MET CS777

Big Data Analytics

Term Paper Presentation

Analyzing and Visualizing Big Data Performance in the Cloud : Google BigQuery vs. Databricks Spark SQL

Team:
Tanvi Thopte
Mehul Bisht

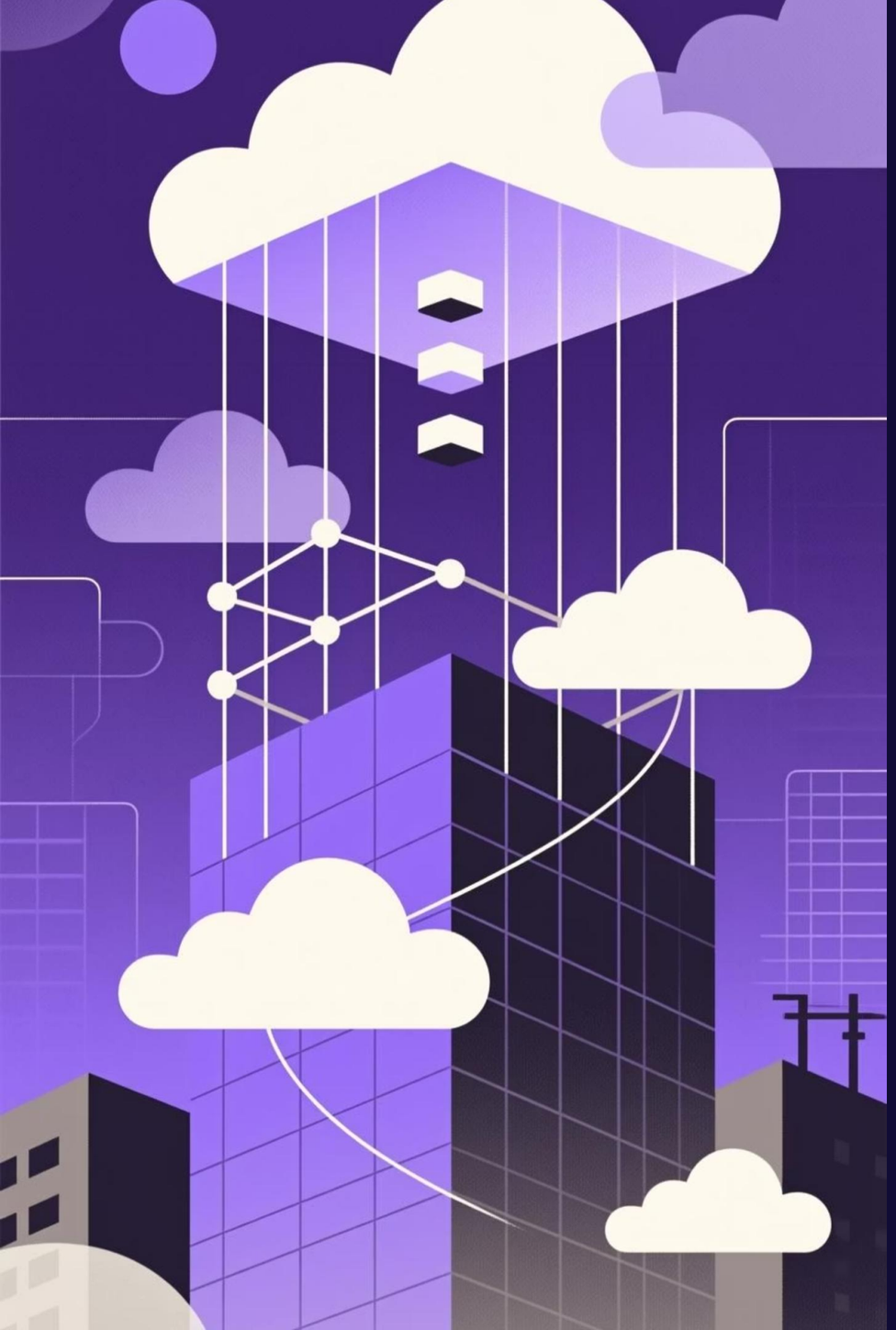
Introduction: Cloud Analytics for Ecommerce Big Data

Project Goal

To evaluate and compare the performance and visualization efficiency of Google BigQuery (serverless) and Databricks Spark SQL (cluster-based) for cloud-based big data analytics on a healthcare dataset, focusing on query execution time, scalability, and insight visualization.

Importance in Ecommerce

Big data analytics helps online retailers analyze user behavior, category performance, and revenue trends across millions of transactions in real time.



Platform Genesis

Both BigQuery and Databricks emerged from the need to manage and query massive datasets efficiently using distributed computing paradigms.

2011: Google BigQuery Launch

Introduced by Google as a serverless, highly scalable, and cost-effective enterprise data warehouse running on the Dremel engine.

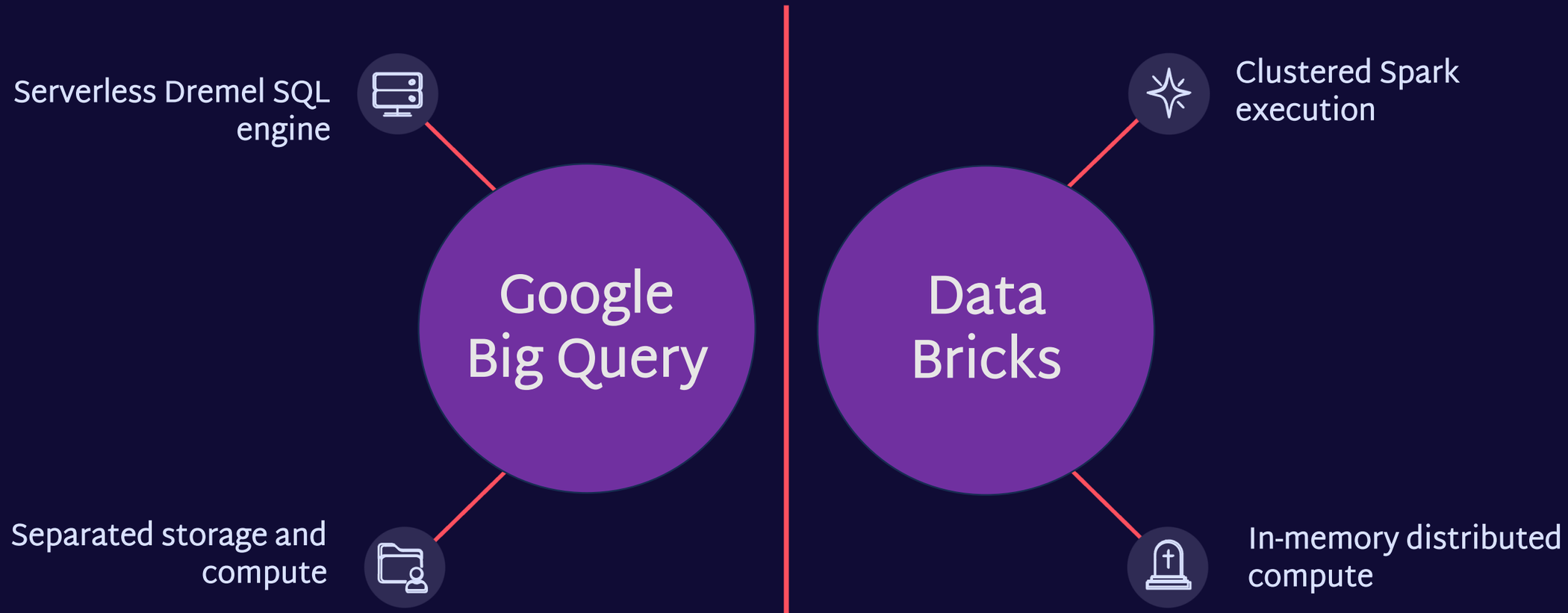
2013: Databricks Founding

Founded by the creators of Apache Spark. Focused on unifying data engineering, data science, and machine learning workloads on a single platform.

Modern Distributed Analytics

Both platforms have evolved to become central components in modern cloud data architectures, offering petabyte-scale analysis capabilities.

Technology Overview and Architectural Comparison



Google BigQuery (GBQ)

- **Engine:** Dremel (SQL-based MPP).
- **Architecture:** Serverless; computation and storage are separated.
- **Data Storage:** Columnar storage for optimized analytical queries.
- **Key Benefit:** Zero-management and automatic scaling.

Databricks Spark SQL

Engine: Apache Spark (unified analytics engine).

Architecture: Cluster-based; requires cluster configuration and management.

Data Storage: Uses Delta Lake for transactional data layer; leverages in-memory processing.

Key Benefit: High flexibility and control over compute resources.

Primary Use Cases and Application Focus

Google BigQuery Focus

Ideal for high-speed, interactive analytics, and Business Intelligence (BI) tools. Excellent for dashboards and ad-hoc SQL queries on structured data.

- Fast Analytics
- Interactive Dashboards (Looker)
- Retail Inventory Analysis

Databricks Focus

Primarily designed for complex data pipelines, large-scale ETL/ELT, and integrated Machine Learning (ML) workflows using notebooks and the Lakehouse architecture.

- Data Engineering / ETL
- Advanced ML Models (MLflow)
- Financial Fraud Detection

Dataset and Experimental Methodology

Our comparison utilized a standardized methodology to ensure parity in execution and measurement.

The Dataset

Dataset: E-commerce event data (Oct 2019–Apr 2020), 7 months of user purchase logs with 9 columns (event_time, event_type, product_id, category_id, category_code, brand, price, user_id, user_session).
Size: tens of GBs (millions of rows).

Context: Focus on patient demographics, diagnoses, procedures, and readmission outcomes.

Data Ingestion

Upload to respective cloud storage: GCS (for GBQ) and DBFS (for Databricks).

Pre-processing

Standardized cleaning, feature engineering, and type casting were performed identically across both platforms.

Query Execution

Four identical, complex analytical SQL queries were executed 10 times to capture average performance metrics.

Metric Recording

Metrics recorded included query execution time (latency), bytes processed, and platform scalability observed.

Visualisation performed using Looker studio and Python

Analytical Tasks: Four Core SQL Queries

The test suite involved increasingly complex SQL queries designed to stress the processing engines' capabilities on a healthcare dataset.

Q1 Daily revenue trend (time-series)

Q2 Category × Brand aggregation

Q3 User-level monetization (total spend)

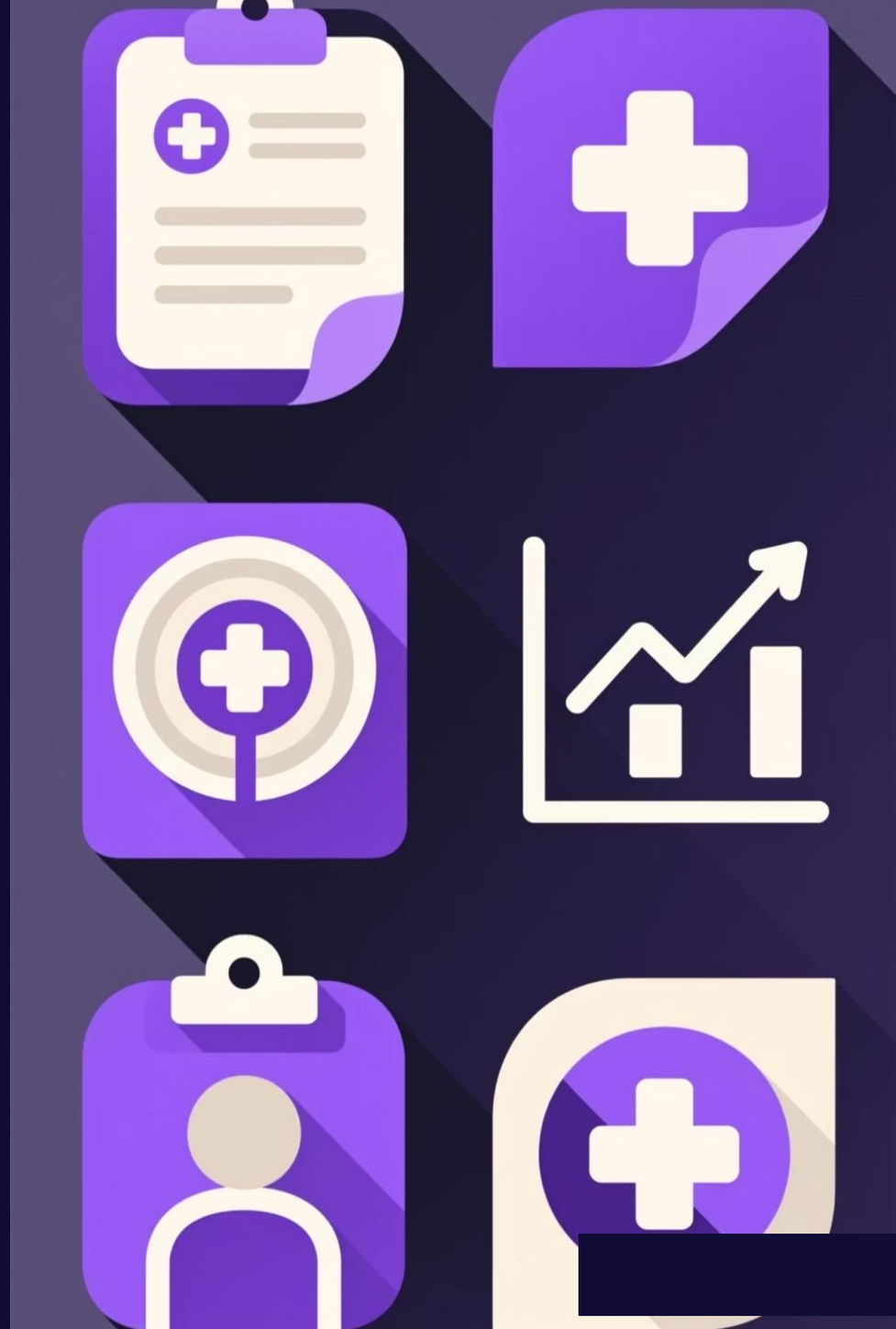
Q4 Session analytics (basket size/value)

Q5 Trend + anomaly detection (z-score method)

Q6 Scalability test (vary date range)

Q7 Top K products by revenue

Q8 Brand-level outlier analysis



Results and Visualization of Query Performance

In this experiment:

BigQuery excelled in speed, scalability, and simplicity, best for analytics at scale.

Databricks excelled in flexibility, caching, and visualization, best for exploration and experimentation.

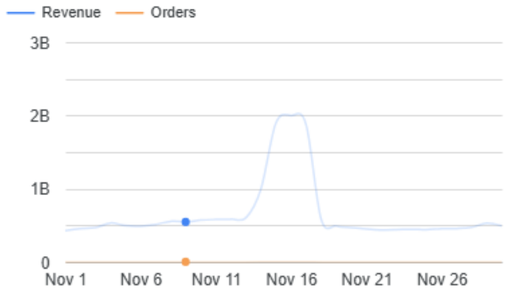
Both platforms are powerful, choosing between them depends on whether your goal is rapid analytics.

Results and Visualization of Query Performance

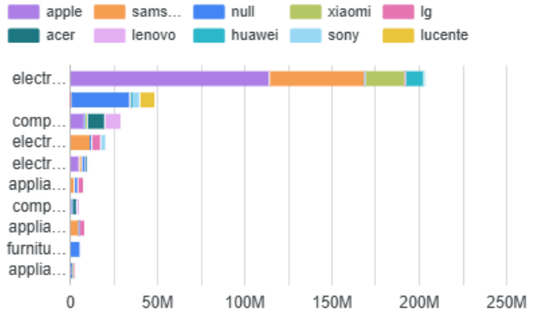
Query / Task	Big Query Execution Time	Databricks Execution Time
Table Creation (events_clean)	~18s	~23 s
EDA Queries (0–4)	12sec total	9sec total
Daily Revenue Trend (5)	~761ms	~648s
Category × Brand Aggregation	~351ms	~1s
User-Level Monetization	~8 s	~3s
Session Analytics	~7s	~7.2s
Anomaly Detection	~513ms	~563ms
Scalability Test (1-Month Data)	~2s	~1 s
Scalability Test (3-Month Data)	~534ms	~2s
Scalability Test (7-Month Data)	~564ms	~3s

Results and Visualization of Query Performance

Daily Revenue Trend



Category × Brand Revenue

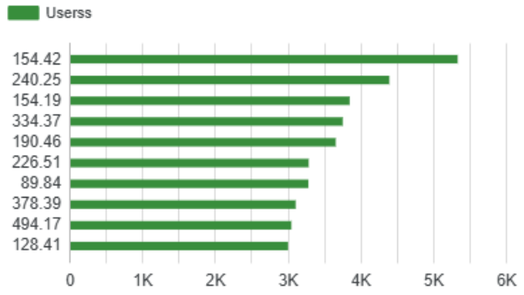


Top users by total spending

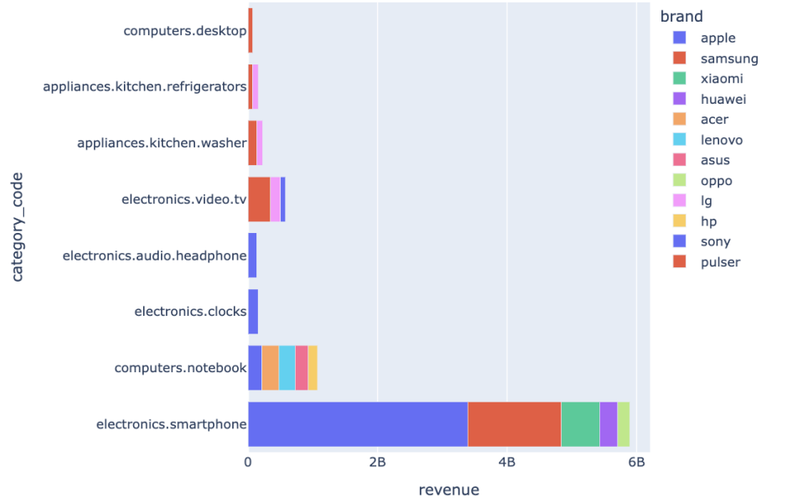
	user_id	Revenue ▾	Avg Tic...
1.	568797382	212,631.41	1,012.53
2.	516054872	191,874.31	1,148.95
3.	569335945	172,308.04	121.09
4.	562850008	168,034.34	852.97
5.	568793129	164,944.46	420.78
6.	554501441	158,196.68	1,068.9
7.	546635249	154,455.24	718.4

1 - 100 / 268643

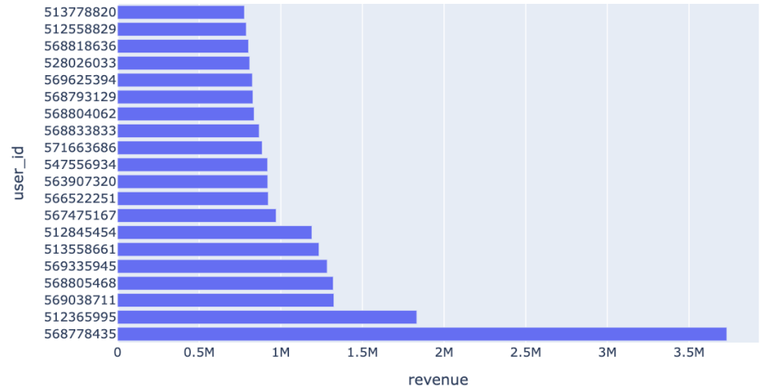
Spending distribution histogram



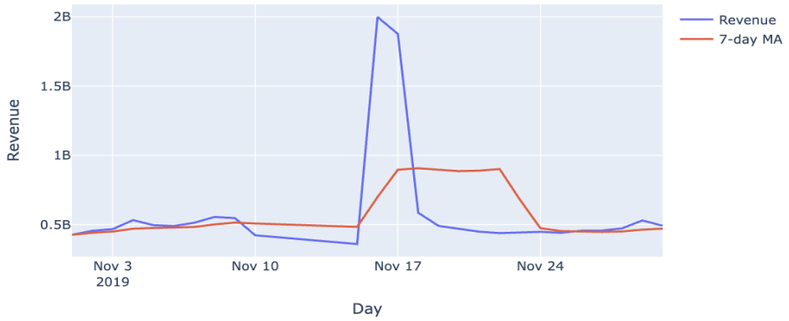
Category × Brand Revenue (Top 20)



Top Users by Total Spending



Daily Revenue Trend (with 7-day MA & Anomalies)



Discussion: Comparing Operational Strengths

Google BigQuery Advantages

- **Ease of Use:** Near-instant setup with serverless provisioning. Minimal overhead.
- **Speed for Ad-Hoc Queries:** Optimized for fast, read-only analytical SQL queries on columnar data.
- **Cost Structure:** Pay-per-query model can be highly cost-effective for burstable, high-volume analytic tasks.

Databricks Spark SQL Advantages

- **Flexibility:** Complete control over cluster types, sizes, and auto-scaling rules.
- **ETL and ML Integration:** Superior platform for complex data transformation pipelines and seamless transition into machine learning model training.
- **Language Support:** Full support for Python, R, Scala, and SQL in notebooks, favouring data scientists.

Both platforms demonstrated excellent scalability and handled the dataset with minimal latency difference once Databricks clusters were provisioned.

Conclusion & Future Work

Summary of Key Findings

BigQuery

Optimal for structured, serverless SQL analytics and immediate BI reporting.

Databricks

Preferred for complex data engineering (ETL) and advanced ML/AI workflows.

Future Scope and Extensions

Predictive Modeling: Incorporate machine learning tasks (e.g., predicting readmission) to evaluate Databricks' MLOps advantage.

Cost-Performance Analysis: Conduct a detailed cost breakdown on petabyte-scale datasets under different pricing tiers.

Multi-Cloud Comparison: Extend the study to include analytics platforms like Snowflake or Azure Synapse for a broader industry benchmark.

Thank You for Your Attention.

THANK YOU