NOVA IMS INFORMATION MANAGEMENT SCHOOL

Big Data Analysis- Spark Applicability Use Cases

Group Project

Antónia Lemos 20231618, Carolina Lee Chin 20231726, Darija Avramoska 20230004, Luís Soeiro 20211536, Matilde Maximiano 20231687,

December 2025

Predicting loan approval: ML Models in Spark

Tools: Spark SQL / DataFrames + MLlib / Spark ML

Use case: Running predictive machine learning models on big data

Benefits of using Pyspark

- PySpark keeps data in computer memory (RAM) making analysis faster than traditional methods
- MLlib provides ready-to-use algorithms optimized for distributed computing
- PySpark automatically splits the work across multiple computers working simultaneously
- PySpark plans the entire workflow first and figures out the most efficient way to process everything

ML Pipeline

PREPROCESSING

StringIndexer / OneHotEncoder can isolate the true impact of each customer group instead of making assumptions.

Vector Assembler enables us to combine all customer signals into one consistent risk profile.

FEATURE SELECTION

UnivariateFeatureSelector assess the relevance of features to clearly explain which customer attributes truly affect approval decisions.

Pipeline standardize how data is processed and explain how a loan decision is generated from raw data.

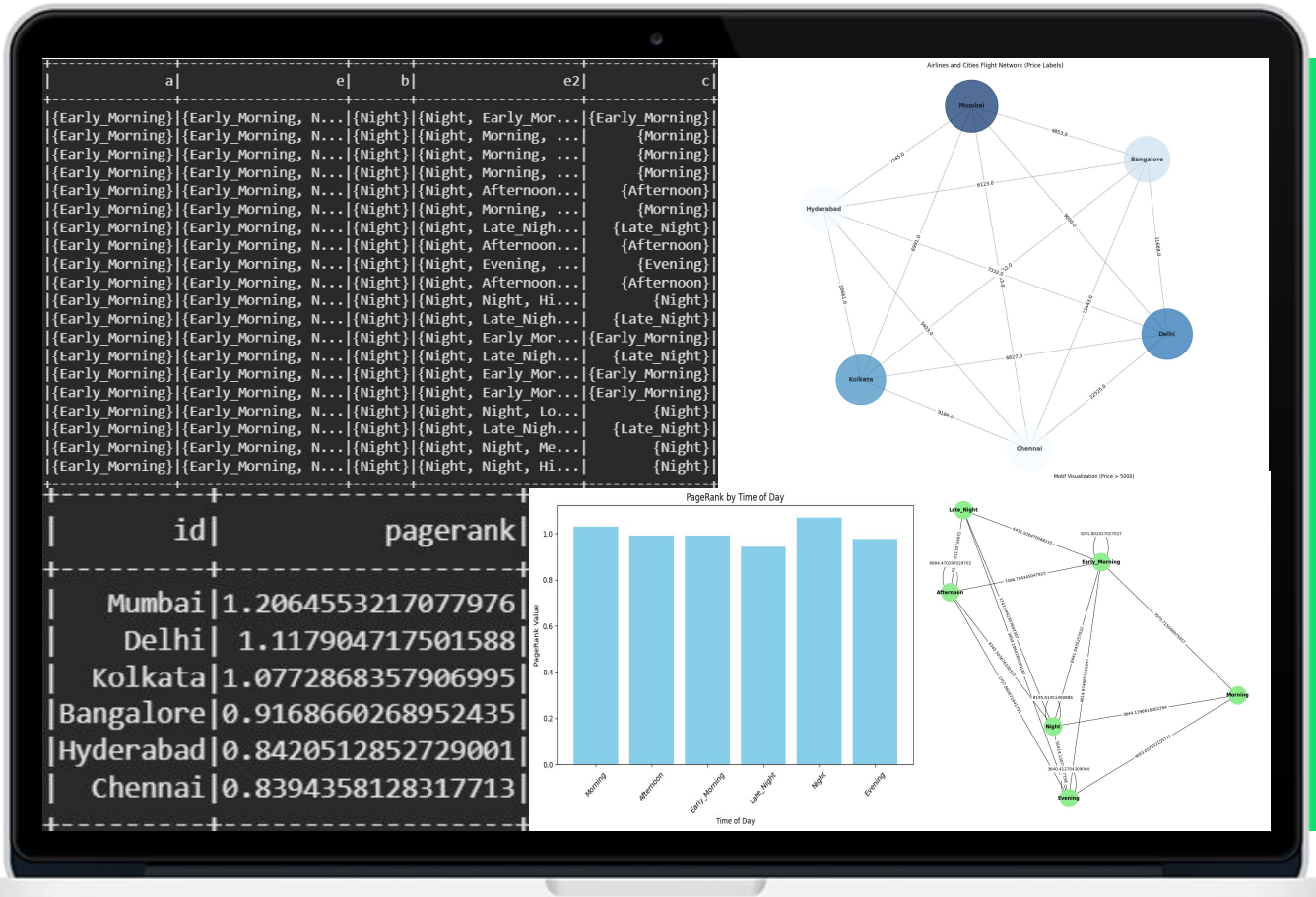
Home ownership: Renting was a surprising positive signal

Past loan defaults had a strong negative effect

Unlocking Business Insights from Flight Data Using Spark

Tools: Spark SQL + GraphFrames

Use case: Analyze large amounts of flight data to uncover valuable business insights and underlying relationships.



Network Analysis

Exploring connections using graph based analysis.

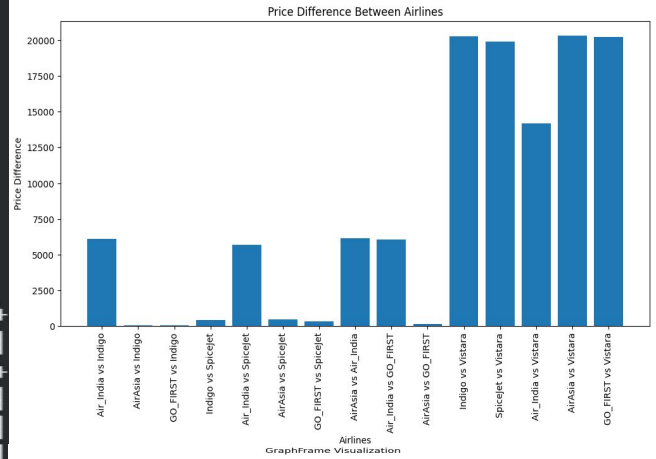
- **GraphFrames** is a great tool for analyzing large scale networks due to its integration with Spark SQL. It was employed to build graphs and allowed us to apply several algorithms.
- **PageRank** was used to determine insights like the most relevant cities and airlines based on their connections within the network.
- **Motifs** was used to identify specific patterns of connections within the network. We filtered motifs where flight prices were higher than a defined threshold to find premium routes and understand price clustering.

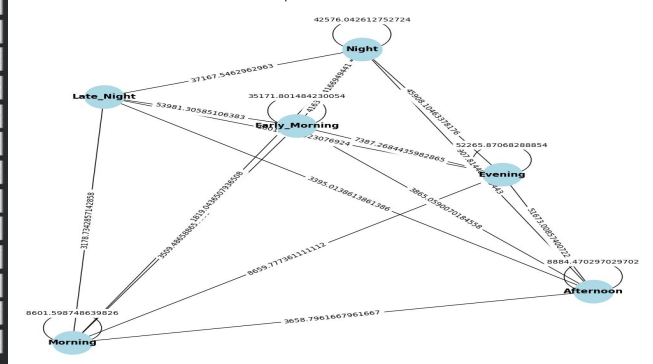
Spark for Price Analysis and Centrality Metrics

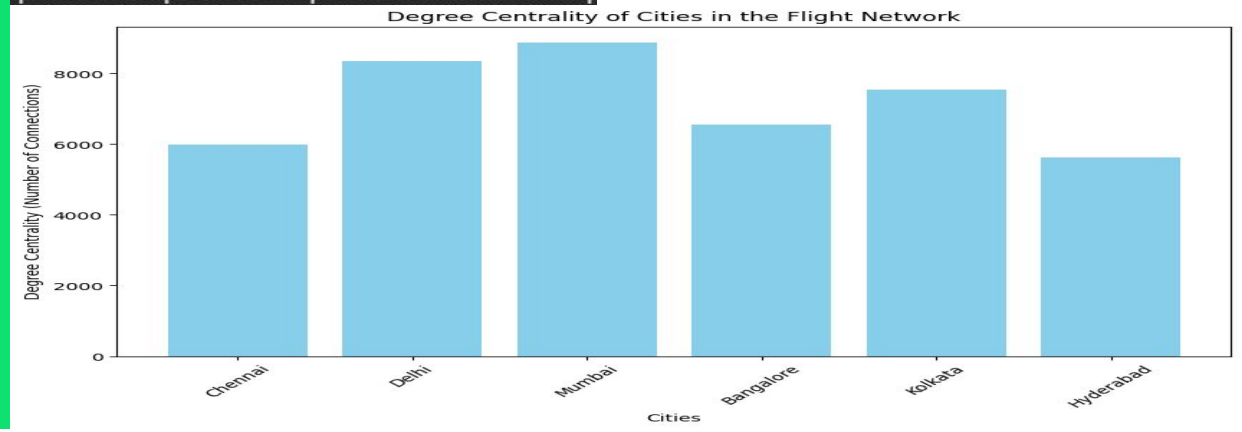
- Spark's ability to handle large scale datasets in a distributed environment makes it an ideal tool for analyzing complex data. Spark's GraphFrames and DataFrames operation allow us to process massive datasets without compromising performance.
- **Spark SQL + DataFrame**: By grouping data based on airlines, departure times and price ranges,we can perform a price analysis across vast amounts of data. This allows airlines to better understand price analysis and optimize pricing strategies.
- **GraphFrames**: It allowed us to model flight networks with cities as vertices and flights as edges. Applying degree centrality and PageRank gave us valuable insights. Cities with high degree centrality are key hubs in the flight network. Those with highest PageRank scores are critical for airlines' strategies.

airline	avg_price
Indigo	8843.91246470351
SpiceJet	9230.846091205212
Air_India	14926.963198546115
AirAsia	8787.304283604137
GO_FIRST	8899.644666666667
Vistara	29130.87941628264

airline1	airline2	price_difference
Air_India	Indigo	6083.050733842605
AirAsia	Indigo	56.608181099372814
GO_FIRST	Indigo	55.73220196315742
Indigo	SpiceJet	386.93362650170275
Air_India	SpiceJet	5696.117107340902
AirAsia	SpiceJet	443.54180760107556
GO_FIRST	SpiceJet	331.2014245385453
AirAsia	Air_India	6139.658914941978
Air_India	GO_FIRST	6027.318531879448
AirAsia	GO_FIRST	112.34038306253024
Indigo	Vistara	20286.966951579132
SpiceJet	Vistara	19900.033325077427
Air_India	Vistara	14203.916217736527
AirAsia	Vistara	20343.575132678503
GO_FIRST	Vistara	20231.234749615975







Hacker News Commentary Sentiment Analysis

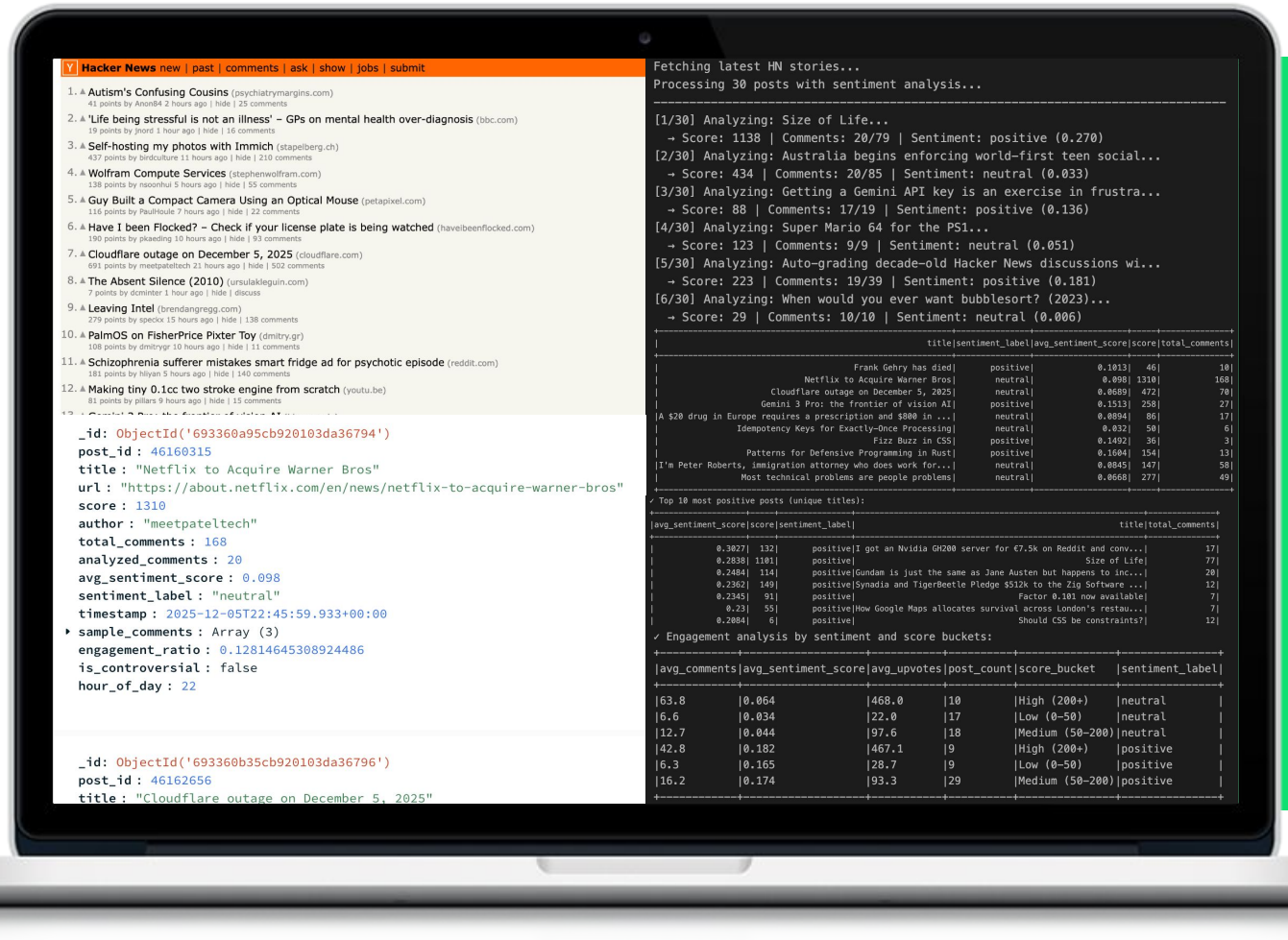
Tools: Kafka (Confluent Cloud) + MongoDB

Use case: Understand public sentiment to social outlets posts by analyzing large volumes of comments in real-time.

> ***Hacker News generates large volumes of data at increasing speed, both in posts and respective comment sections. For a business, it is hard to grasp how the public is reacting to the news in the comment section.***

Hacker News

Business relevance: Competitive intelligence in real-time, preventive crisis management, market trend detection



Spark on: Unified Batch + Streaming Processing

Enabling real time sentiment analysis + historical queries

- **PySpark** Structured Streaming was used to handle real-time Hacker News data, allowing continuous sentiment analysis at scale. It provided the framework for handling streaming DataFrames with windowed aggregations and watermarking for data handling.
- **Kafka** enabled a reliable data stream from the producer (sentiment analyzer) to the consumer (PySpark), ensuring no data loss during transmission.
- **MongoDB** was used to persist enriched sentiment analysis results for historical analytics.

Big Data Optimizations for Production-Scale Streaming

fetch posts and comments | **analyze sentiment (Producer side)** | **stream enriched data with sentiment** | **read pre-enriched data, aggregate, analyze** | **persist results**

- Spark's ability to handle large-scale streaming datasets in a distributed environment makes it ideal for real-time analysis. PySpark's Structured Streaming API allows us to process continuous data flows without compromising performance through batch processing controls.
- **Kafka + PySpark Integration**: TextBlob performs sentiment analysis on Hacker News comments in the producer before streaming enriched data to Kafka. By configuring maxOffsetsPerTrigger (1000 records) and trigger intervals (10 seconds), PySpark consumes already enriched messages in batch sizes.
- **MongoDB for Analytics**: As an alternative to temporary windowed views or in-memory Spark SQL tables, MongoDB provides persistent storage with server-side aggregation pipelines. This enabled efficient engagement analysis by sentiment and score buckets, identifying top positive posts with bulk writes (2500 records/batch) optimizing throughput for historical queries.

```
Fetching latest HN stories...
Processing 30 posts with sentiment analysis...
------------------------------------------------------------------------------
[1/30] Analyzing: Size of Life...
  → Score: 1138 | Comments: 20/79 | Sentiment: positive (0.270)
[2/30] Analyzing: Australia begins enforcing world-first teen social...
  → Score: 434 | Comments: 20/85 | Sentiment: neutral (0.033)
[3/30] Analyzing: Getting a Gemini API key is an exercise in frustra...
  → Score: 88 | Comments: 17/19 | Sentiment: positive (0.136)
[4/30] Analyzing: Super Mario 64 for the PS1...
  → Score: 123 | Comments: 9/9 | Sentiment: neutral (0.051)
[5/30] Analyzing: Auto-grading decade-old Hacker News discussions wi...
  → Score: 223 | Comments: 19/39 | Sentiment: positive (0.181)
[6/30] Analyzing: When would you ever want bubblesort? (2023)...
  → Score: 29 | Comments: 10/10 | Sentiment: neutral (0.006)
```

Real-time streaming Kafka with TextBlob sentiment analysis

```
✓ Top 10 most positive posts (unique titles):
+-------------------+-----+---------------+--------------------+--------------+
|avg_sentiment_score|score|sentiment_label|               title|total_comments|
+-------------------+-----+---------------+--------------------+--------------+
|             0.3027|  132|       positive|I got an Nvidia GH...|            17|
|             0.2838| 1101|       positive|        Size of Life|            77|
|             0.2484|  114|       positive|Gundam is just the...|            20|
|             0.2362|  149|       positive|Synadia and TigerB...|            12|
|             0.2345|   91|       positive|Factor 0.101 now ...|             7|
|              0.23|   55|       positive|How Google Maps al...|             7|
|             0.2084|    6|       positive|Should CSS be cons...|            12|
|             0.2059|   44|       positive|Scientists create ...|             6|
|             0.1912|   62|       positive|EFF Launches Age V...|             6|
|             0.1909|   79|       positive|Show HN: HCB Mobil...|            10|
+-------------------+-----+---------------+--------------------+--------------+
```

MongoDB querying

```
_id: ObjectId('693360a95cb920103da36794')
post_id : 46160315
title : "Netflix to Acquire Warner Bros"
url : "https://about.netflix.com/en/news/netflix-to-acquire-warner-bros"
score : 1310
author : "meetpateltech"
total_comments : 168
analyzed_comments : 20
avg_sentiment_score : 0.098
sentiment_label : "neutral"
timestamp : 2025-12-05T22:45:59.933+00:00
▸ sample_comments : Array (3)
engagement_ratio : 0.12814645308924486
is_controversial : false
hour_of_day : 22
```

Example of extracted insights per post (MongoDB collection)