INNOPOLIS UNIVERSITY, BIG DATA
SPRING 2025

Predicting Flight Delays

TEAM 21

DMITRY ENIN
EGOR LEBEDEV
ALIE ABLAEVA
ANASTASIA ANKUDINOVA

Objective

Our goal is to use **Big Data** technology to predict flight delays using historical data (e.g., weather, airlines, airports) to:

- Reduce Costs*
- Improve Passenger Experience

Why Big Data?

- Dataset size (3 million records) demands distributed processing

* US economy suffers a $32.9 billion annual loss due to airplane delays

Our Plan

01

Data Collection

Downloading dataset to a database

02

Data Optimization

Preparing for efficient analysis

03

EDA

Analyze delay patterns

04

Model Building

Train machine learning models to predict delay

05

Dashboard Creation

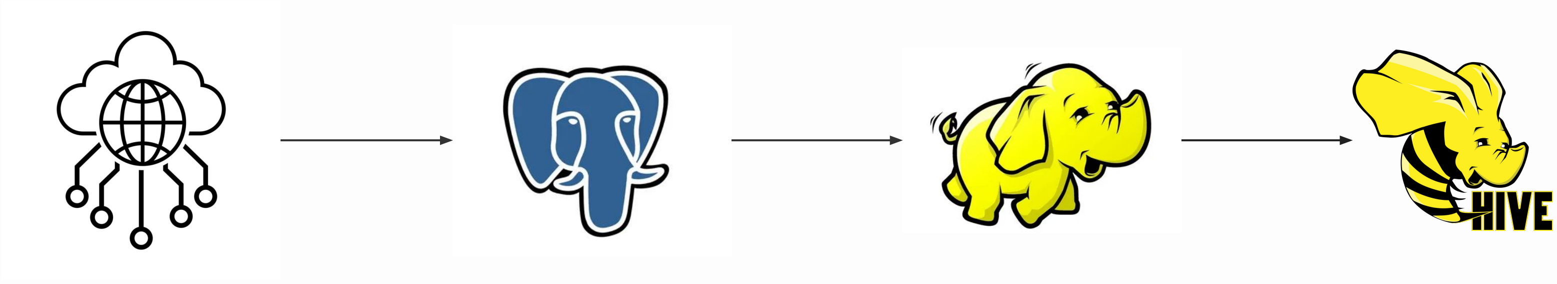
Vidualize findings

Stage-wise Results

Data Collection and Optimization

- Downloaded dataset using *wget*
- Loaded dataset to PostgreSQL

- Loaded data from PostgreSQL to HDFS as Parquet
- Hive optimization:
  1. Partitioning by *origin*
  2. Bucketing by *flight number*



Data Analysis

Dataset Characteristics

Title: Flight Delay and Cancellation Dataset (2019-2023)

Features:

Categorical: AIRLINE, ORIGIN, DEST

Numerical: DEP_TIME, DEP_DELAY

DateTime: FL_DATE

Target: DEP_DELAY - numerical, mean = 10.1, std = 49.3

Threshold for classification - 0

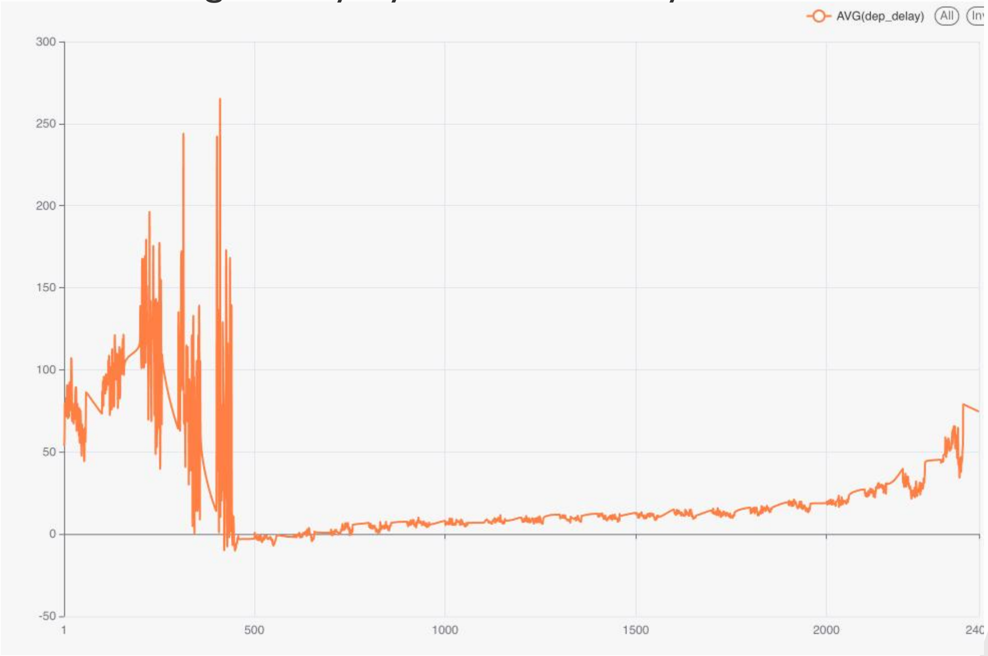
Data Analysis

Key Delay Patterns

Insight:

Delay starts slowly increasing after 10 am and gets extreme values between 1 am and 5 am

Average delay by time of the day



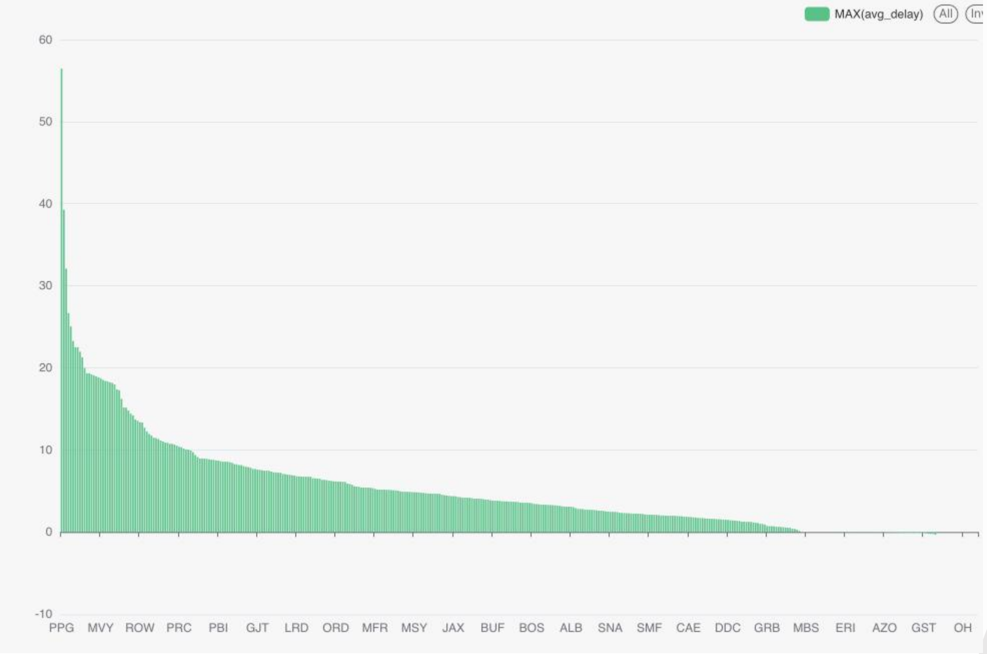
Data Analysis

Key Delay Patterns

Insight:

The airport affects the delay time

Average delay by airport



Dashboard

Conclusion

We have built

- Automatic data pipeline
- Machine learning model

Learned to handle large datasets with distributed tools (Spark, Hive).

Improved skills in performance optimization

Improved a skill of working in a team