Flight Cancellation Prediction Using Domestic US Flights Data

Team 15
Nazgul Salikhova, Chulpan Valiullina, Milyausha Shamsutdinova, Diana Vostrova

Introduction

Problem: According to the Federal Aviation Administration (FAA), approximately 20% of flights in the United States experienced delays in 2019, resulting in an estimated **$32.9 billion** in costs to airlines, airports, and passengers.

Our goal: build a scalable Big Data pipeline to predict flight cancellations using US domestic flight data (2016–2018) and machine learning.

1. Perform data analysis to identify key patterns
2. Ingest and process millions of flight records
3. Store and query the data efficiently using Hive and HDFS
4. Train a machine learning classifier to predict flight cancellation
5. Present the analysis and predictions through an interactive dashboard

Dataset Characteristics

Source: Domestic US Air Flights 2016–2018 – *Kaggle*

Key stats

- 18.5M total rows
- 26 features
  Covers 3 years (2016–2018)
  Includes: times, carriers, delays, cancellations, reasons

Target variable

- `Cancelled` (1 = cancelled, 0 = not)

Architecture of Data Pipeline

Pipeline stages

- **Ingestion**: CSV → PostgreSQL → HDFS using Sqoop
- **Storage**: Hive tables (partitioned, Parquet using Snappy)
- **EDA**: HiveQL & Tez engine
- **Modeling**: Spark MLlib
- **Visualization**: Apache Superset dashboards

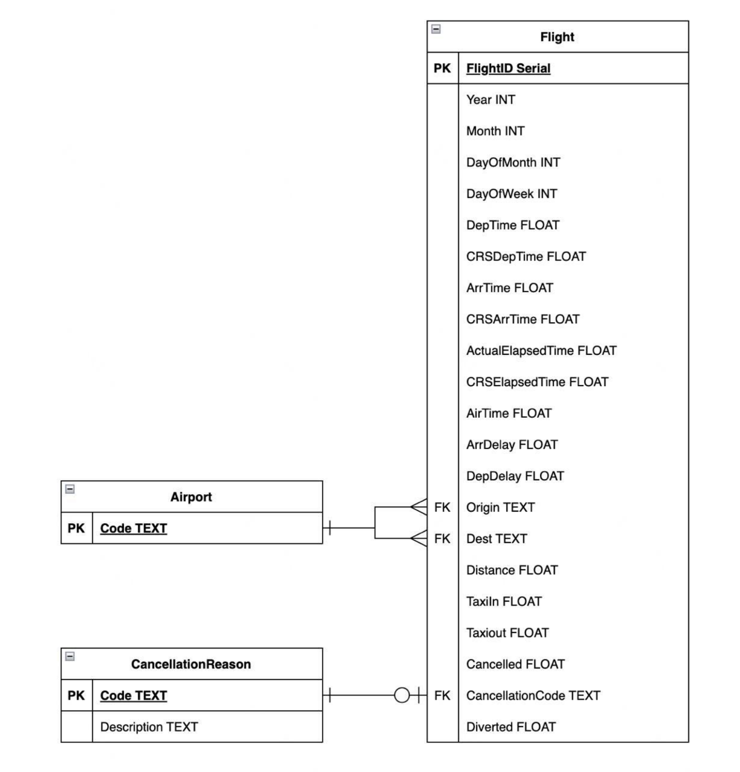
Tools used

- Apache Sqoop
- Hive
- Spark
- Apache Superset

The results by stages

01 Data Ingestion

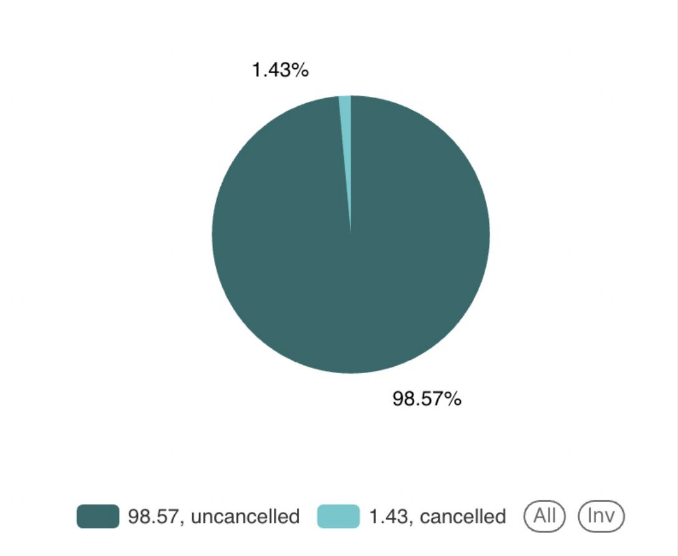
- Successfully processed raw CSV file (18,5M rows - 1.9 GB)
- Built ER diagram and relational model
- Removed features 'carrierdelay', 'weatherdelay', 'nasdelay', 'securitydelay', 'lateaircraftdelay'
- Loaded data into PostgreSQL
- Transferred to HDFS using Apache Sqoop and Hadoop MapReduce engine
- Stored in AVRO format with SNAPPY compression



ER diagram

Analysis results

Insight 1: Flight cancellation rate



Flight cancellations rate is pretty high 1.43%, which affect hundreds of thousands of flights and cause significant disruption.

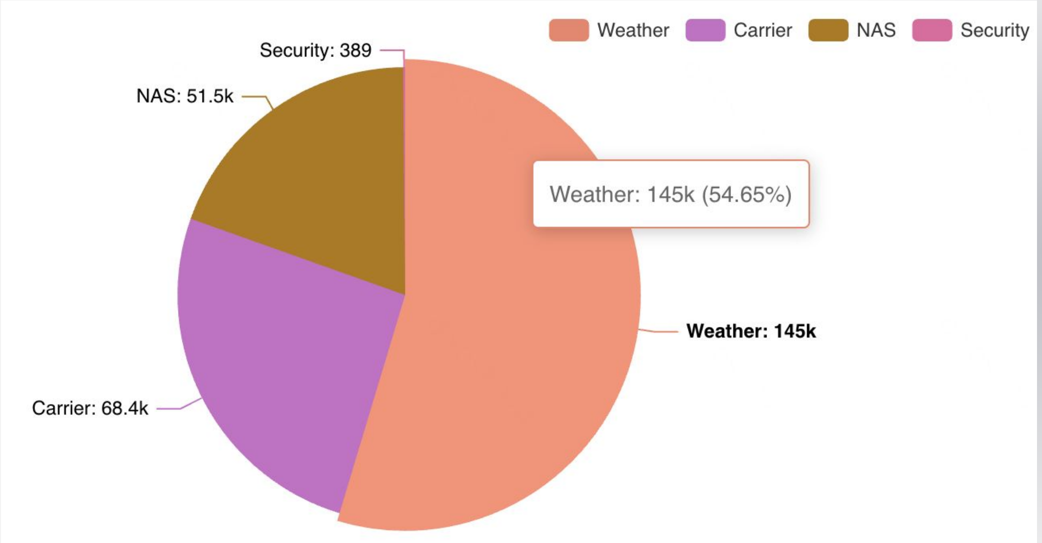
cancelled_count	not_cancelled_count	total_flights
265138	18240587	18505725

Analysis results

Insight 4: Cancellation Reasons

Weather is the dominant cause of cancellations, far ahead of carrier or airspace issues.

- Weather – 54.65%
- Carrier – 25.78%
- Problems within the National Airspace System – 19.43%
- Security – 0.15%



The results by stages

02 Data Preparation & EDA

- Created Hive tables
- Optimized hive tables with partitioning and bucketing
- Stored data in data warehouse Hive
- Performed Exploratory Data Analysis with HiveQL
- Created insightful charts in Apache Superset

Data preparation for ML modelling

Data preprocessing pipeline

1. Drop columns with large number of missing values which strongly correlate with label
2. Drop rows with other missing values
3. Decompose time features
4. Encode categorical features
5. Scale numerical features
6. Balance dataset by down-sampling the majority class
7. Split dataset into train / test

Initially	18505725
After drop missing values	18505702
Balanced	531167

Data preparation for ML modelling

Table flights

year	month	dayofmonth	dayofweek	crsdeptime	crsarrtime	crselapsedtime	origin	dest	distance	cancelled
2017	11	12	7	2055.0	2205.0	130.0	DCA	MKE	634.0	0.0
2017	11	12	7	1430.0	1535.0	125.0	DCA	MKE	634.0	0.0
2017	11	12	7	1305.0	1500.0	175.0	DCA	MSY	969.0	0.0
2017	11	12	7	1855.0	2050.0	175.0	DCA	MSY	969.0	0.0
2017	11	12	7	1645.0	1845.0	180.0	DCA	OMA	1012.0	0.0
2017	11	12	7	1955.0	2115.0	80.0	DCA	PVD	356.0	0.0
2017	11	12	7	1145.0	1305.0	80.0	DCA	PVD	356.0	0.0
2017	11	12	7	755.0	915.0	140.0	DCA	STL	719.0	0.0
2017	11	12	7	1155.0	1320.0	145.0	DCA	STL	719.0	0.0
2017	11	12	7	1855.0	2015.0	140.0	DCA	STL	719.0	0.0

ML Modelling

Models and metrics

Model	AUC	Accuracy	Precision	Recall	F1-score
Logistic Regression (elasticNetParam: 0.0, maxIter: 50, regParam: 0.01)	69.58%	64.16%	63.79%	65.18%	64.48%
Random Forest (maxDepth: 15, numTrees: 50, featureSubsetStrategy: sqrt)	**76.49%**	**69.56%**	**69.39%**	**69.78%**	**69.59%**

The results by stages

03 Modeling in Spark ML

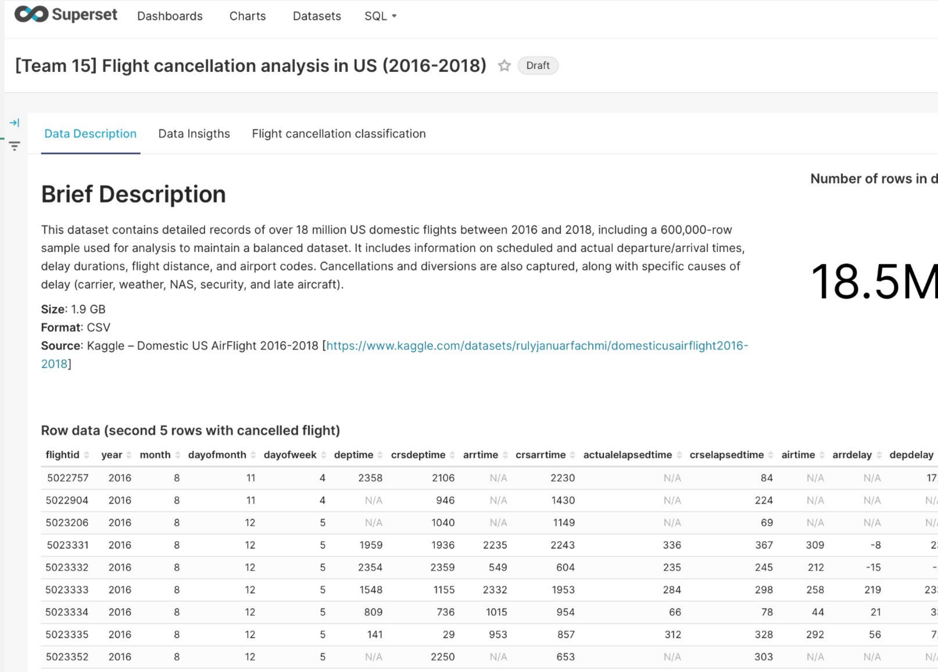
- Performed data preprocessing with PySpark (handle missing values, time feature decomposing, one-hot encoding, etc.)
- Trained classification models:
  - Logistic Regression
  - Random Forest
- Evaluated models using accuracy, F1-score, and ROC-AUC

Best results achieved with **Random Forest** classifier

The results by stages

04 Superset Dashboard

- Built an interactive dashboard using Apache Superset
- Visualized tabs:
  - Data description
  - Data insights
  - Flight cancellation classification
- Added valuable text blocks
- Added evaluation metrics of ML model training



flightid	year	month	dayofmonth	dayofweek	deptime	crsdeptime	arrtime	crsarrtime	actualelapsedtime	crselapsedtime	airtime	arrdelay	depdelay
5022757	2016	8	11	4	2358	2106	N/A	2230	N/A	84	N/A	N/A	17
5022904	2016	8	11	4	N/A	946	N/A	1430	N/A	224	N/A	N/A	N/
5023206	2016	8	12	5	N/A	1040	N/A	1149	N/A	69	N/A	N/A	N/
5023331	2016	8	12	5	1959	1936	2235	2243	336	367	309	-8	2
5023332	2016	8	12	5	2354	2359	549	604	235	245	212	-15	-
5023333	2016	8	12	5	1548	1155	2332	1953	284	298	258	219	23
5023334	2016	8	12	5	809	736	1015	954	66	78	44	21	3
5023335	2016	8	12	5	141	29	953	857	312	328	292	56	7
5023352	2016	8	12	5	N/A	2250	N/A	653	N/A	303	N/A	N/A	N/

Demo

http://hadoop-03.uni.innopolis.ru:8808/superset/dashboard/169/

Conclusion

We learned how to

- Ingest large-scale datasets into HDFS using PostgreSQL and Sqoop
- Store and prepare data using Hive with AVRO and SNAPPY compression
- Perform exploratory data analysis using Hive and Spark
- Train and evaluate machine learning models using Spark MLlib
- Visualize insights with interactive dashboards in Apache Superset

Challenges Encountered

- Hive tables optimization
- Long model training time
- Working in cluster in the first time
- Superset is not intuitively understable
- Cluster is down several times

Thanks!

- Our GitHub repository

Contacts:

- n.salikhova@innopolis.university
- c.valiullina@innopolis.university
- m.shamsutdinova@innopolis.university
- d.vostrova@innopolis.university