Flight Delay & Cancellation Analysis (2019–2023)

Presented by:

Poojitha Jayareddygari

Navya Reddy Thadisana

Sahit Cheeka

Sai Kiran Jagini

Jeevith Gowda

The Problem: Flight Disruptions

Flight delays and cancellations are major pain points for both airports and passengers, leading to significant operational inefficiencies, increased costs, and decreased customer satisfaction.

Operational Impact

Disruptions strain airport resources and schedules.

Economic Cost

Financial losses for airlines and passengers.

Customer Dissatisfaction

Negative travel experiences and reduced loyalty.

Our Solution: Real-time Delay Prediction

We developed an end-to-end big-data pipeline using Apache Spark to analyze, model, and simulate flight delay patterns, providing real-time predictions.

Data Ingestion

Large-scale data collection.

Cleaning & Engineering

Preparing data for analysis.

Live Visualizations

Interactive dashboards.

EDA & SQL Analytics

Deep insights into patterns.

Structured Streaming

Real-time predictions.

Predictive ML Modeling

Classification & regression.

Dataset Overview: 3 Million Flights

Our analysis is based on a comprehensive dataset of U.S. domestic airline flight delays and cancellations from Kaggle.

Source & Scope

- Kaggle: U.S. Domestic Airline Flight Delays & Cancellations
- Period: Jan 2019 – Dec 2023
- Size: 3 million rows (2.1GB CSV)

Key Features

FL_DATE: Flight date

OP_UNIQUE_CARRIER: Airline code

ORIGIN, DEST: Airport codes

DEP_DELAY, ARR_DELAY: Delays in minutes

CANCELLED, DIVERTED: Event indicators

DISTANCE: Mileage

Data Preprocessing & Insights

We standardized data, handled missing values, and engineered new features to enhance predictive power.

1

Standardized Data

Cleaned column names and types.

2

Handled Nulls

Removed missing data for critical fields.

3

Derived Fields

Added year, month, day_of_week.

4

Categorical Encoding

Transformed airline, origin, destination.

Key EDA Insight: Arrival delay is almost entirely driven by departure delay (correlation ≈ 0.95).

Key EDA Findings

Our exploratory data analysis revealed critical patterns in flight delays.

Seasonal Peaks

Summer months (June–July) show the highest average delays.

Airline Performance

Low-cost carriers (Frontier, Allegiant, Spirit) have higher delay frequencies.

Hub Congestion

Major airports (DFW, ORD, ATL) experience significant congestion.

Predictive Modeling: High Accuracy

We applied both classification and regression models using MLlib and Sklearn to predict delays.

Best Classification Model

Logistic Regression: ROC-AUC = 0.9334

Strong Regression Model

Random Forest Regressor: R^2 = 0.8139

Overall Accuracy

Classification accuracy reaches 89%.

Feature Importance (Random Forest)

dep_delay	0.9109
year	0.0047
distance	0.0015
month	0.0007

Departure delay is the dominant predictor.

Real-time Streaming & Visualization

Our system generates real-time delay predictions and visualizes evolving patterns through a live dashboard.

Stream Batch Generation

Creates micro-batches of flight data.

ML Prediction

Real-time predictions per batch.

Live Visualization

Evolving delay patterns displayed.

Limitations & Future Enhancements

While robust, our system has areas for improvement.

Current Limitations

- Weather data absent (major predictor)
- Heavy dependence on departure delay
- Simulated streaming, not real API
- Resource bottlenecks on Spark/Codespaces

Future Enhancements

- Integrate real-time weather APIs
- Explore external factors for early prediction
- Connect to live airline data feeds
- Optimize Spark for larger datasets

Conclusion: A Powerful Predictive Tool

This project successfully demonstrates a full big-data analytics and ML pipeline for flight delay analysis.

Comprehensive Insights

Historical analysis reveals clear operational trends.

High Accuracy

Predictive models achieve strong performance.

Real-time Capability

Spark enables real-time predictions for proactive management.

Our system provides valuable insights for optimizing airline operations and enhancing passenger experience.

Thank You!