

# FACTORED DATATHON

## TEAM SEED42

approaches for  
sanction analysis

# OUR APPROACH

## Data Acquisition and Delivery

Utilizing a web scraper to automatically collect, organize, merge, and deliver data through the Spark framework.

## Data Preparation and Initial Analysis

Preparing the data was essential to tailor it for our application development needs. Initial analysis was made using Jupyter Notebooks over PySparkSQL.

## Goldstein Scale Prediction

A model capable of predicting the score was created and delivered to users who need to perform simulations and analysis over this metric.

## RAG Based Chat Bot Over the Data

A LLM was deployed using LangChain agents and tool to query the data and run analysis over it, answering the users questions with our chatbot

## Infrastructure

The initial version was developed on a simple architecture using Spark and accessed via Streamlit, APIs, or CLI. However, a more robust architecture has been planned for future implementation in the project. Currently, the project is only running locally.

# EDA

Our data analysis focused on sanction events, particularly by combining GDELT databases with CAMEO codes.

The analysis covers data from January 1, 2023, to October 16, 2024, and includes:

- Geographic Distribution of Sanction-Related Events
- Temporal Trend of Sanction-Related Events
- Sentiment Analysis Related to Sanctions
- Identifying Key Actors in Sanction-Related Events
- Sentiment Analysis Over Time for Specific Countries
- Co-occurrence Analysis of Actors
- Temporal Analysis of Sanction Events by QuadClass
- Impact Analysis: Sanctions and Media Attention
- Analyzing the Effect of Sanctions on International Relations

The screenshot shows a Jupyter Notebook cell with the following code:

```
gkg_files = glob.glob('/home/tiago/factored-datathon-2024-*.csv')
csv_files = [file for file in gkg_files if file.lower().endswith('.csv')]
events_df = spark.read.csv(csv_files, sep='\t', header=True)
cameo_df = spark.read.csv('/home/tiago/factored-datathon-2024.csv')
cameo_df = cameo_df.withColumnRenamed("EventCode", "EventCode_join")
merged_df = events_df.join(cameo_df, events_df['EventCode'] == cameo_df['EventCode'])
gdelt_df = merged_df.drop("EventCode_join")

# Create year and month columns for easier analysis
gdelt_df = gdelt_df.withColumn("GoldsteinScale", col("GoldsteinScale").cast("float"))
gdelt_df = gdelt_df.withColumn("NumMentions", col("NumMention").cast("float"))
gdelt_df = gdelt_df.withColumn("NumSources", col("NumSources").cast("float"))
gdelt_df = gdelt_df.withColumn("NumArticles", col("NumArticles").cast("float"))
gdelt_df = gdelt_df.withColumn("AvgTone", col("AvgTone").cast("float"))

# Show the first few rows
gdelt_df.show(5)
```

The output of the code shows the first 5 rows of the `gdelt_df` DataFrame:

GLOBALEVENTID	SQDDATE	MonthYear	Year	FractionDate	Actor1Code
NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL
1161944228	20140309	201403	2014	2014.189	CAN
1161944228	20140309	201403	2014	2014.189	CAN

only showing top 5 rows

# GOLDENSTEIN SCALE PREDICTOR

To help organizations simulate and predict the impact of sanctions in the global scale, we developed a Machine Learning model capable of predicting the Goldenstein Scale using the following variables

The model used was and achieved XXX accuracy on test data.

```
Outputs > sanction_analysis.ipynb > M4 Analyzing the Effect of Sanctions on Inte...  
Code | Markdown | Run All | Restart | Clear All Outputs | Variables  
  
gkg_files = glob.glob('/home/tiago/factored-datathon-2024-1/*.csv')  
csv_files = [file for file in gkg_files if file.lower().endswith('.csv')]  
events_df = spark.read.csv(csv_files, sep='\t', header=True)  
cameo_df = spark.read.csv('/home/tiago/factored-datathon-2024-1/cameo.csv')  
cameo_df = cameo_df.withColumnRenamed("EventCode", "EventCode_join")  
merged_df = events_df.join(cameo_df, events_df['EventCode'] == cameo_df['EventCode_join'])  
gdelt_df = merged_df.drop("EventCode_join")  
  
# Create year and month columns for easier analysis  
gdelt_df = gdelt_df.withColumn("GoldsteinScale", col("GoldsteinScale").cast(IntegerType()))  
gdelt_df = gdelt_df.withColumn("NumMentions", col("NumMentions").cast(IntegerType()))  
gdelt_df = gdelt_df.withColumn("NumSources", col("NumSources").cast(IntegerType()))  
gdelt_df = gdelt_df.withColumn("NumArticles", col("NumArticles").cast(IntegerType()))  
gdelt_df = gdelt_df.withColumn("AvgTone", col("AvgTone").cast(IntegerType()))  
  
# Show the first few rows  
gdelt_df.show(5)
```

✓ 21.7s

GLOBALEVENTID	SQDDATE	MonthYear	Year	FractionDate	Actor1Code
NULL	NULL	NULL	NULL	NULL	Actor1Code
NULL	NULL	NULL	NULL	NULL	Actor1Code
NULL	NULL	NULL	NULL	NULL	Actor1Code
1161944228	20140309	201403	2014	2014.189	CAN
1161944228	20140309	201403	2014	2014.189	CAN

only showing top 5 rows

# GDELT CHATBOT

Additionally, we have developed a chatbot using a Streamlit application to simplify data analysis and handling. This chatbot enables users to input a data range for analysis and ask questions to an LLM model, which utilizes Langchain Agents and Tools in a RAG-like structure for data analysis and reasoning.

## GDELT Data Scraper and Chatbot

### Get your base data

Start Date

2024/08/24

End Date

2024/08/24

Fetch Data

### GDELT Chatbot

Enter question:

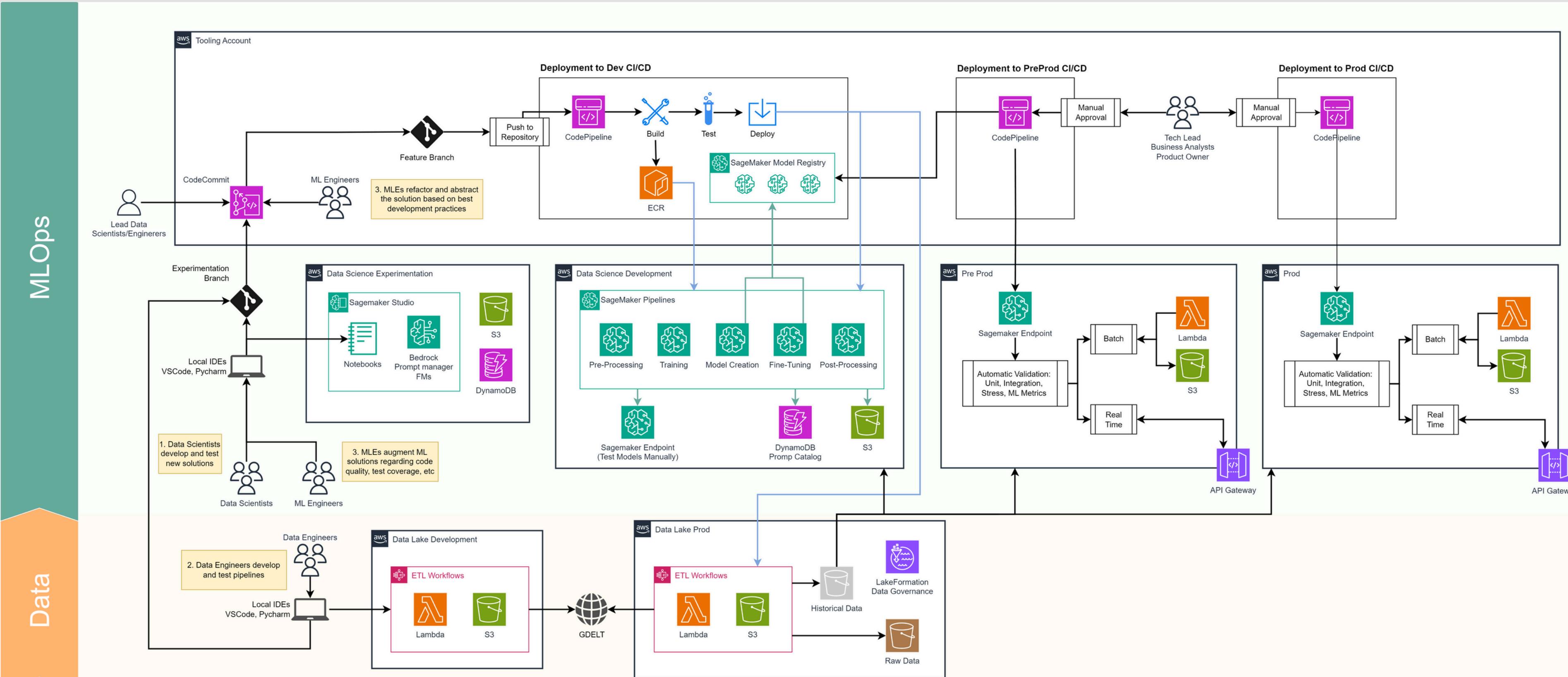
How many sanctions impact Venezuela?

Submit

There are 2709 sanctions that impact Venezuela.

# INFRASTRUCTURE

Although for this first version we opted for a local and simplified infrastrucutre, we propose the following design to make this solutions to production.



# OUR GITHUB REPOSITORY

Our github repository can be accessed through the following link and contain the content below

<https://github.com/glev1/factored-datathon-2024-seed42>

Content:

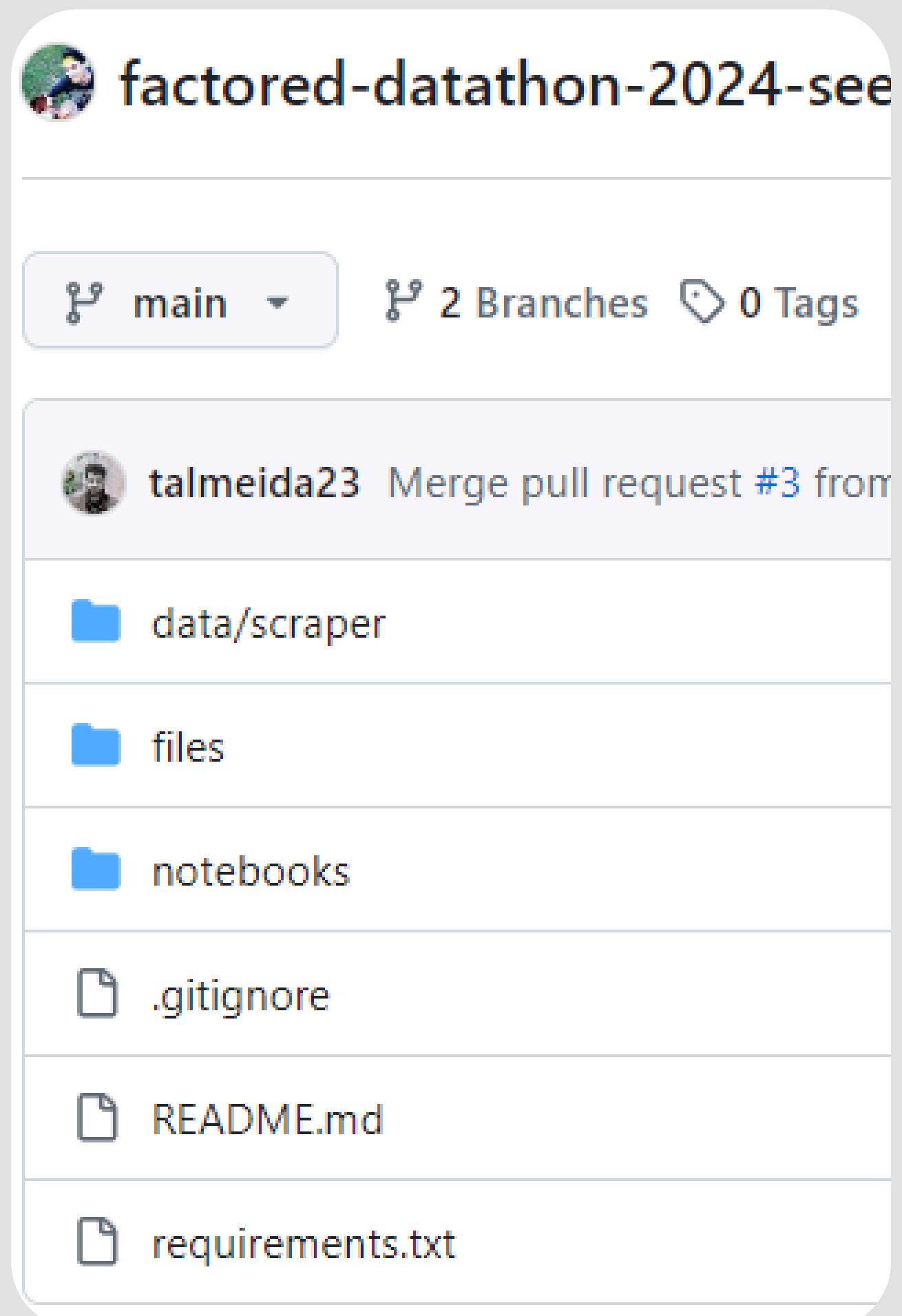
app - Our Streamlit application

notebooks - Our Jupyter EDA notebooks

packages - The libraries we built to support our development

README.md - Our general documentation

requirements.txt - The Python packages we used



# THANK YOU

team seed42