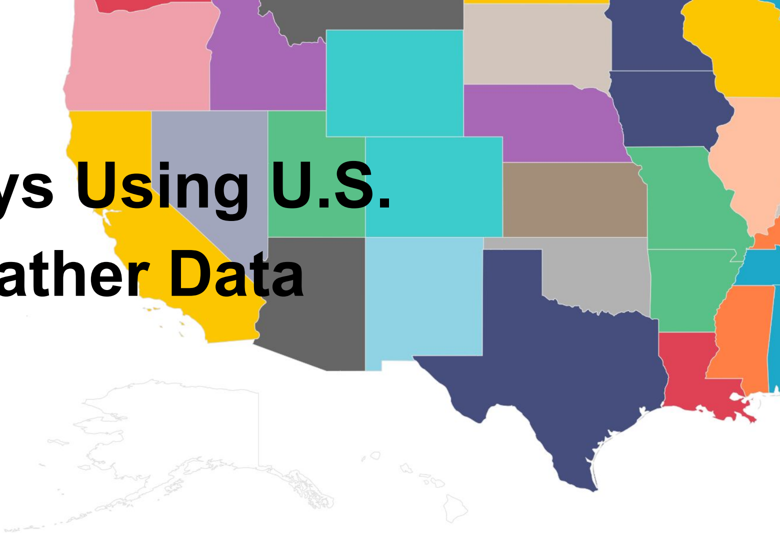


Predicting Traffic Delays Using U.S. Congestion and Weather Data

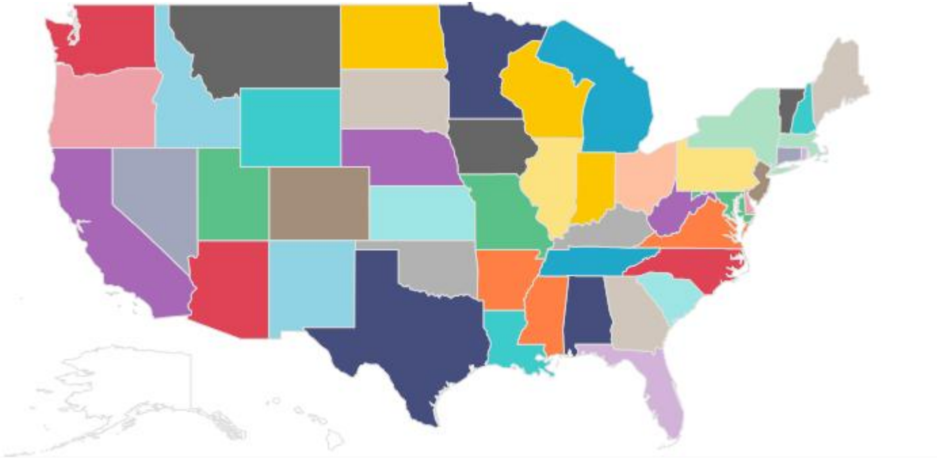
Team 26

Name	Role	Innomail
Alexey Tkachenko	Data engineer	a.tkachenko@innopolis.university
Daniil Abrosimov	ML specialist	d.abrosimov@innopolis.university
Egor Machnev	Data scientist	e.machnev@innopolis.university
Apollinaria Chernikova	Tester and Technical writer	a.chernikova@innopolis.university

+ CI bot

Introduction

- Analyze traffic congestion on US roads
- Identify key factors influencing traffic delays (weather, time, location).
- Predict traffic jam duration with reasonable accuracy.
- Develop a big data pipeline from raw CSVs to analytical dashboard.
- Provide decision support for urban planners and transportation authorities.

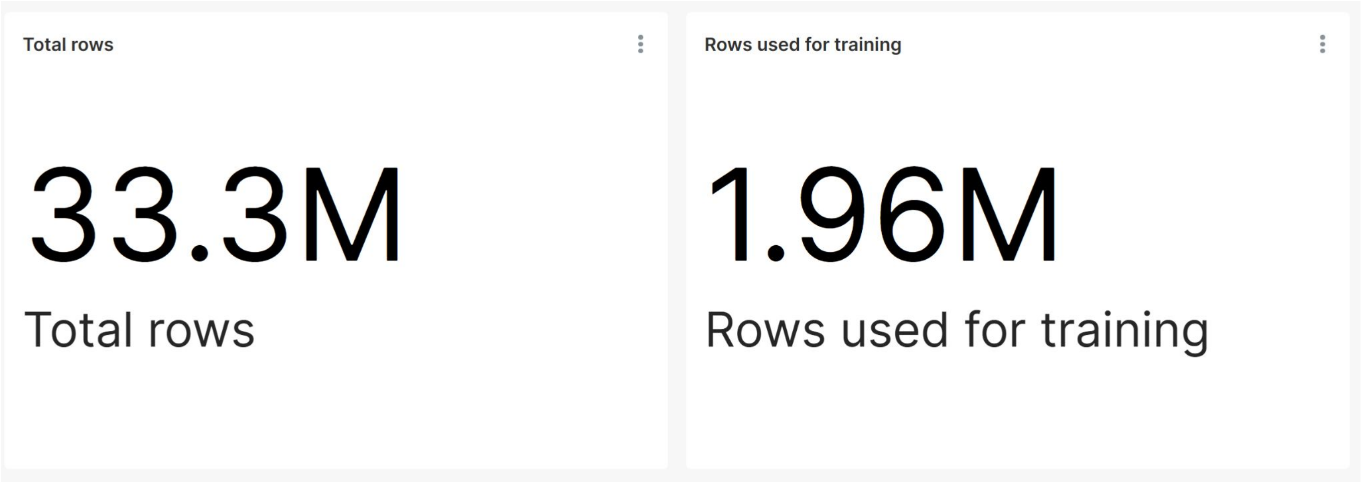


Data Description

US Traffic Congestions (2016-2022)

Comprehensive Dataset of 33 Million U.S. Traffic Congestion Events

Kaggle



Total rows

33.3M

Total rows

Rows used for training

1.96M

Rows used for training

Architecture of data pipeline

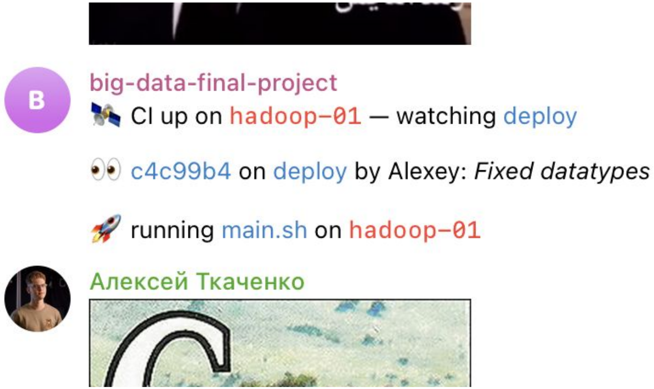
Stage	Input	Processing	Output
Stage I	Kaggle ZIP archive (CSV, ~12.8 GB)	- Load and clean data - Import to PostgreSQL - Export to HDFS using Sqoop → convert to Parquet (Snappy-compressed)	Raw lake: /project/warehouse/traffic (Parquet)
Stage II	Parquet files from Stage I	- Create external Hive table traffic - Create optimized partitioned & bucketed table traffic_partitioned - Auto-ingest per-state - Generate 14 EDA views (CSV + Hive)	Hive warehouse + /output/dashboard/qX.csv
Stage III	50k-row-per-state sample from traffic_partitioned Hive table. ~2mil rows in total	- Spark MLlib pipeline with custom transformers: - Cyclical time encoder - GeoToECEF spatial transformer - Word2Vec embedding - OneHotEncoding for low cardinality categorical features - Feature hashing for high cardinality features - Grid search for LR and RF models	Trained models + evaluation metrics (RMSE, R², MAE)
Stage IV	Analytical results + model outputs	- Load into Apache Superset - Configure dashboards, filters, drilldowns	Interactive BI dashboard

GitHub

Development Automation ✨

To streamline development and ensure continuous integration throughout the project lifecycle, we implemented a custom CI script tailored for the cluster environment.

- Fetches and resets to the latest commit;
- Executes the main pipeline script (main.sh);
- Sends logs and outputs to a Telegram chat for real-time team visibility.



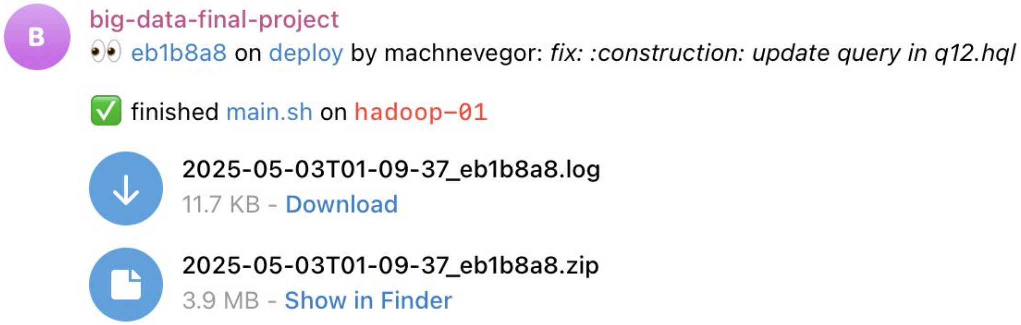
big-data-final-project

🛰 CI up on hadoop-01 — watching deploy

👀 c4c99b4 on deploy by Alexey: *Fixed datatypes*

🚀 running main.sh on hadoop-01

Алексей Ткаченко



big-data-final-project

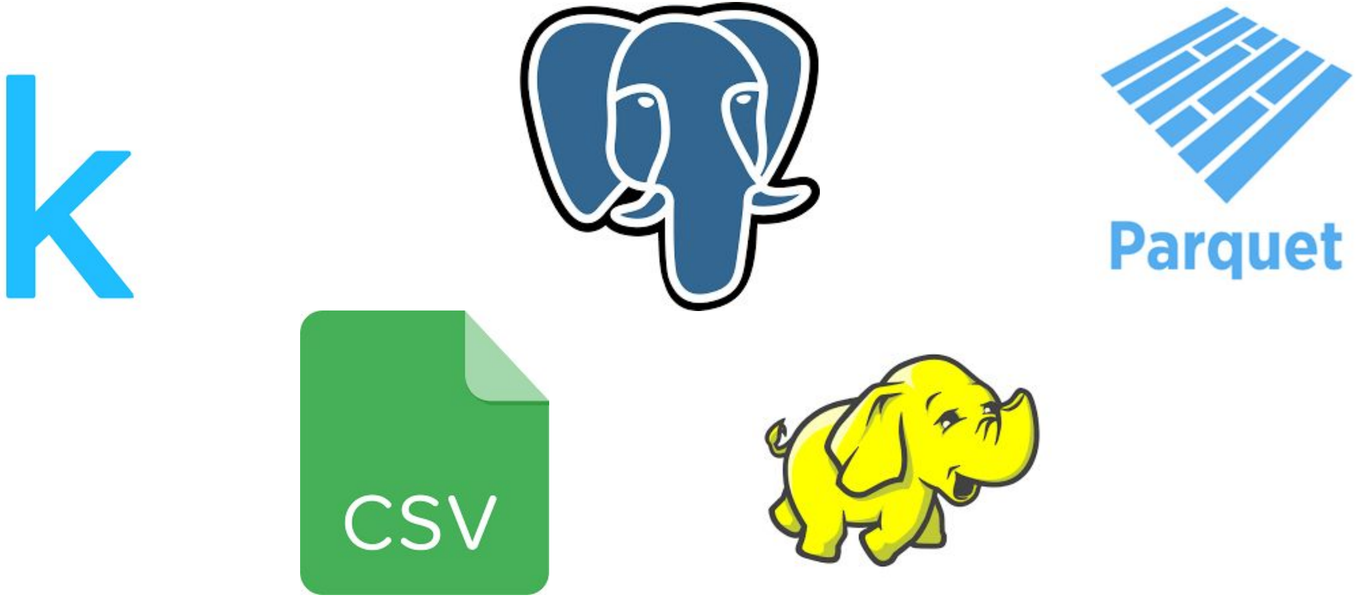
👀 eb1b8a8 on deploy by machnevegor: *fix: :construction: update query in q12.hql*

✅ finished main.sh on hadoop-01

2025-05-03T01-09-37_eb1b8a8.log
11.7 KB - Download

2025-05-03T01-09-37_eb1b8a8.zip
3.9 MB - Show in Finder

Data preparation



Data analysis

Initially formulated **14** hypotheses about factors that might influence traffic delays.

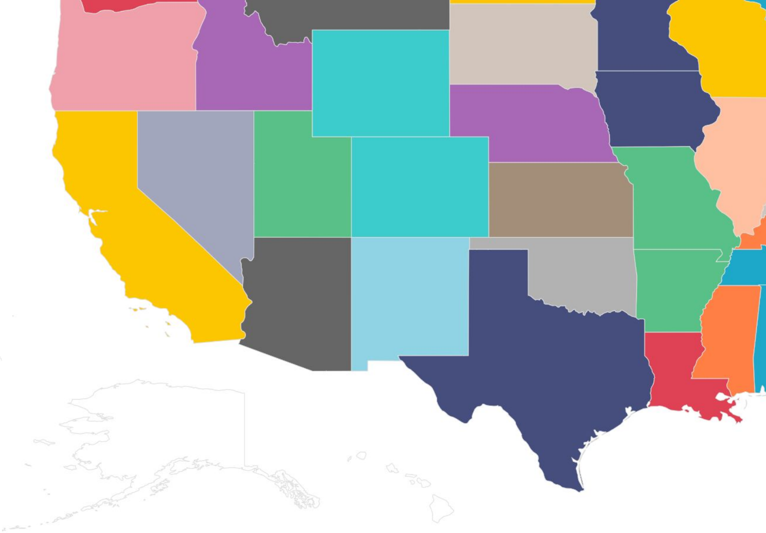
7 were tested and validated through visual analysis using charts and dashboards.

ML Modeling

- **Random Forest Regressor**:
  - maxDepth 10;
  - numTrees 50.
- **Linear Regression**:
  - regParam 0.01;
  - elasticNetParam 0.0 (pure L2 regularization)

Model	**RMSE**	**R²**	**MAE**
Random Forest	2.9146	0.5787	1.5438
Linear Regression	3.3102	0.4566	1.8117

Data presentation



Superset
Dashboard

Conclusion

Our contributions include:

- ➔ End-to-end automation for data ingestion, partitioning, and transformation;
- ➔ Targeted HQL queries for exploratory data analysis (EDA);
- ➔ A trained machine learning model predicting traffic delays using geographic, temporal, and weather-related features;
- ➔ An interactive Superset dashboard for intuitive data visualization and decision support.

Reflections on own work

A solid and awesome team!