Title: Real-Time Ratings Analysis using Spark, Kafka, and PySpark

Problem Description

In the current digital age, real-time analysis of user ratings is crucial for businesses to understand customer sentiment and improve their services. Our project aimed to analyze user ratings in real-time using Spark, Kafka, and PySpark.

Streaming Source

Our data source was a Kafka topic where each message represented a user rating. Each message was a JSON object with the following schema:

JSON

```
{
  "rating_id": Integer,
  "user_id": Integer,
  "stars": Integer,
  "route_id": Integer,
  "rating_time": Integer,
  "channel": String,
  "message": String
}
```

A sample message might look like this:

JSON

```
{
  "rating_id": 123,
  "user_id": 456,
  "stars": 4,
  "route_id": 789,
  "rating_time": 1591234567,
  "channel": "web",
  "message": "Great service!"
}
```

Design Considerations for Pipelines

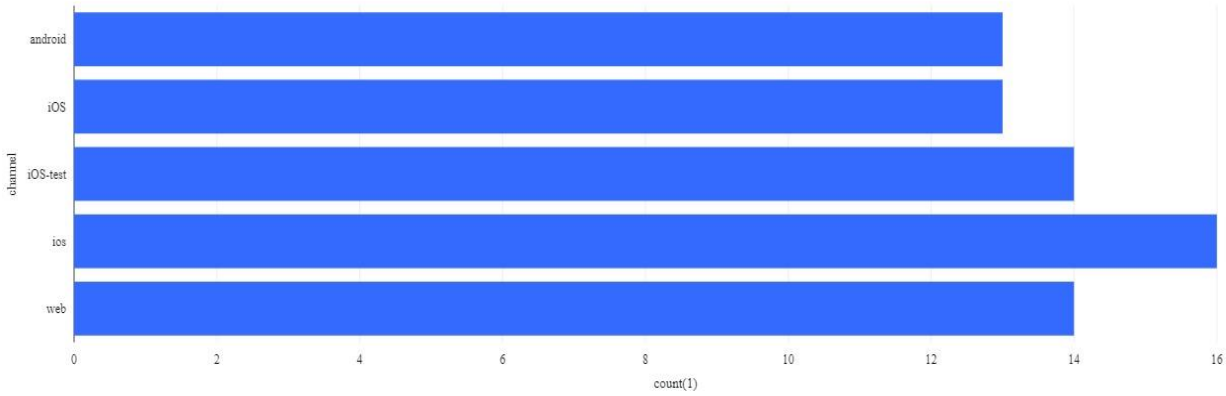
We designed two pipelines for our project. The first pipeline ingested streaming data from the Kafka topic into a staging zone periodically. We used Spark DataFrame APIs to read from Kafka and write to the staging zone.

The second pipeline read data from the staging zone, created insights using Spark SQLs, and wrote the results to a new location. We chose to use Spark SQLs for this pipeline because of its expressive power and simplicity.

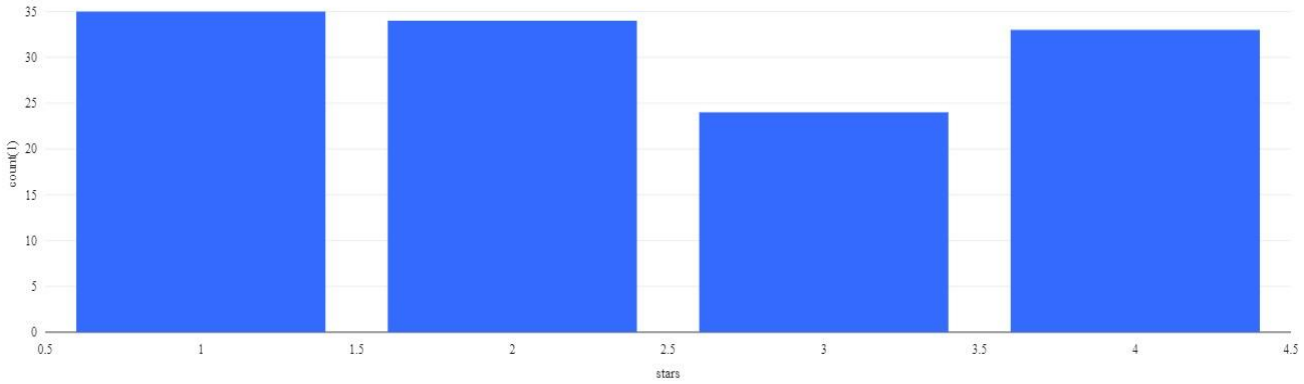
Business Insights

We derived several business insights from our data:

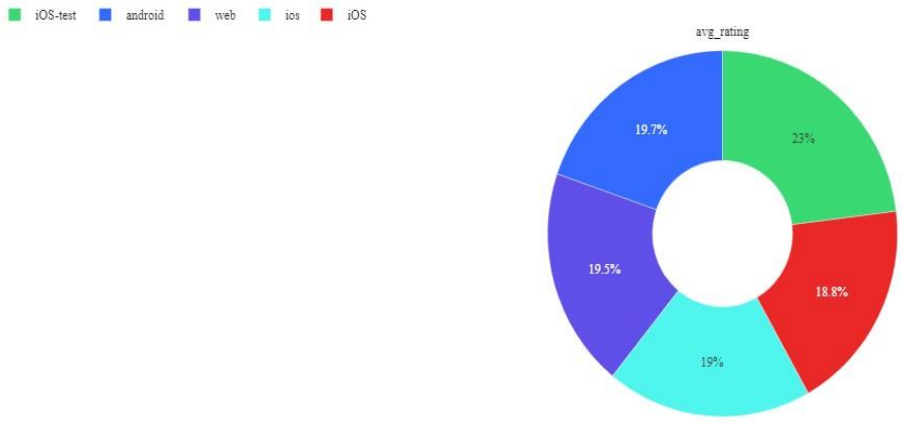
1. **Channels Used:** We analyzed the distribution of ratings across different channels. This helped us understand which channels were most popular among users.



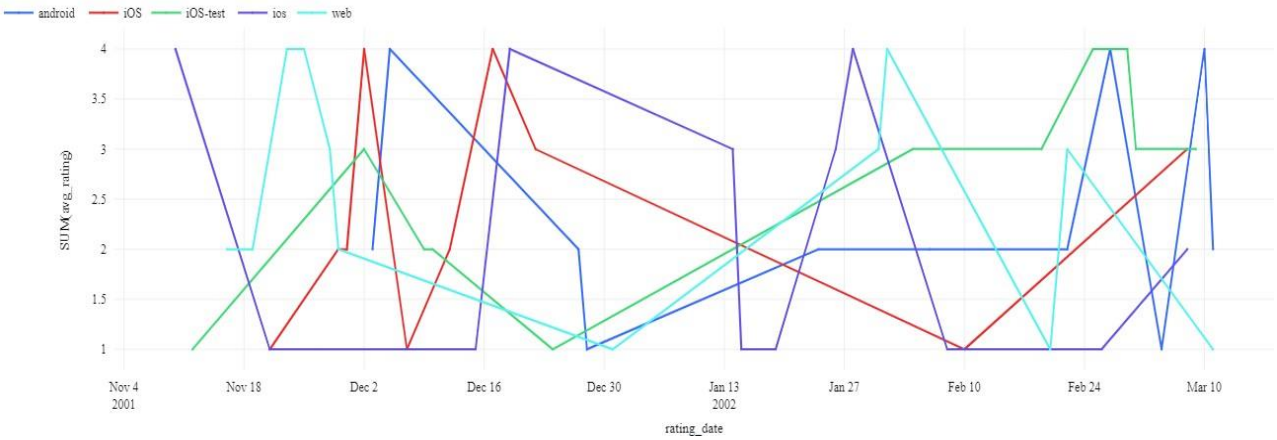
2. **Ratings:** We looked at the distribution of star ratings. This gave us an idea of overall user satisfaction.



3. **Channel wise avg rating:** We calculated the average star rating for each channel. This showed us how user satisfaction varied across different channels.



4. **Change in channel ratings over time (trend):** We analyzed how the average star rating for each channel changed over time. This helped us identify trends and patterns in user satisfaction.



We visualized these insights using various types of charts: bar charts for the distribution of channels and ratings, a pie chart for channel-wise average ratings, and a line chart for the trend of channel ratings over time.

Summary of Accomplishments

Our project was successful in achieving its goal of real-time ratings analysis. We were able to ingest streaming data, process it in real-time, and derive valuable insights. Our visualizations provided a clear and intuitive understanding of user ratings.

Lessons Learned

- Real-time data processing presents unique challenges and requires careful design of data pipelines.
- Spark and Kafka are powerful tools for working with streaming data.
- Choosing the right visualization can greatly enhance the understanding of data.

Team-

Salmoli Chandra 24PGAI0057

Meera Karamta 24PGAI0003

Kishnu Srivastava 24PGAI0042

Praveen Kumar Singh 24PGAI0068