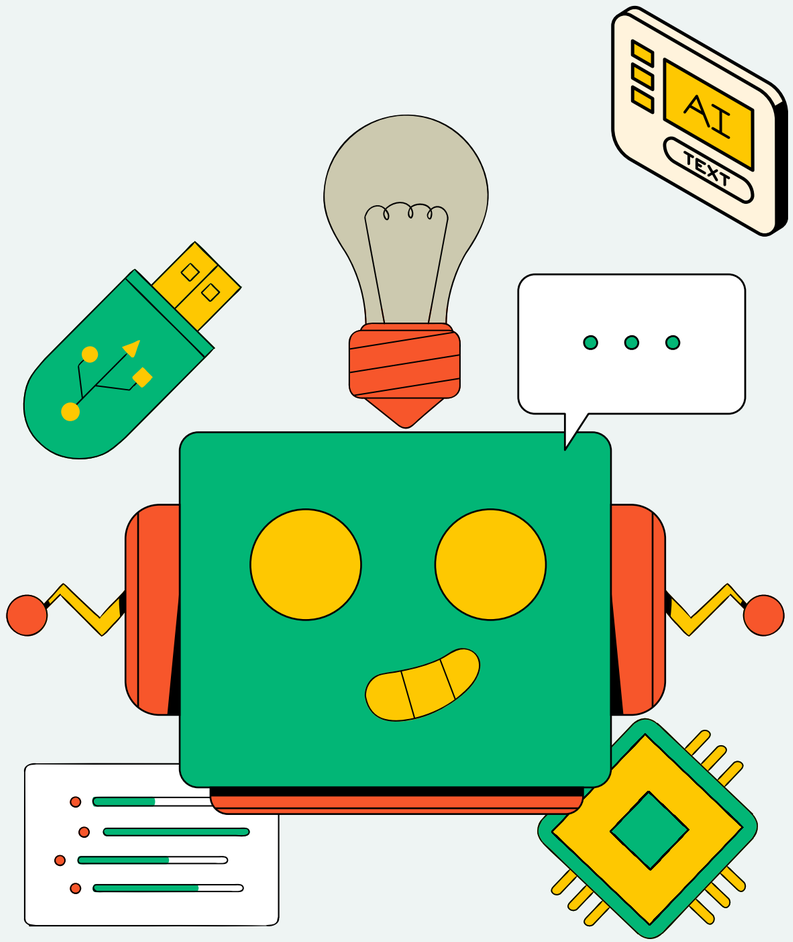
BIG DATA STORAGE AND PROCESSING

GROUP 11



UNITED STATES FLIGHTS DATA

PRESENTATION

LECTURER: MR. TRAN VIET TRUNG

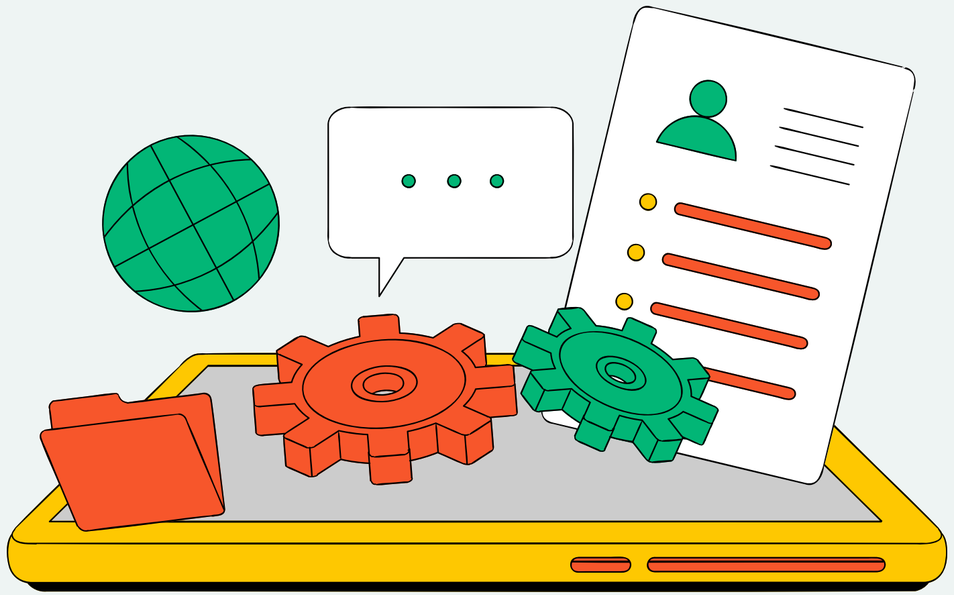
VU VIET LONG 20225508
PHAM LONG HAI 20225492
PHAM CONG HIEU 20225493
TRAN VUONG HUNG 20225496
NGUYEN PHAN THANG 20225529

OUTLINE

Problem Statement & Objectives

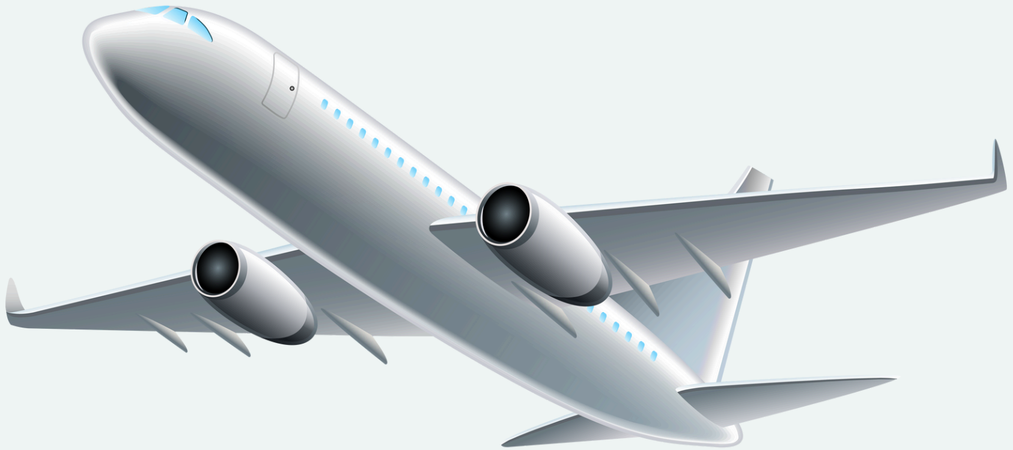
Dataset

Architecture



PROBLEM STATEMENT

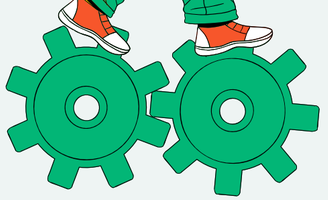
- The aviation industry plays a vital role in global connectivity and economic growth.
- Flight delays and cancellations cause **major inconvenience** for passengers and **financial losses** for airlines.
- Factors such as **weather, technical issues, airport operations, and air traffic control** increase the risk of disruptions.
- There is a strong need to **analyze flight data** to identify trends, causes, and potential issues in advance.



OBJECTIVES

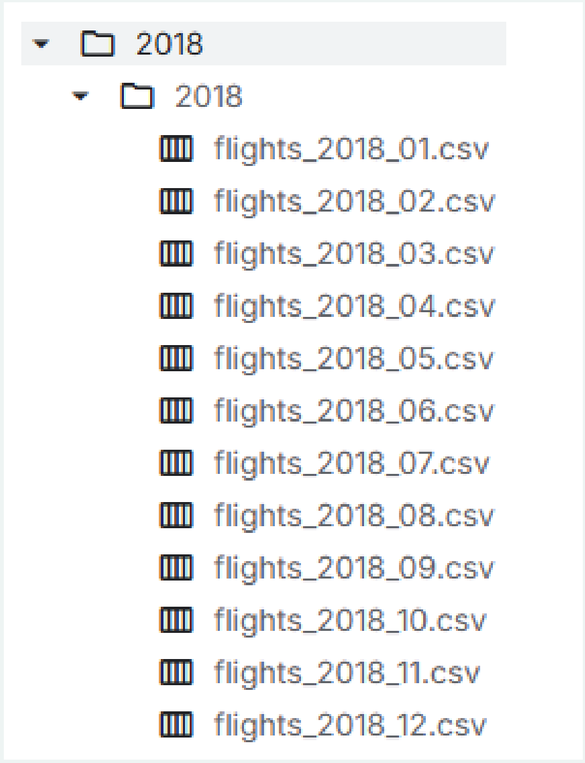
- Big Data and real-time processing enable continuous monitoring and faster decision-making.
- Real-time analytics help:
  - Detect and respond to delays **earlier**.
  - **Optimize** flight operations and resource allocation.
  - Improve passenger satisfaction and system efficiency.

- This project simulates **streaming flight data** and applies a **Lambda architecture** combining **batch and stream processing** to achieve efficient, scalable, and accurate analysis.



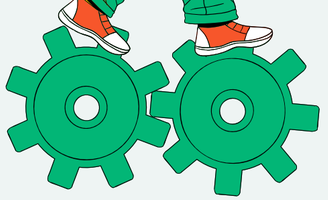
DATASET

- **Source:** Airline Delay Analysis Dataset from the U.S. Bureau of Transportation Statistics (BTS)
- **Official website:** transtats.bts.gov
- **Focus:** Domestic U.S. flight performance — focusing on flight delays.
- **Time range: January 2018 – June 2025**
- **Format:** Each month stored in a separate .csv file
- **Publicly uploaded version:** *Kaggle Dataset* (our customized version)
- Used to **simulate real-time streaming** for Big Data analysis and visualization



DATA EXAMPLE

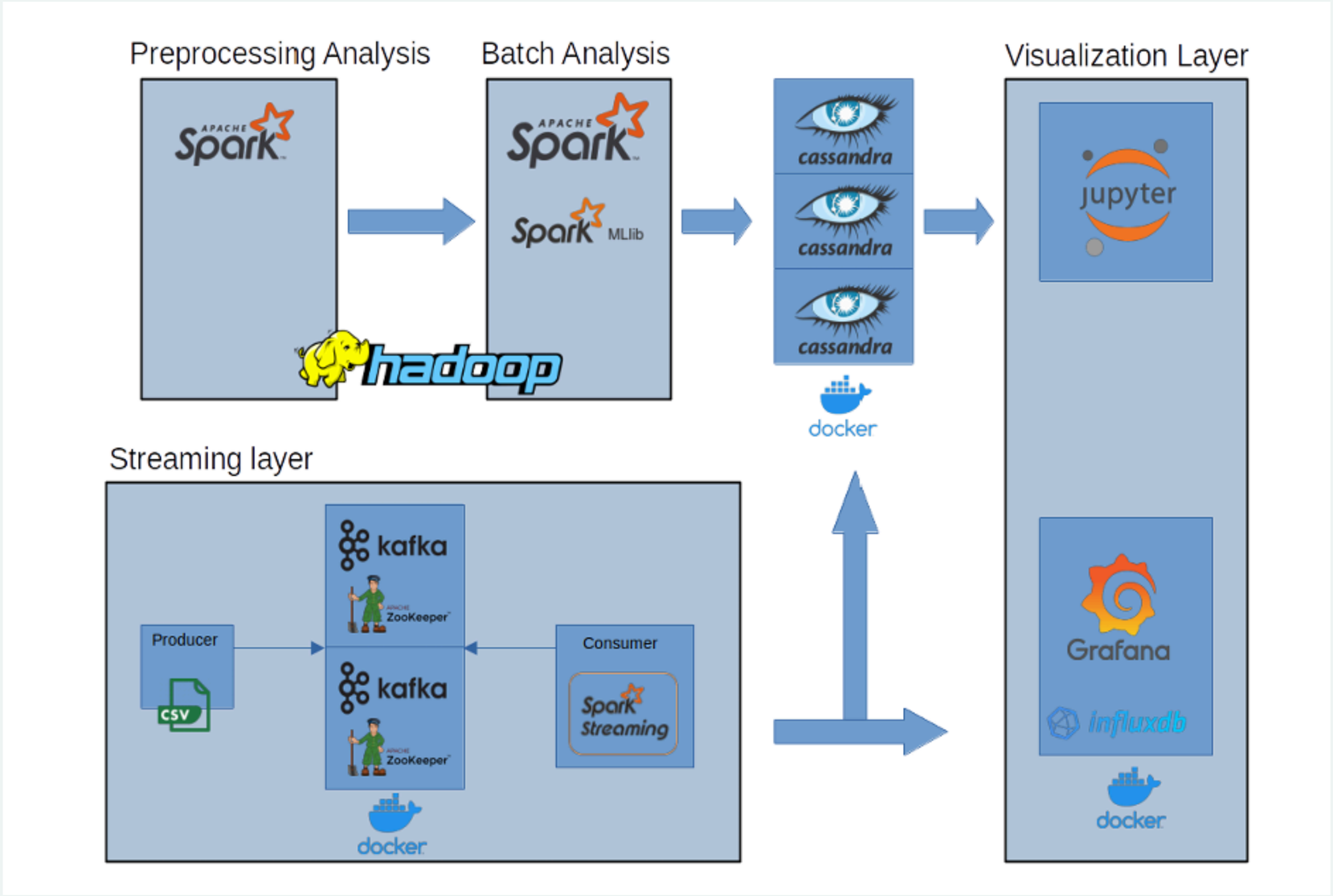
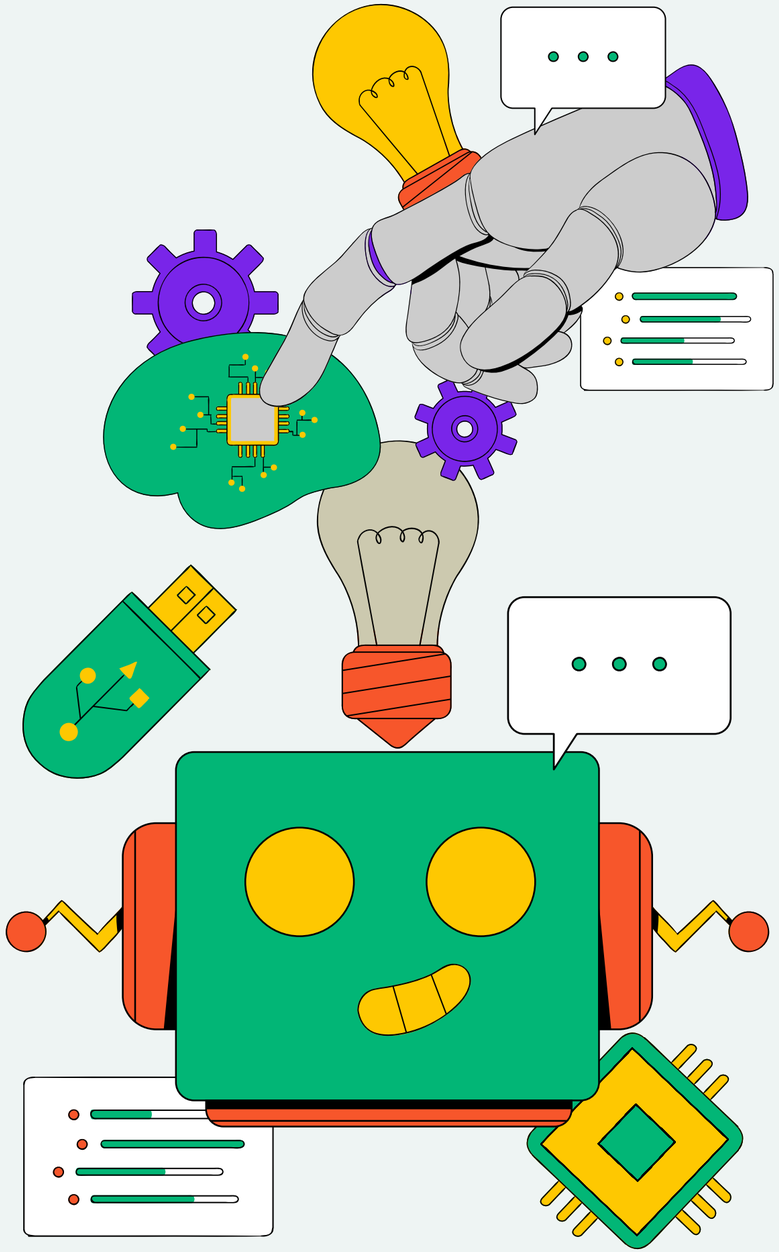
DATASET (2)



The dataset contains **26 attributes** describing various aspects of U.S. domestic flights, including:

- **Flight details** such as date, carrier, flight number, and aircraft tail number.
- **Airport information** for both origin and destination (airport code and city name).
- **Timing data** like scheduled and actual departure/arrival times, taxi times, and flight duration.
- **Delay-related fields** indicating departure and arrival delays, as well as the specific causes of delay (carrier, weather, system, security, or late aircraft).
- **Operational indicators** such as flight distance, cancellation status and reason, and diversion status.

ARCHITECTURE



THANK YOU FOR YOUR ATTENTION!