TITLE: REAL-TIME DATA PROCESSING AND ADVANCED ANALYTICS

SUBTITLE : FINAL REPORT PRESENTATION

NAME: SUDHIR YADAV

DATE : 27/02/2025

AGENDA

- Introduction and Problem Statement
- Data Source and Ingestion
- Pipeline Architecture
- Data Storage and Management
- Data Processing and Transformation
- Key Insights and Data Analysis
- Results and Business Impact
- Scalability and Future Work
- Challenges and Lessons Learned
- Conclusion and Final Takeaways

Azure DataServices Used

Blob Storage

Data Factory

DataBricks

Storage Account

Microsoft Azure

Event Hub

SQL Database

PowerBI

CosmosDB

INTRODUCTION:

IN TODAY'S FINANCIAL LANDSCAPE, REAL-TIME DATA PROCESSING AND ADVANCED ANALYTICS ARE ESSENTIAL FOR DETECTING FRAUD AND UNDERSTANDING CUSTOMER BEHAVIOR. THIS PROJECT LEVERAGES AZURE SERVICES TO BUILD A SCALABLE SOLUTION FOR REAL-TIME TRANSACTION MONITORING, PREDICTIVE ANALYTICS, AND INSIGHTFUL REPORTING.

PROBLEM STATEMENT:

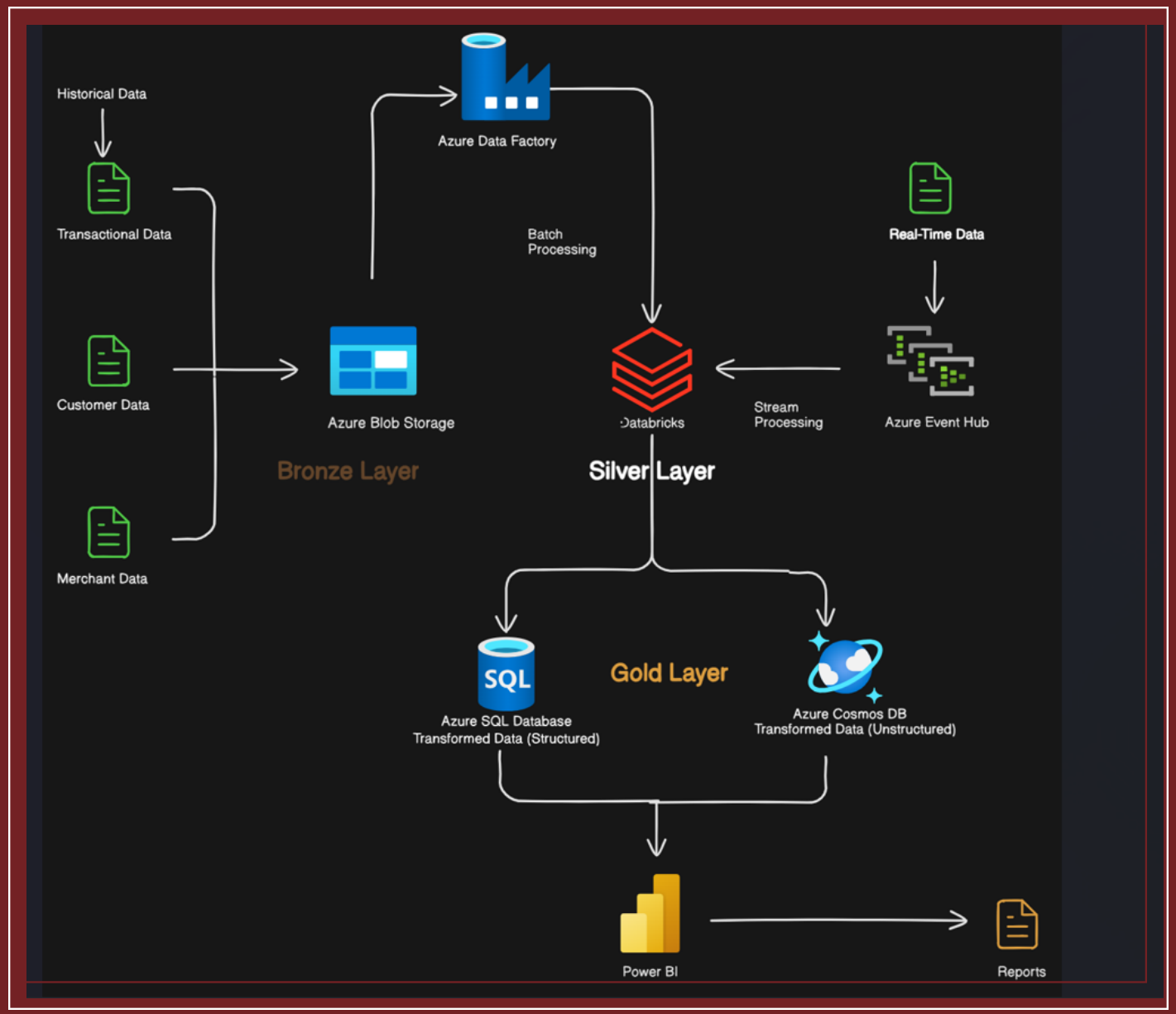
A FICTIONAL FINANCIAL INSTITUTION PROCESSES MILLIONS OF TRANSACTIONS DAILY AND NEEDS TO ANALYZE THIS DATA FOR FRAUD DETECTION, RISK ASSESSMENT, AND REGULATORY COMPLIANCE. THE PRIMARY CHALLENGE IS TO EFFICIENTLY AGGREGATE, PROCESS, AND STORE BOTH HISTORICAL AND REAL-TIME DATA STREAMS TO DERIVE ACTIONABLE INSIGHTS AND OPTIMIZE OPERATIONAL EFFICIENCY.

OBJECTIVES:

- INGEST HISTORICAL DATA FROM AZURE BLOB STORAGE.
- STREAM REAL-TIME DATA FROM AZURE EVENT HUB FOR REAL-TIME ANALYTICS.
- TRANSFORM DATA USING AZURE DATABRICKS FOR CLEANING, AGGREGATION, AND OPTIMIZATION.
- OPTIMIZE DATA STORAGE BY IMPLEMENTING COST-EFFICIENT MEDALLION ARCHITECTURE.
- GENERATE VISUAL REPORTS USING POWER BI TO PROVIDE BUSINESS INSIGHTS.

DATA SOURCE AND INGESTION

- Data Sources: Historical data (fraud detection and compliance data) and real-time data (transaction alerts).
- Ingestion Methods: Use of Azure Event Hub for real-time data streaming and Azure Data Factory for batch processing.



ARCHITECTURE

DATA STORAGE AND OPTIMIZATION

STORAGE ARCHITECTURE:

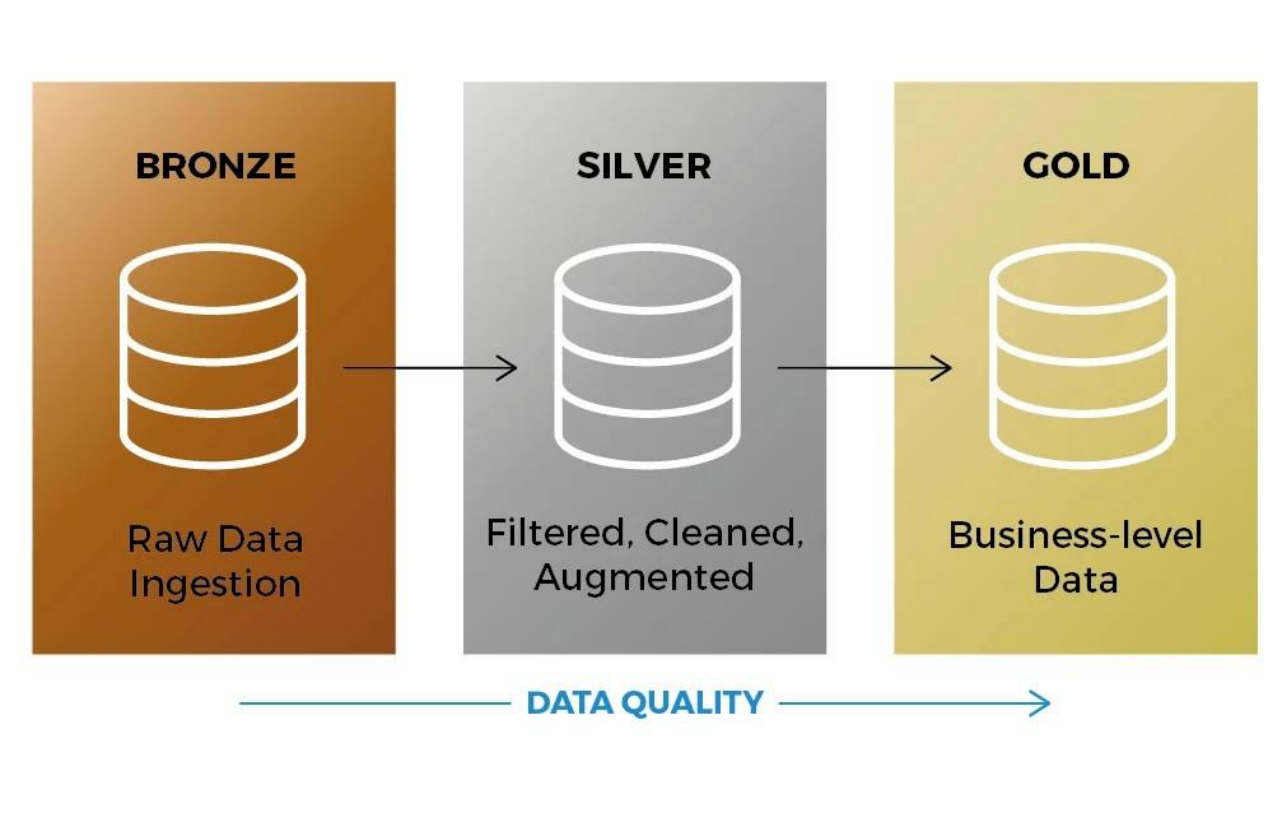
MEDALLION APPROACH

- BRONZE LAYER: AZURE BLOB STORAGE FOR RAW HISTORICAL AND REAL-TIME DATA.
- SILVER LAYER: AZURE DATABRICKS FOR CLEANING AND TRANSFORMING DATA.
- GOLD LAYER: AZURE SQL DATABASE AND AZURE COSMOS DB FOR AGGREGATED DATA ANALYSIS.

OPTIMIZATION TECHNIQUES:

- PARTITIONING AND INDEXING FOR FAST QUERYING IN SQL.
- HIERARCHICAL NAMESPACE TO SHOW MEDALLION ARCHITECTURE.
- EVENT HUB AUTO-SCALING TO REDUCE REAL-TIME DATA INGESTION COSTS.

MEDALLION ARCHITECTURE



REAL-TIME DATA PROCESSING AND TRANSFORMATION

Processing Pipeline:

- Azure Event Hub receives JSON data from real-time sources.
- Azure Databricks performs transformation and cleaning.
- Azure SQL Database stores processed real-time data for reporting.

Triggering Events:

- Data Arrival: New JSON transaction added, process and store data.
- Fraud Alert: High-value anomaly detected, trigger alert.

Transformation:

- Transformation: Explanation of data transformation processes and ensuring data quality.
- Measures: Steps taken to maintain data integrity and accuracy.

KEY INSIGHTS AND DATA ANALYSIS

High-Level Insights

- **Transaction Volume Trends**: Identified peak transaction periods and analyzed growth trends.
- **Fraud Detection Rates**: Highlighted periods with high anomaly detection and common fraud patterns.
- **Customer Analytics**: Top customers, demographic insights, and churn analysis.

Visuals

- **Transaction Volume**: Line chart showing transaction volumes over time and bar chart comparing segments.
- **Fraud Detection**: Heatmap for high-risk periods and pie chart for fraud detection rates.
- **Customer Analytics**: Bubble chart for top customers and histogram for demographic distribution.

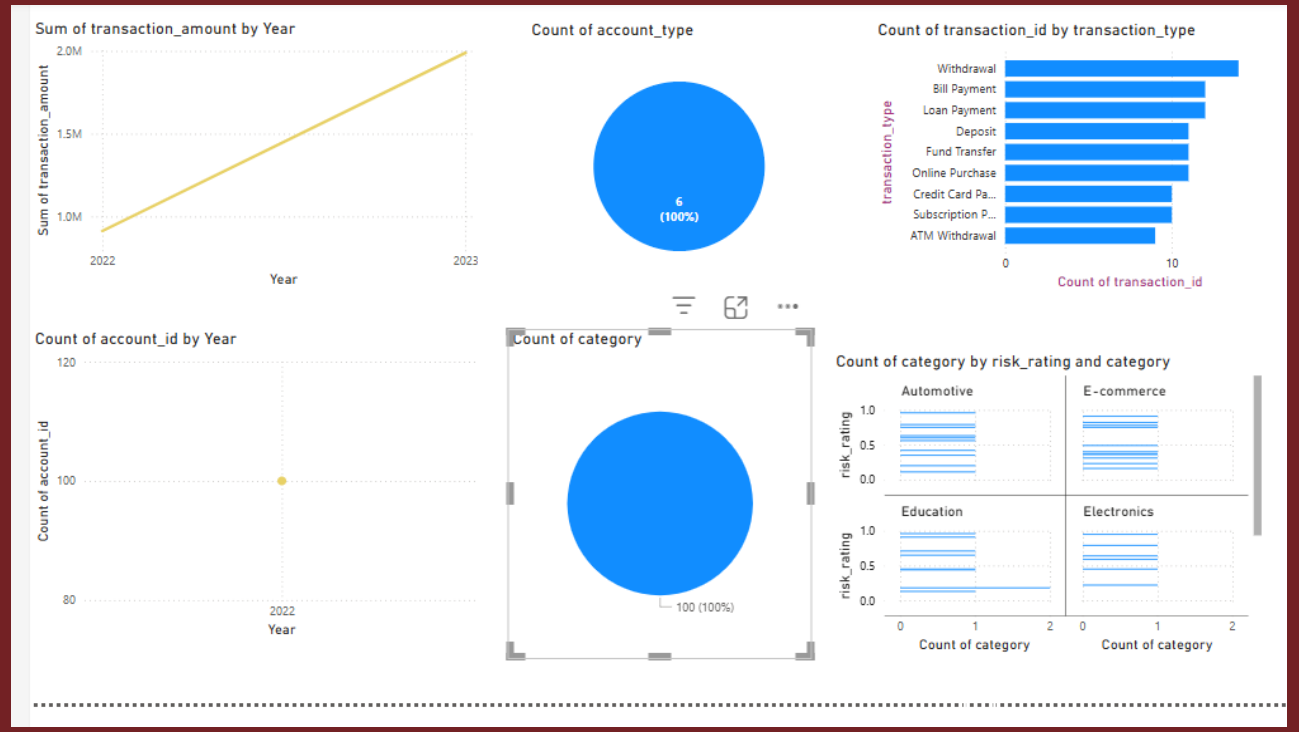
RESULTS AND BUSINESS IMPACT

Main Outcomes

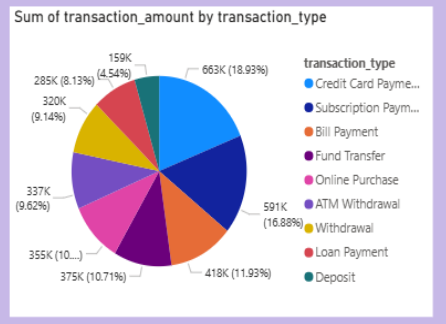
- **Optimized Data Pipeline: Achieved faster query performance and improved real-time processing.**
- **Accurate Fraud Detection: Enhanced anomaly detection and identified common fraud patterns.**
- **Customer Insights: Gained valuable demographic and behavioral insights for better customer targeting.**

Business Relevance

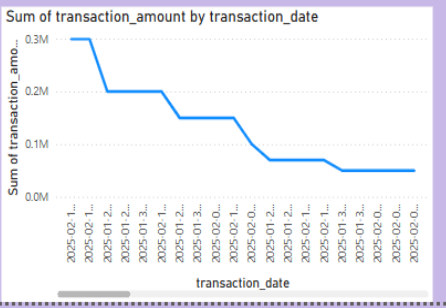
- **Operational Efficiency**: Streamlined data processing, resulting in time and cost savings.
- **Risk Management**: Strengthened fraud detection mechanisms and proactive risk assessment.
- **Decision-Making**: Provided data-driven insights for strategic business decisions.
- **Customer Satisfaction**: Increased satisfaction through tailored services and personalized experiences.



transaction_id	transaction_type	merchant_id	account_id	Sum of transaction_amount	transaction_date
RT_TXN20037	Bill Payment	MER2074	ACC1004	300000	2025-02-13 16:21
RT_TXN20054	Fund Transfer	MER2065	ACC1084	300000	2025-02-17 02:58
RT_TXN20004	Subscription Payment	MER2060	ACC1011	200000	2025-01-30 18:34
RT_TXN20015	Credit Card Payment	MER2042	ACC1097	200000	2025-01-26 05:48
RT_TXN20040	Credit Card Payment	MER2013	ACC1042	200000	2025-01-28 15:02
RT_TXN20084	Loan Payment	MER2087	ACC1081	200000	2025-02-15 15:01
RT_TXN20008	Withdrawal	MER2043	ACC1056	150000	2025-01-30 12:26
RT_TXN20031	ATM Withdrawal	MER2035	ACC1060	150000	2025-01-29 12:39
RT_TXN20080	Online Purchase	MER2074	ACC1071	150000	2025-02-19 15:57
RT_TXN20092	Subscription Payment	MER2061	ACC1019	150000	2025-02-06 10:26
RT_TXN20034	Credit Card Payment	MER2006	ACC1066	100000	2025-02-01 22:43
RT_TXN20036	Credit Card Payment	MER2018	ACC1078	70000	2025-02-10 01:41
Total				3502900	



4M
Sum of transaction_amount

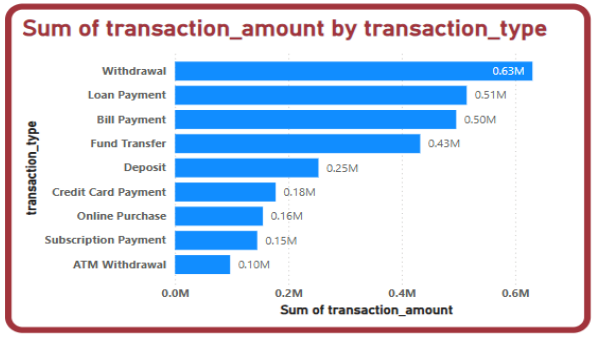


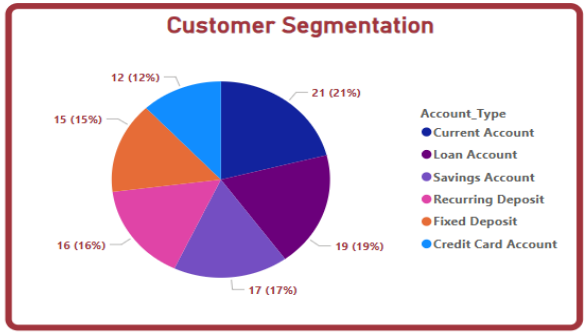
3M
Total Transaction Amount

29.01K
Average Transaction Amount

Total Transaction Amount and Target Transaction Amount by transaction_date
50000
Goal: 40000 (+25%)

100
Number of Transactions



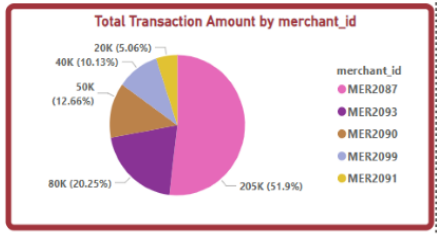


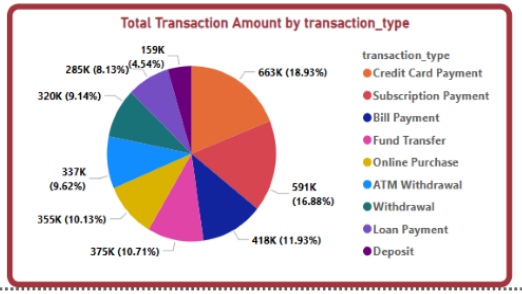
Real Time Transaction Amount by time and type

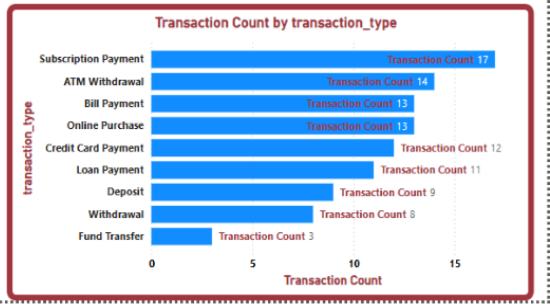
transaction_type	January	February	Total
ATM Withdrawal	205500	131500	337000
Bill Payment	14000	404000	418000
Credit Card Payment	485000	178000	663000
Deposit	35000	124000	159000
Fund Transfer		375000	375000
Loan Payment	41200	243500	284700
Online Purchase	102500	252500	355000
Subscription Payment	276000	315200	591200
Withdrawal	185000	135000	320000
Total	1344200	2158700	3502900

REAL-TIME DATA POWER BI REPORT

4M
Total Transaction Amount







SCALABILITY AND FUTURE WORK

- **Scalability**: Use of Azure Virtual Machine Scale Sets and Databricks auto-scaling.
- **Future Enhancements**: Suggestions for future improvements, such as advanced security features and additional data integrations.

CHALLENGES AND LESSONS LEARNED

Challenges

1. **Real-Time Processing**: Ensuring low latency for real-time data streams.
2. **Data Integration**: Integrating diverse data sources seamlessly.

Lessons Learned

1. **Scalability**: Importance of designing scalable architectures.
2. **Automation**: Benefits of automating data processes.
3. **Data Quality**: Ensuring data quality for accurate analytics.

CONCLUSION AND FINAL TAKEAWAYS

Summary

- **Project Findings**:
  - Implemented a scalable ETL and real-time analytics solution for BFSI.
  - Optimized data processing, storage, and real-time fraud detection.
  - Provided valuable business insights through data analysis.
- **Recommendations**:
  - Integrate machine learning for predictive fraud detection.
  - Optimize data storage costs with tiered solutions.
  - Enhance security with role-based access control (RBAC).

Final Takeaways

- **Efficient Data Processing**: Managed large volumes of transaction data effectively.
- **Real-Time Insights**: Enabled real-time analytics for proactive fraud and compliance management.
- **Business Impact**: Delivered insights that improved efficiency and customer satisfaction.
- **Scalability**: Created a solution that scales with growing data demands.

THANK YOU