Big Data Project

Big Data Project August 2025

Real-Time Hotel Review Monitoring

Presented by: Sharon, Roni & Isaac

Problem Statement

The Problem: Overwhelming Reviews – No Real-Time Insight

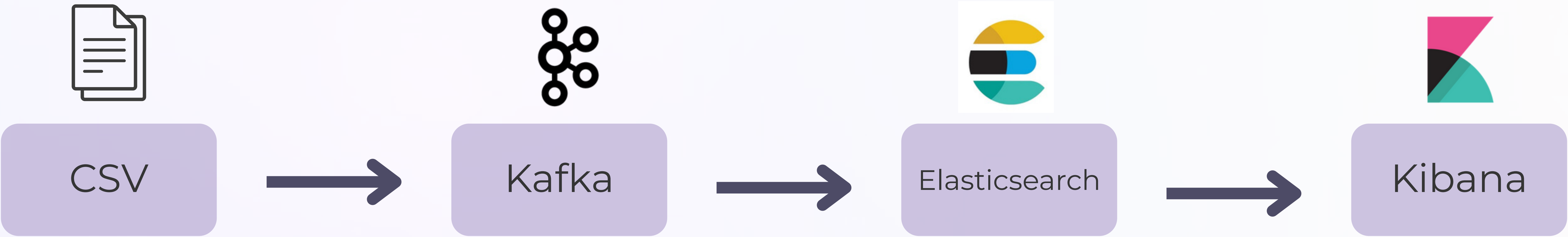
- Hundreds of thousands of hotel reviews are written daily
- Hotels can’t keep up with the volume and extract meaningful feedback
- Critical insights are missed or discovered too late

Our project focuses exactly on this gap:

turning ongoing reviews into near real-time signals that hotels can act on right away.

The Solution

Our Solution: A Near Real-Time Review Monitoring Pipeline



CSV → Kafka → Elasticsearch → Kibana

Overview of the Data Flow:

Data Producer (Python + Kafka):

- Reads hotel reviews from a large CSV file (500K+ reviews).
- Sends each review as a message to a Kafka topic (hotel-reviews), simulating a real-time stream.

Kafka Broker:

- Acts as a buffer/messaging layer between the data source and the processing engine.
- Ensures scalable, fault-tolerant data delivery.

Consumer (Python Kafka Consumer):

- Listens to the Kafka topic.
- Performs light preprocessing and data cleaning
- Sends the structured data to Elasticsearch.

Elasticsearch:

- Stores all processed reviews as JSON documents.
- Enables fast full-text search and aggregations.

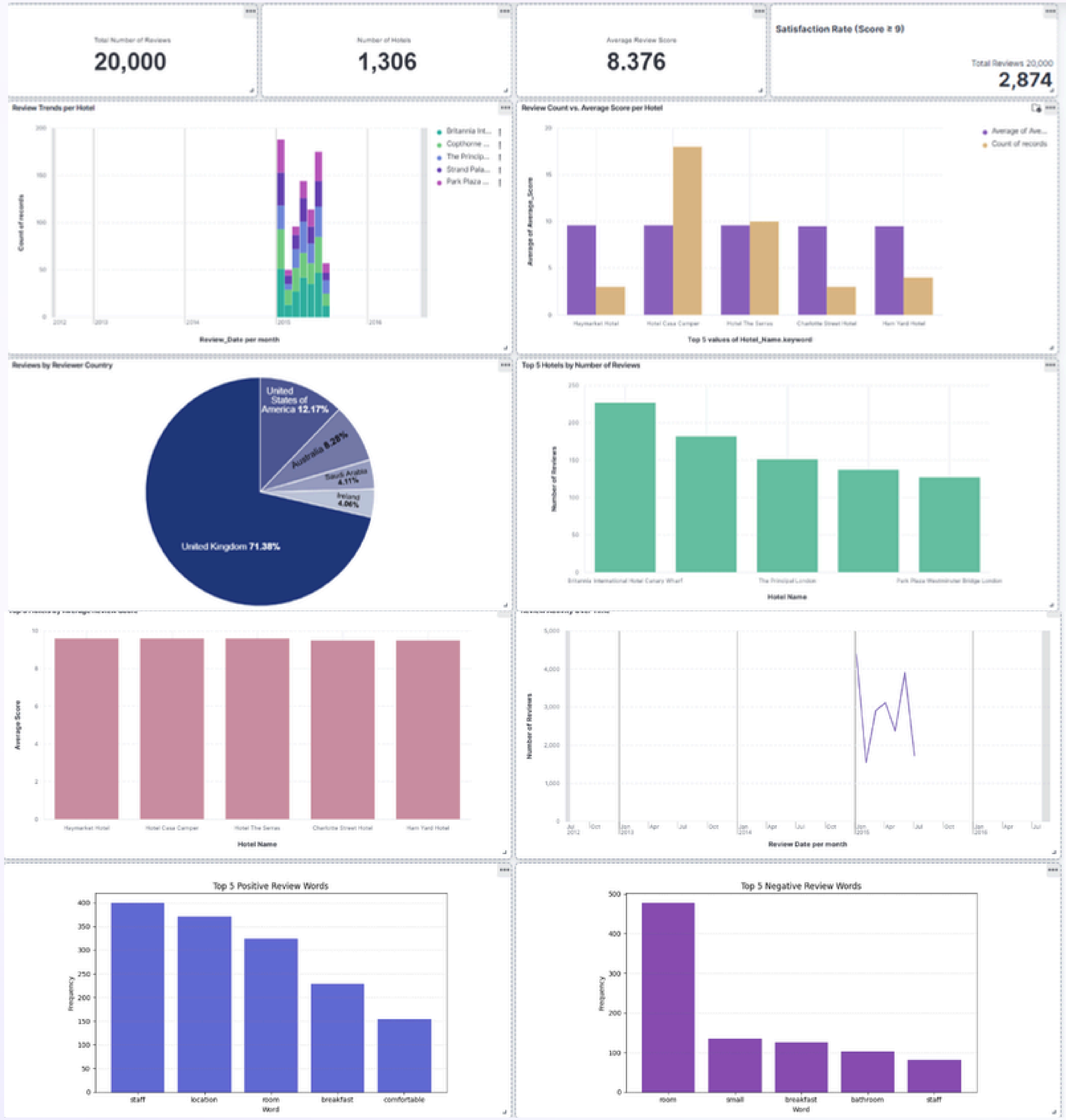
Kibana Dashboard:

- Visualizes the data in real time.
- Includes time series trends, keyword frequency analysis, reviewer demographics, and top hotels by rating and popularity.

Dashboard & Insights

Live Dashboard – Key Insights from the Data

Full Dashboard Overview



Total Number of Reviews
20,000

Number of Hotels
1,306

Average Review Score
8.376

Satisfaction Rate (Score ≥ 9)
Total Reviews 20,000
2,874

KPI's

Trends

Keywords

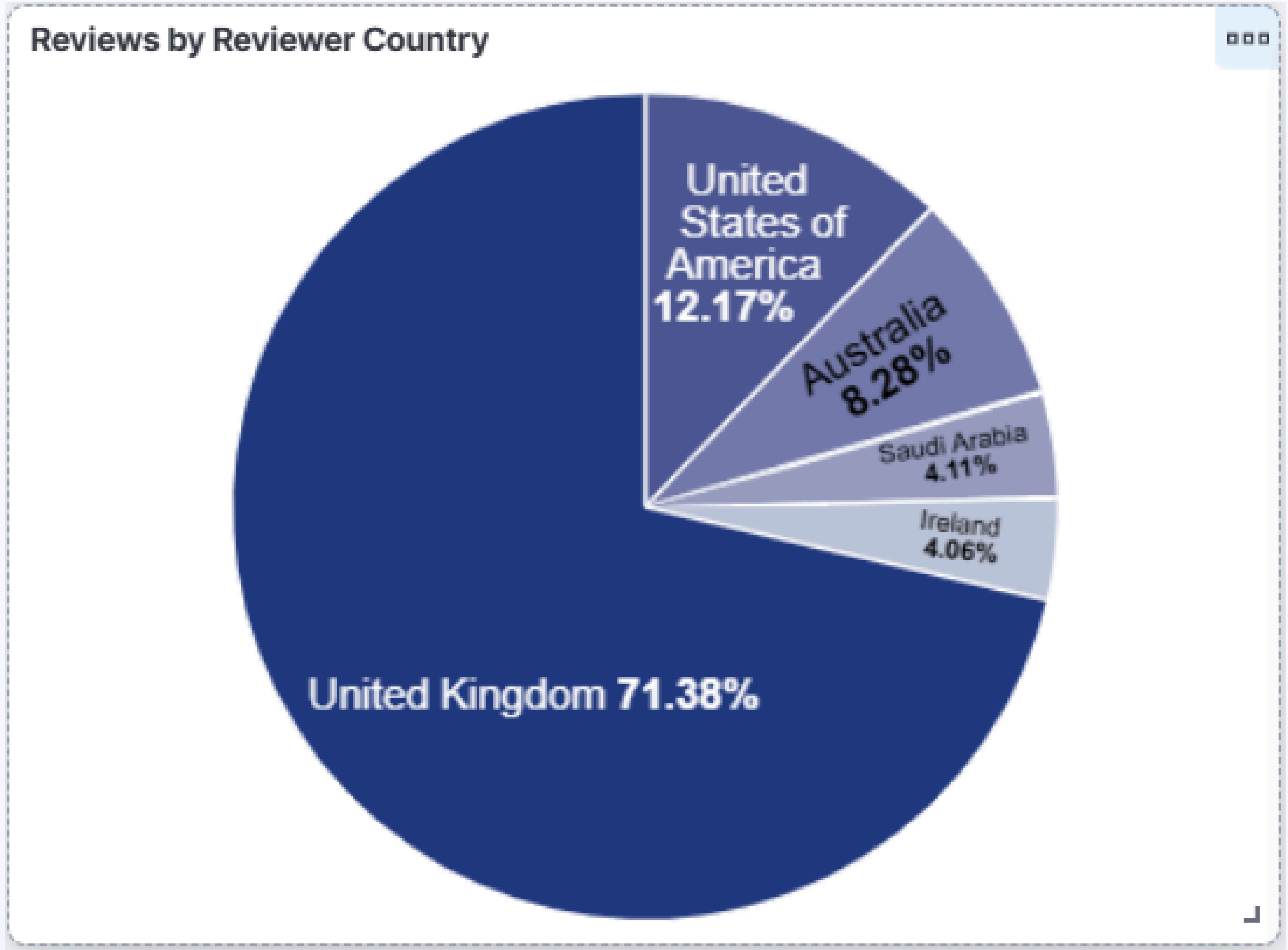
Key Insights from the Dashboard 💡

What Can We Learn from the Dashboard?

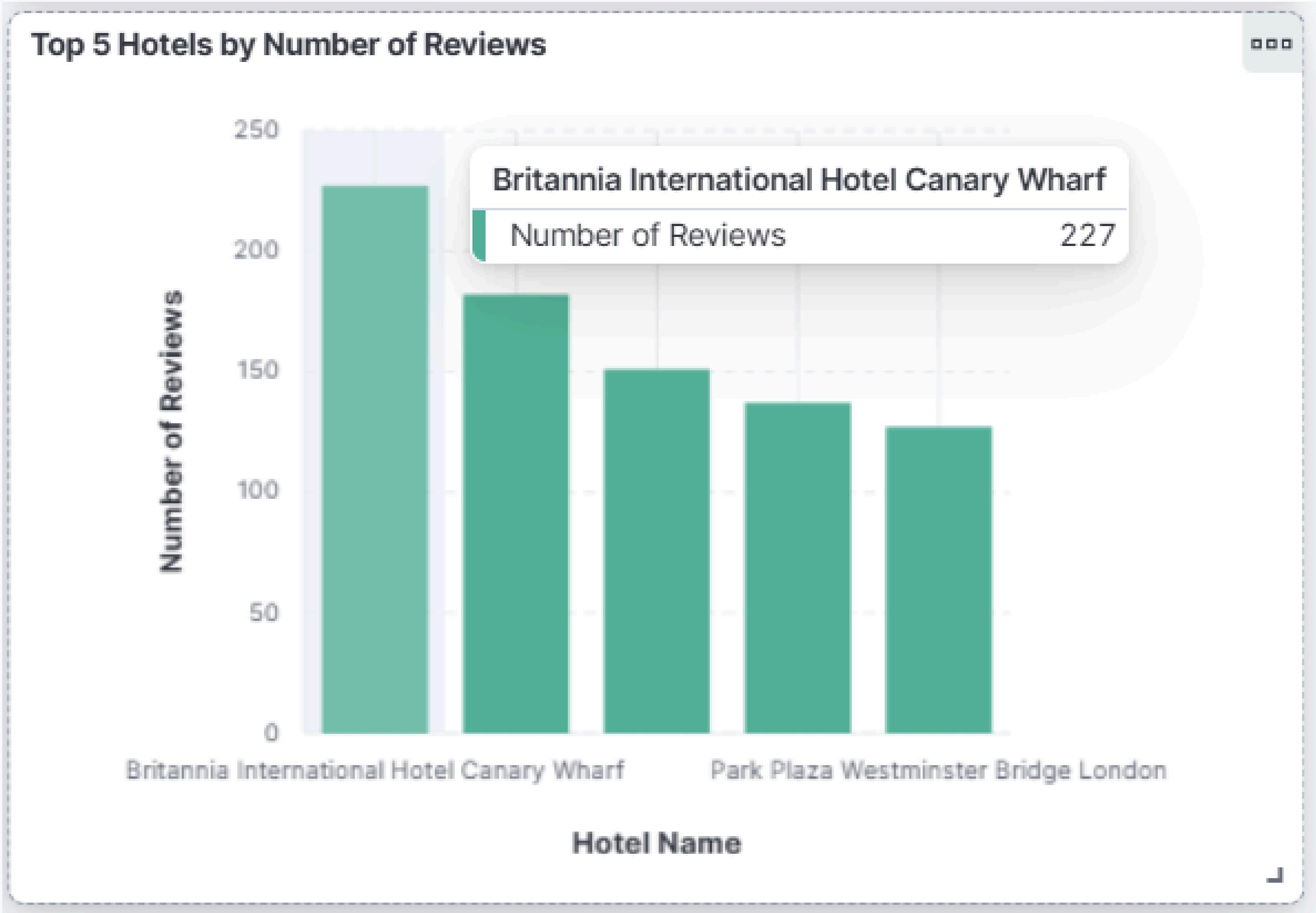
- Identify which hotels get the most attention and satisfaction
- Track review trends over time
- Understand the geographic distribution of reviewers
- Analyze what aspects guests care about – both positive and negative
- Compare score patterns across hotel types or regions

This makes it useful not just for analysis, but also for real decisions and improving the guest experience.

Examples from Our Dashboard



- Most reviewers are from the UK (71%)
- showing clear geographic dominance in the dataset



- Top-reviewed hotels are all in London
- with Britannia International leading the list

Challenges & Solutions

Challenge	Solution
• Learning unfamiliar tools (Kafka, Docker, Kibana)	• We started from scratch, setting up multi-container environments and debugging configuration issues using docker-compose.
• Preventing duplicate reviews	• We added a unique Review_ID to each review and set it as the document ID in Elasticsearch, ensuring safe re-ingestion.
• Missing libraries in Docker (e.g., pandas)	• We centralized all dependencies in requirements.txt and used it for automatic installation during container startup.
• Kibana not displaying live data	• We preprocessed and formatted Review_Date correctly, sorted the data, and adjusted Kibana’s time filters.
• Cleaning and preparing real-world text	• We built logic in the producer to standardize and clean the review fields (e.g., missing values, formatting).

Conclusion:

A Scalable, Flexible Monitoring Solution

- Demonstrates a near real-time pipeline for extracting insights from hotel reviews
- Useful for hotel chains, review platforms, and customer experience teams
- The dashboard is modular – easy to adapt, expand, and customize
- Anyone can add features, filters, or track specific trends
- The system is scalable, open-ended, and ready for real-world use

Most importantly, it shows how Big Data tools can turn raw reviews into actionable insights.

Questions?