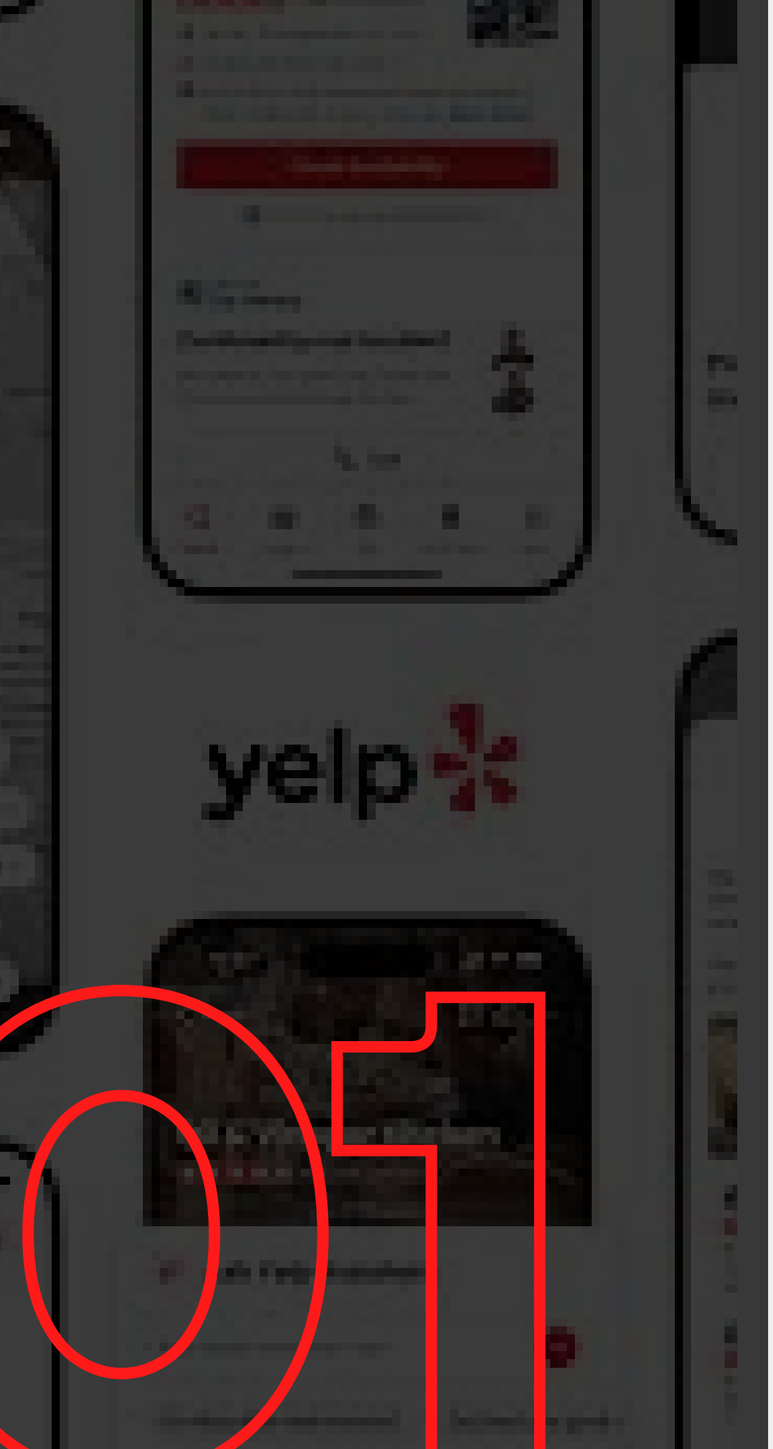
Scalable Big Data Analytics on Yelp Dataset :

A Pipeline for Real-Time Sentiment Analysis

Presented By :
Imen Kenza Chouaieb
Sara Sbissi

Supervised by : Pr. Manel Abdelkader



Business context

In today’s economy, user-generated content is the biggest driver of a business's reputation. However, the sheer scale like the 7 million reviews in the Yelp dataset, creates a massive 'data overhead. Manual analysis is no longer feasible, which means critical customer complaints often go unnoticed.

The objective of this project is to design and implement a big data analytics pipeline that processes large-scale review data efficiently and extracts meaningful insights for business decision-making.

yelp

Tech Stack

01. DATABRICK S :
unified orchestration layer for the entire pipeline.

02. APACHE SPARK :
a unified platform for batch processing, SQL analytics, and ML (MLlib)

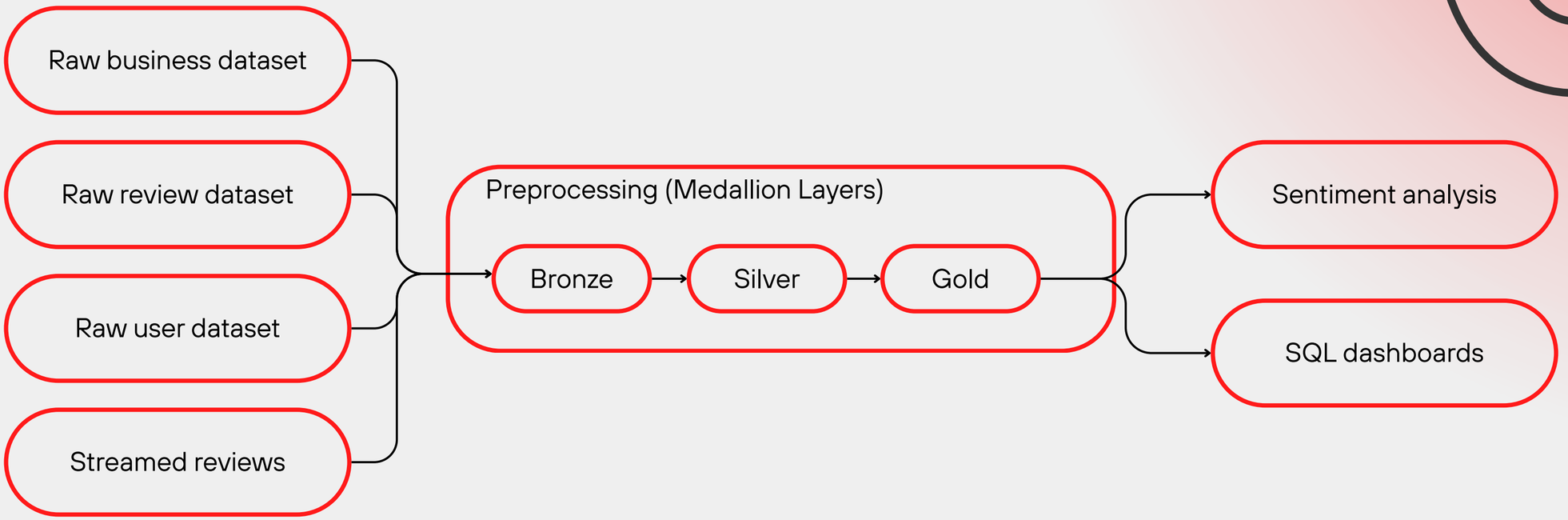
03. DELTA LAKE :
ACID-compliant storage layer on top of cloud object storage, addressing critical limitations of traditional data lakes

04. Unity Catalog :
provides enterprise-grade data governance capabilities, essential for production data platforms:

02

yelp

03



Pipeline Architecture

Data ingestion

The ingestion phase establishes the **Bronze layer** within a dedicated Unity Catalog namespace to ensure enterprise-grade governance.

Raw JSON files are staged in high-performance Unity Catalog Volumes, providing a secure landing zone decoupled from compute resources. To ensure reliability, the large-scale dataset was partitioned into manageable chunks for efficient cloud transfer and parallel processing.

Finally, explicit PySpark schemas were enforced during ingestion to guarantee structural integrity and accurate data typing from the outset.

yelp

Medallion Architecture

-The use of explicit PySpark StructTypes to ensure ingestion accuracy, the persistence of data in Delta tables for ACID compliance

-**Business Refinement:** Flattening complex nested maps like Operating Hours and Attributes (e.g., BusinessParking, Ambience, and GoodForMeal).

-Review Standardization: Handling the deduplication of identical reviews, using the absolute value function to fix negative engagement counts (useful/funny/cool).

User data normalisation ;
Flattending nested JSON into arrays in attributes like friends and elite.
Handling duplicate values
Derived new metrics, friend_count and elite_count,

Join the fact table (reviews) with dimension tables (business and user) to create high-performance tables optimized for your Databricks SQL Dashboards.

05

yelp

06

ML Results

- **Core objective:** Classify reviews into star ratings (1 to 5) to identify service failures or exceptional performance in real-time.
- **Selected algorithms:** Logistic regression & Random forest
- **Preprocessing Pipeline:**
  - StringIndexer: Converts star ratings into numerical labels.
  - Tokenizer: Breaks sentences into individual words and removes punctuation.
  - StopWordsRemover: Filters out non-semantic words (e.g., "the", "a").
  - HashingTF & IDF: Converts text into numerical feature vectors using TF-IDF logic.
- **Result: 46.6%** accuracy for RF and 65% accuracy for LR

07

Real-time streaming

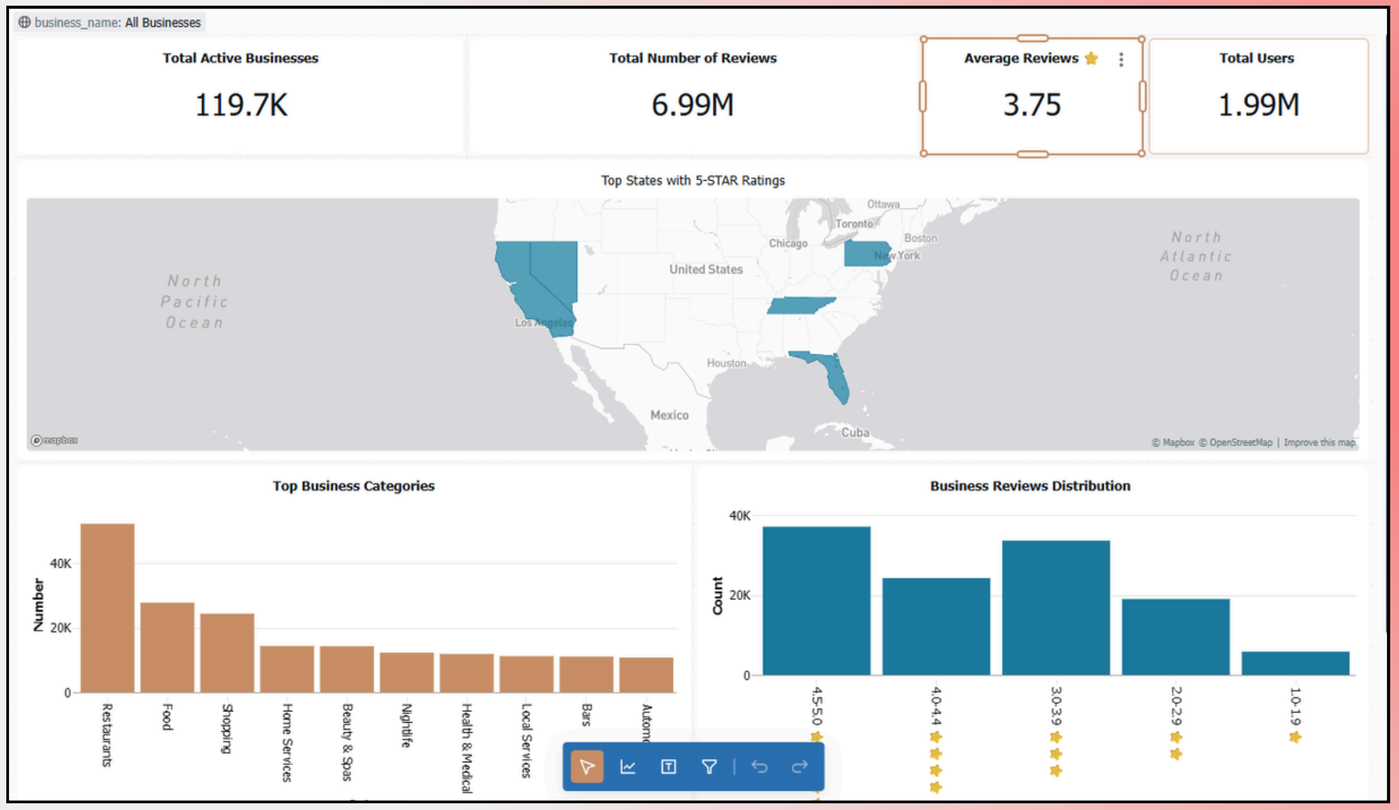
- **Ingestion Engine:** Databricks Auto Loader is used to bypass file listing bottlenecks and automatically detect schema changes in raw JSON data.
- **Framework:** DLT manages the data flow via a Directed Acyclic Graph (DAG), handling orchestration and error handling automatically.
- **Data Quality:** Implementation of DLT expectations to ensure records lacking a review_id are filtered out before reaching the analytics stage.
- **Fault Tolerance:** Manual Structured Streaming with a checkpointLocation tracks offsets, allowing the system to resume exactly where it left off after any restart.
- **Real-time Business Value:** The final Gold Layer performs running calculations of average stars and counts the reviews, reflecting the current reputation as soon as new reviews arrive.

08

KPIs & Dashboard

yelp

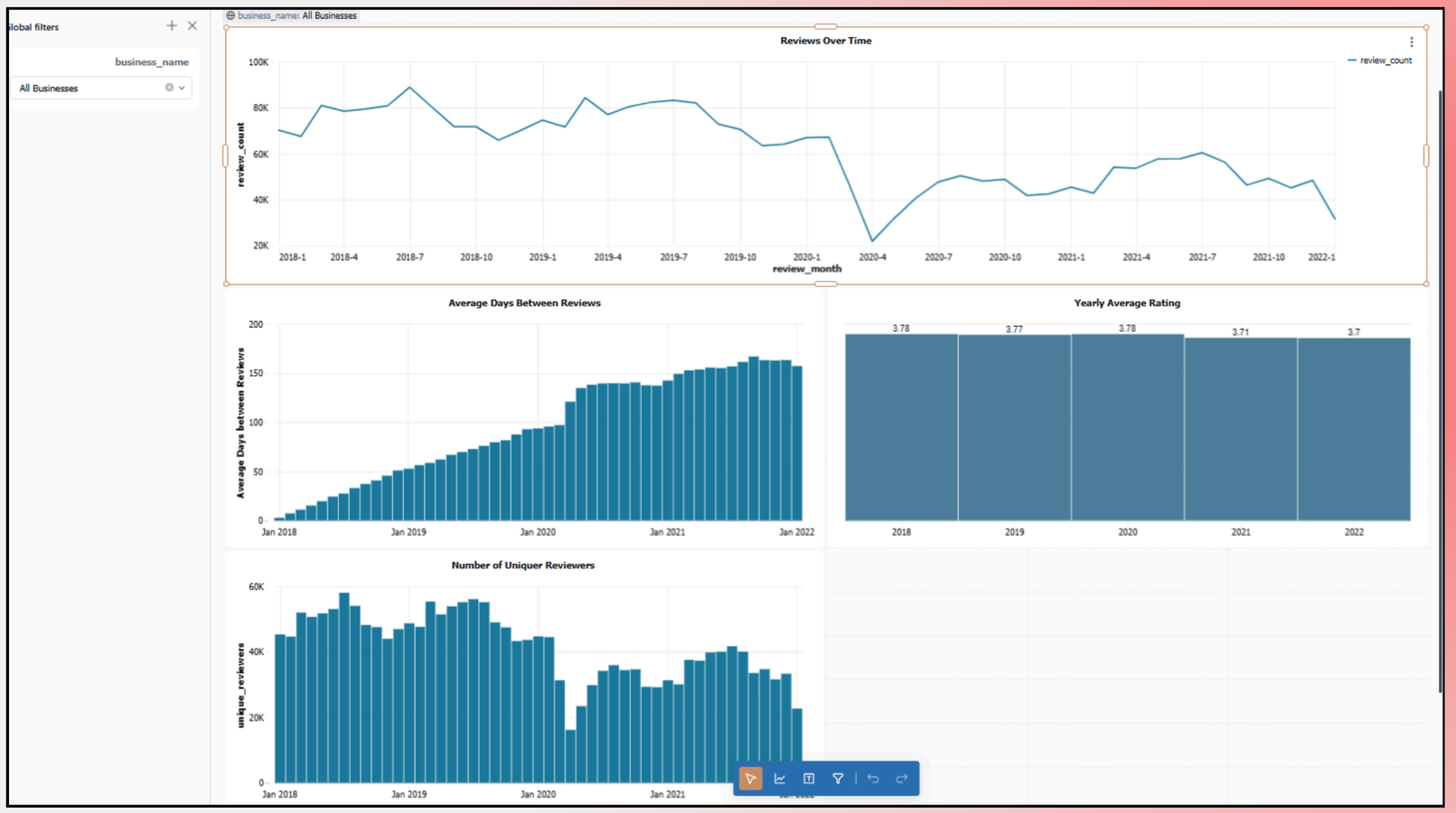
Businesses Overview Dashboard



08

KPIs & Dashboard

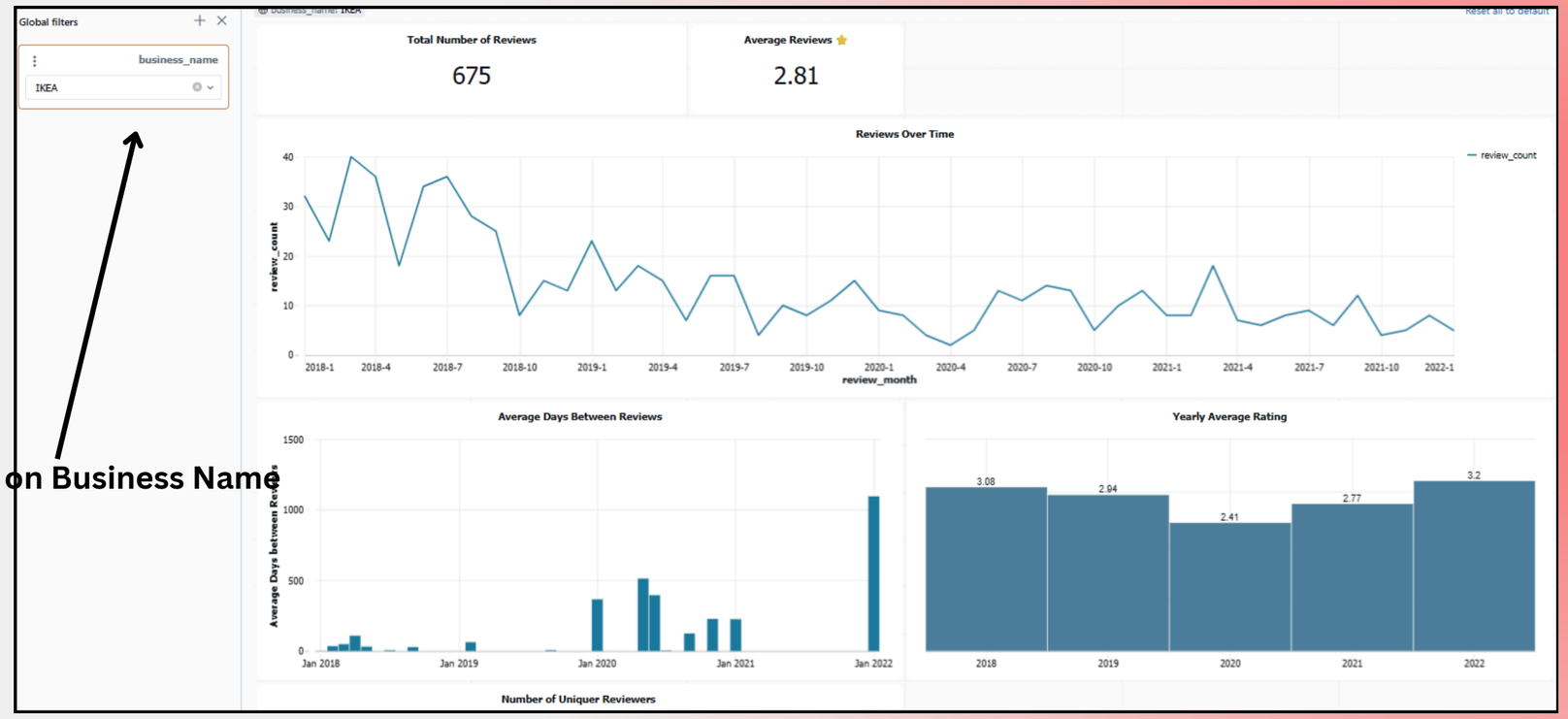
Reviews Dashboard



08

KPIs & Dashboard

Reviews Dashboard

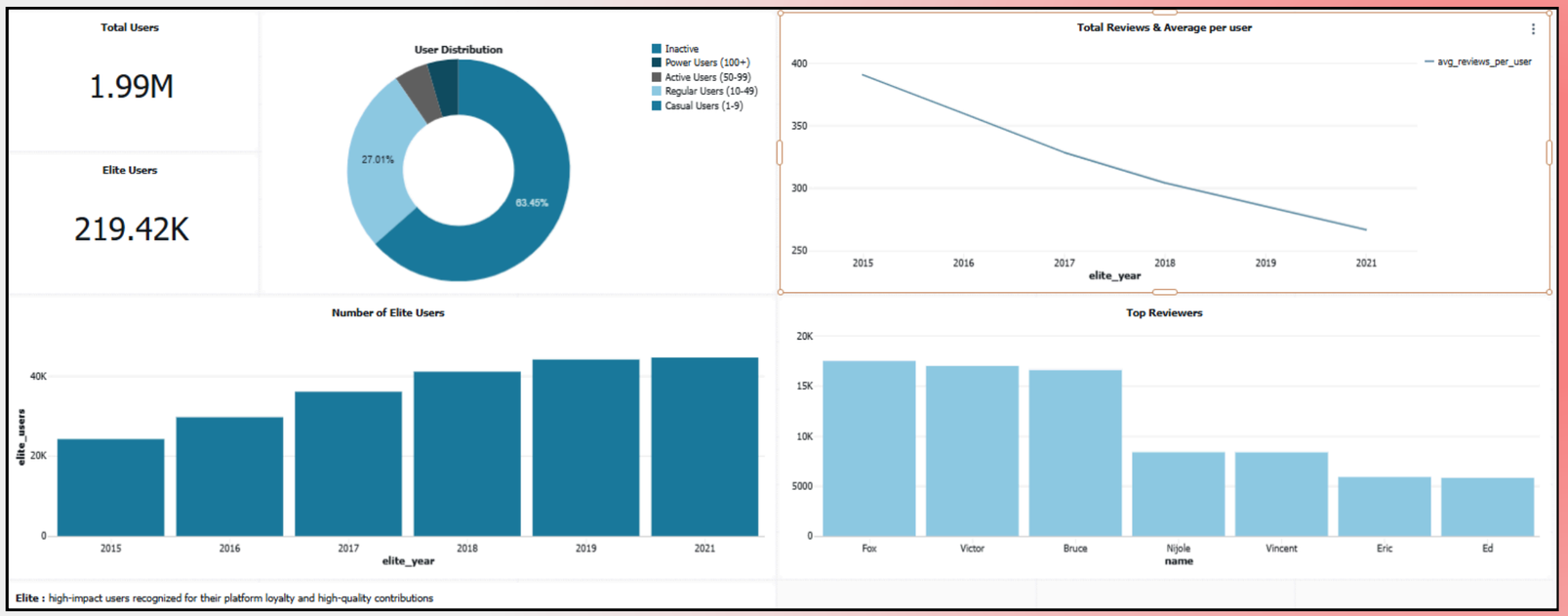


Filter on Business Name

08

KPIs & Dashboard

Users Dashboard



09

Conclusion

- **Scalable Engineering Foundation:** Developed a production-ready Medallion architecture using Apache Spark and Delta Lake to transform 7M+ raw records into a governed, high-performance "Single Version of Truth."
- **Automated Intelligence & BI:** Integrated PySpark MLlib sentiment classification and interactive SQL Dashboards to convert unstructured reviews into actionable KPIs for over 119,000 businesses.
- **Future-Proof Roadmap**: Designed a horizontally scalable system ready for real-time structured streaming, advanced NLP migration, and multi-node cluster expansion for datasets exceeding 50M+ records.

Thank You

Any Questions?