CAPSTONE PROJECT

Credit Card Approval Analysis

BY VALLURU UDAY KIRAN

PROBLEM STATEMENT

To create a comprehensive Data Pipeline with

- **Azure Data Factory** and
- **Databricks**

in order to create a **Dashboard** for Credit Card Approval *Analysis* and to derive useful business insights from it

DATA SOURCES

The data sources used are:

- **Applicants Data** –Blob
- **Applicants Data**–HTTP (GitHub)
- **Credit Data**– – SQL table – SQL DB

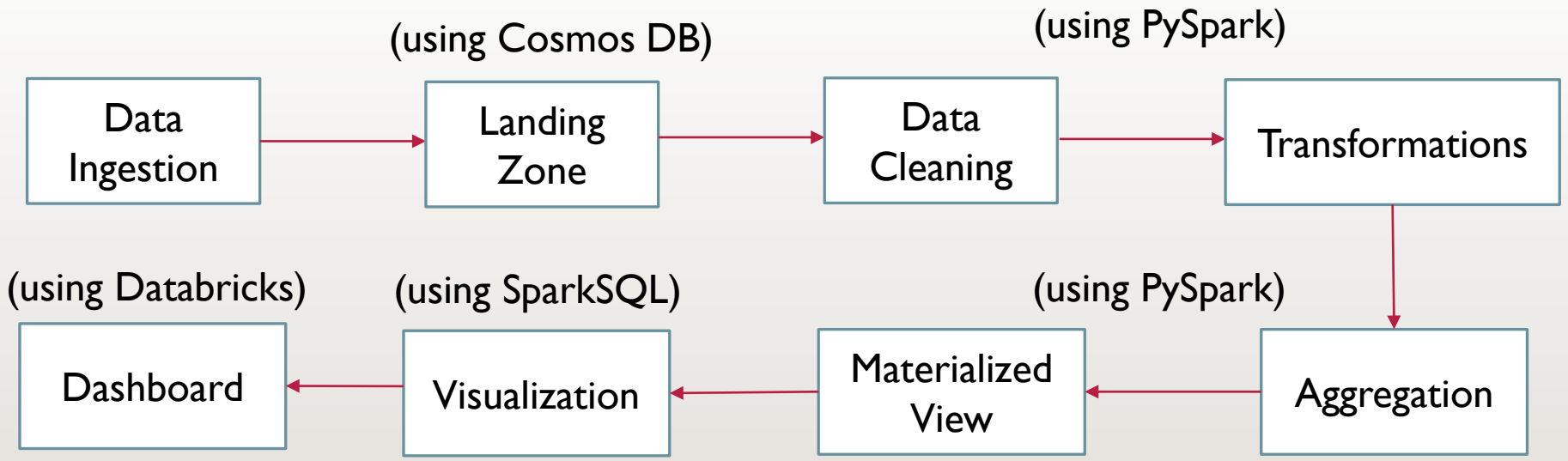
SOLUTION FLOW

Tech Stack

- **Azure Storage** – For Data Ingestion & Storage
- **Azure Data Factory** – For Pipeline
- **PySpark** – For Transformations and Aggregation
- **SparkSQL** – For Visualization
- **Azure Databricks** – For Dashboard

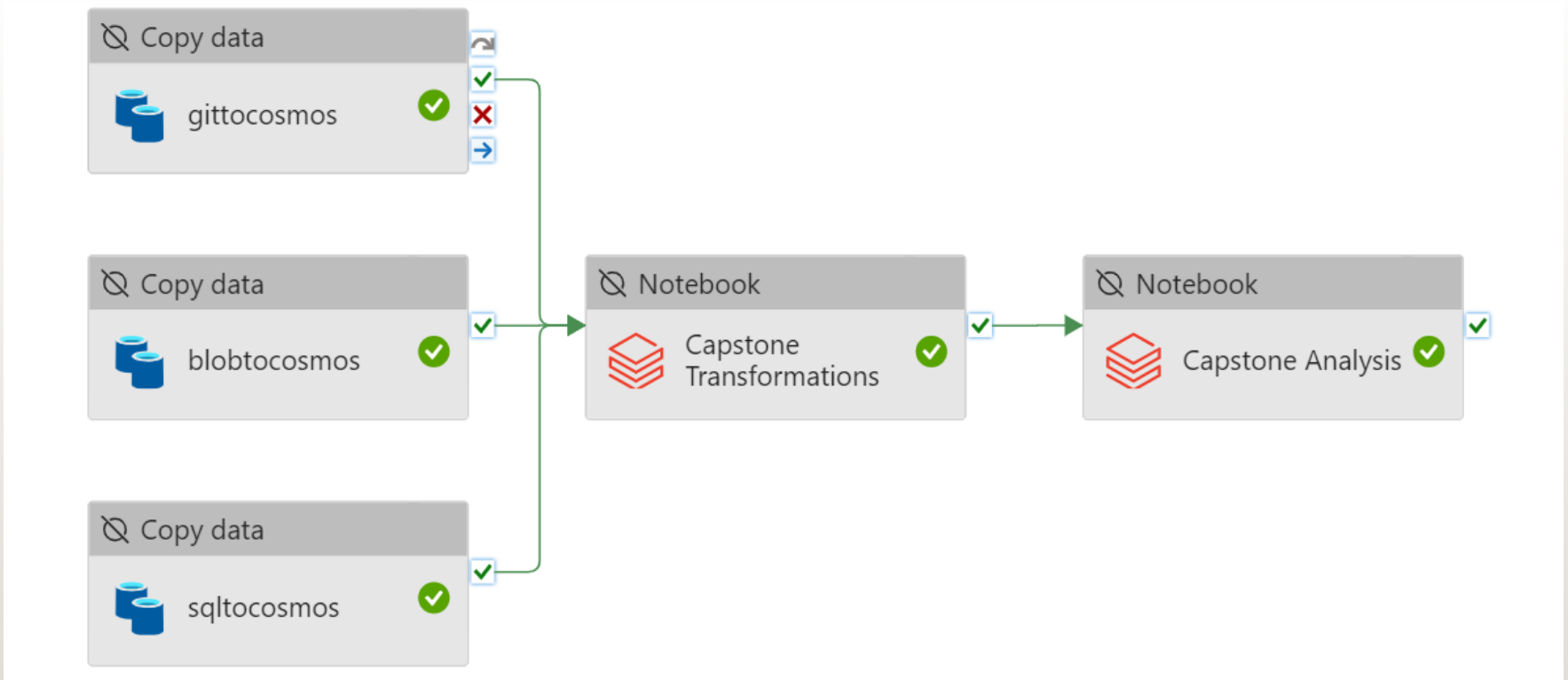
SOLUTION FLOW

Pipeline



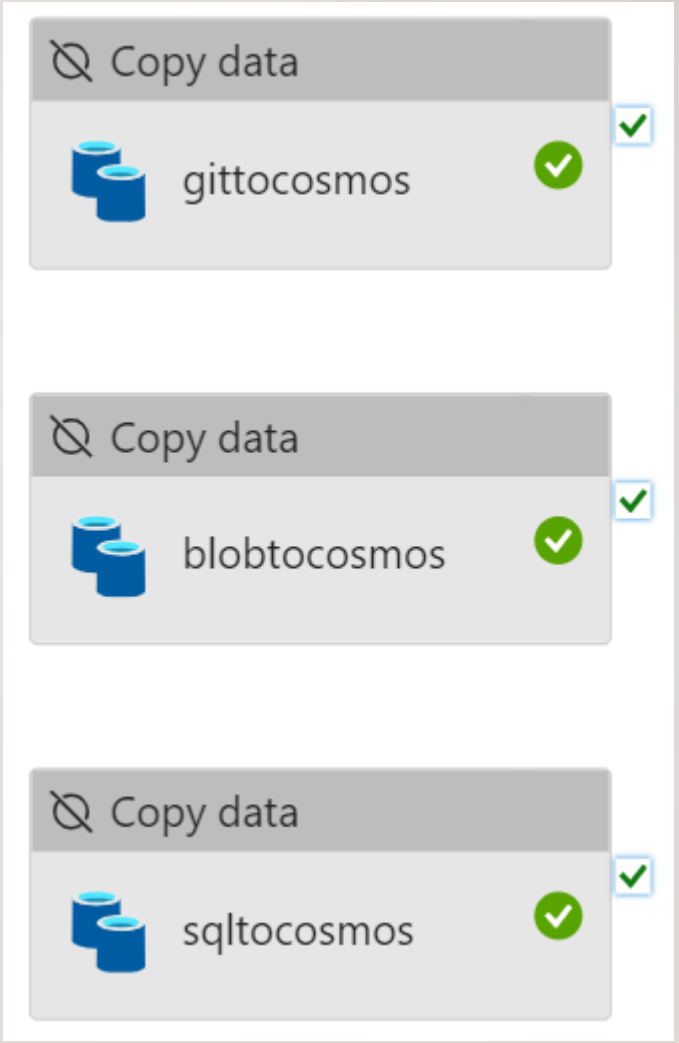
SOLUTION FLOW

Pipeline



SOLUTION FLOW

Pipeline – Data Ingestion



Azure Cosmos DB

Globally distributed

Multi Model

High Throughput

WHY NOSQL COSMOS DB FOR THE LANDING ZONE ?

- Global Distribution and Scalability
- Low Latency with High Availability
- Cosmos DB is schema-less, meaning you can have different fields in different records without any schema enforcement.

HTTP

Succeeded

Azure IR region: Central US ⓘ

Azure Cosmos DB for NoSQL

Data read: ⓘ	23.062 MB	**Data written:** ⓘ	127.292 MB
Files read: ⓘ	1	**Rows written:** ⓘ	1,86,168
Rows read:	1,86,168	**Peak connections:** ⓘ	4
Peak connections: ⓘ	2		

Copy duration	05:02:13
Throughput: ⓘ	1.272 KB/s

HTTP → Azure Cosmos DB for NoSQL

Start time	9/28/2024, 12:33:53 PM
Used DIUs ⓘ	4
Used parallel copies ⓘ	4
Duration	05:02:13

Azure Blob Storage
Region: Central US

Succeeded

Azure IR region: Central US ⓘ

Azure Cosmos DB for NoSQL

Data read: ⓘ	31.282 MB	**Data written:** ⓘ	172.604 MB
Files read: ⓘ	1	**Rows written:** ⓘ	2,52,389
Rows read:	2,52,389	**Peak connections:** ⓘ	4
Peak connections: ⓘ	8		

Copy duration	03:57:42
Throughput: ⓘ	2.194 KB/s

Azure Blob Storage → Azure Cosmos DB for NoSQL

Start time	9/28/2024, 5:36:10 PM
Used DIUs ⓘ	4
Used parallel copies ⓘ	4
Duration	03:57:42

Azure SQL Database
Region: Central US

Succeeded

Azure IR region: Central US

Azure Cosmos DB for NoSQL

Data read:	22.346 MB	Data written:	68.483 MB
Rows read:	10,48,575	Rows written:	10,48,575
Peak connections:	1	Peak connections:	4

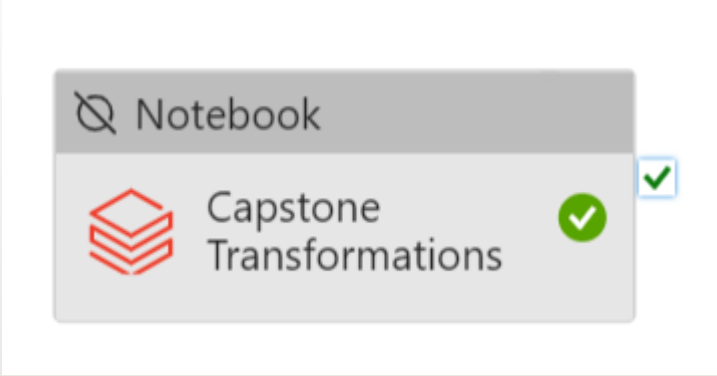
Copy duration	08:10:23
Throughput:	760 bytes/s

Azure SQL Database → Azure Cosmos DB for NoSQL

Start time	9/28/2024, 12:33:53 PM
Used DIUs	4
Used parallel copies	4
Duration	08:10:23

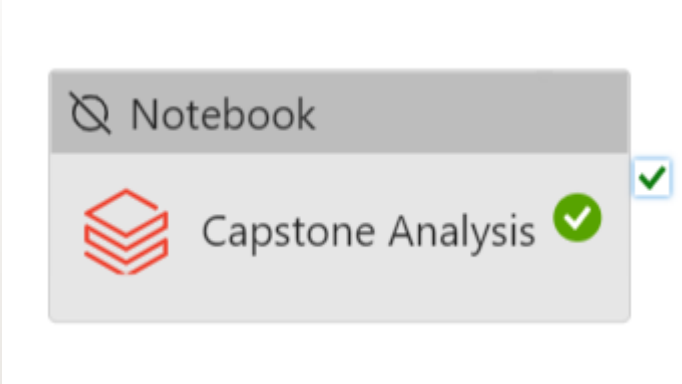
SOLUTION FLOW

Pipeline – Data Cleaning, Transformation & Aggregation

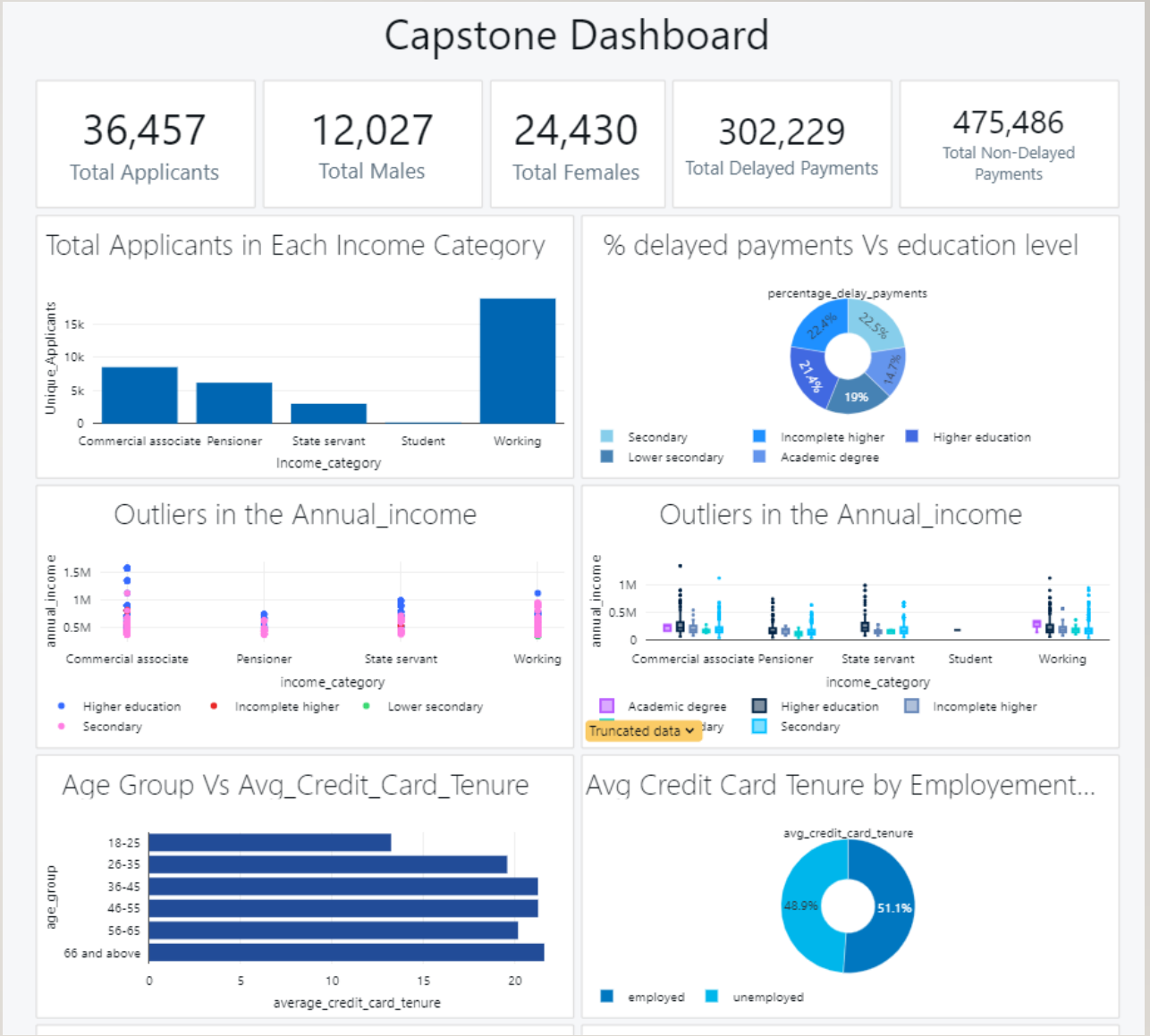


SOLUTION FLOW

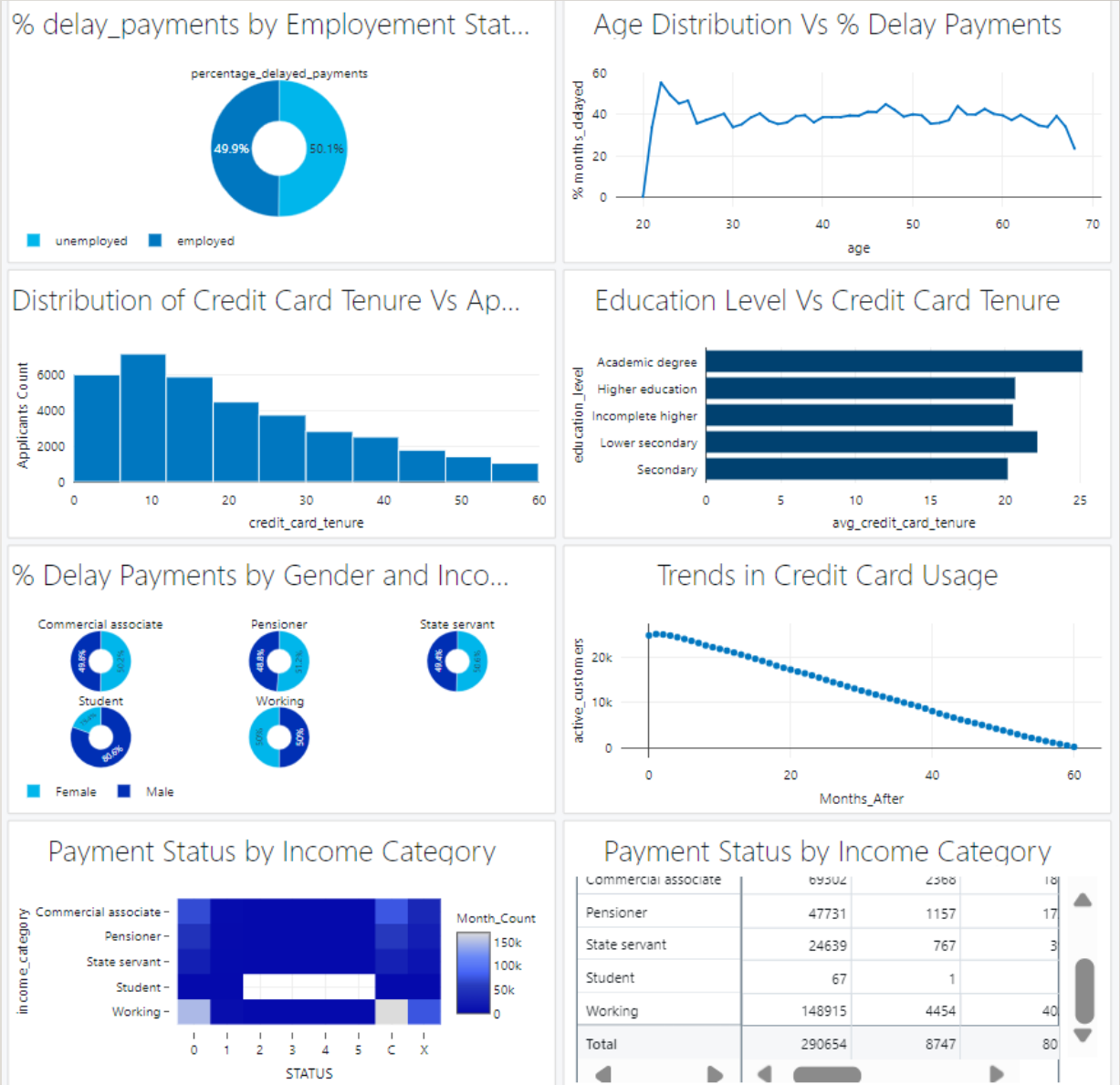
Pipeline – Data Visualization



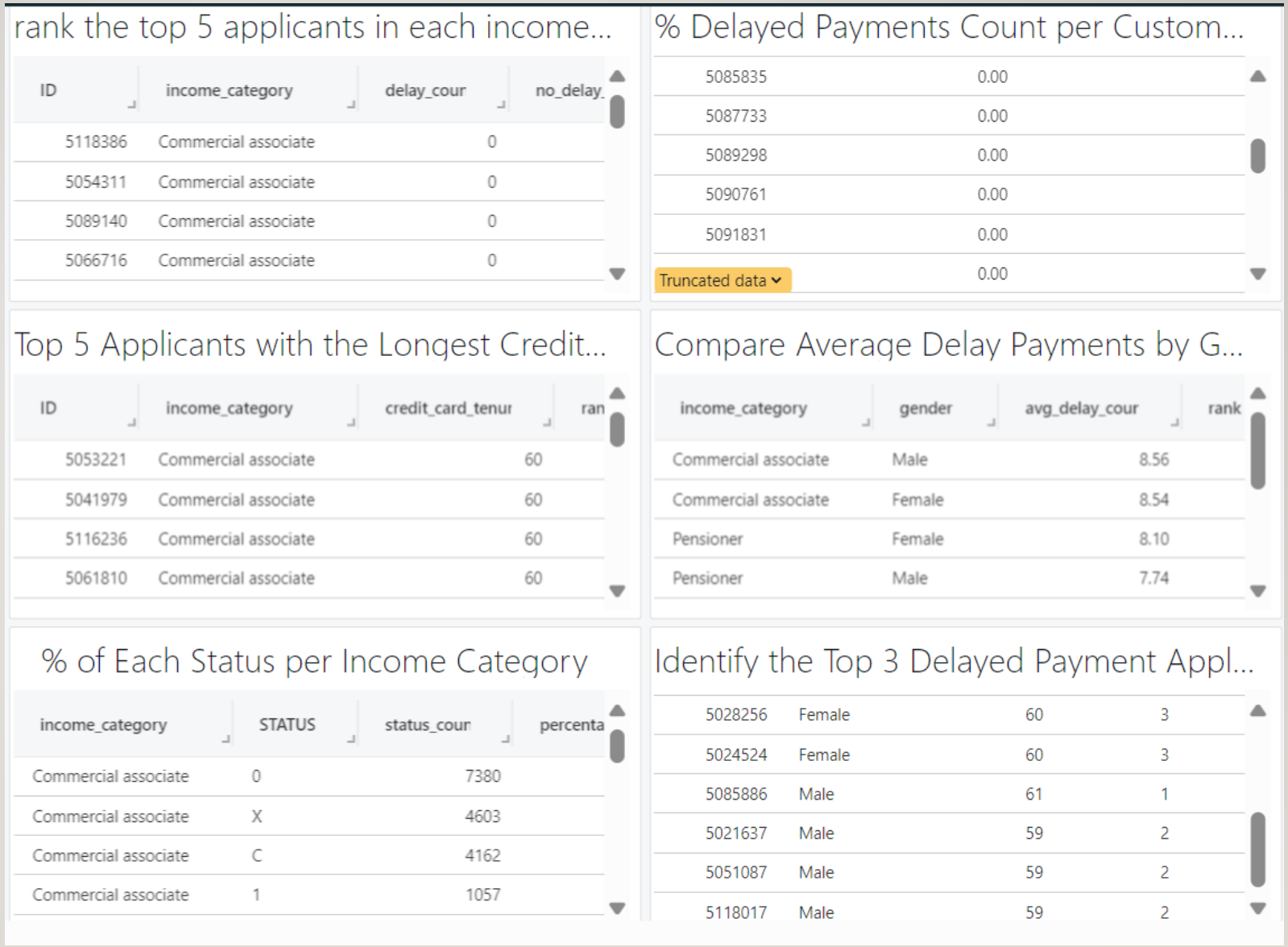
OUTPUT DASHBOARD



OUTPUT DASHBOARD



OUTPUT DASHBOARD



rank the top 5 applicants in each income...

ID	income_category	delay_cour	no_delay
5118386	Commercial associate	0	
5054311	Commercial associate	0	
5089140	Commercial associate	0	
5066716	Commercial associate	0	

% Delayed Payments Count per Custom...

5085835	0.00
5087733	0.00
5089298	0.00
5090761	0.00
5091831	0.00
Truncated data	0.00

Top 5 Applicants with the Longest Credit...

ID	income_category	credit_card_tenur	ran
5053221	Commercial associate	60	
5041979	Commercial associate	60	
5116236	Commercial associate	60	
5061810	Commercial associate	60	

Compare Average Delay Payments by G...

income_category	gender	avg_delay_cour	rank
Commercial associate	Male	8.56	
Commercial associate	Female	8.54	
Pensioner	Female	8.10	
Pensioner	Male	7.74	

% of Each Status per Income Category

income_category	STATUS	status_cour	percenta
Commercial associate	0	7380	
Commercial associate	X	4603	
Commercial associate	C	4162	
Commercial associate	1	1057	

Identify the Top 3 Delayed Payment Appl...

5028256	Female	60	3
5024524	Female	60	3
5085886	Male	61	1
5021637	Male	59	2
5051087	Male	59	2
5118017	Male	59	2

RESULTS/INSIGHTS

From the data the following insights have been derived

- By the analysis we can find that education level correlates with payment behavior. It appears that higher education levels like graduate school are having much more non-defaulter compared to lower education levels.
- Customers with an **academic degree** have the longest average credit card tenure (25.1 months), suggesting that higher education correlates with long-term financial stability and responsible credit use.
- Younger adults, especially those in their **mid to late twenties**, show a higher rate of defaulting. This could reflect early-career financial instability. In contrast, older customers (above 40) display more financial maturity, with lower default rates.
- Both unemployed and employed individuals have similar delayed payment rates (~39%), indicating that credit management issues rises regardless of employment status. However, unemployed individuals tend to have shorter credit card tenures, reflecting financial instability.

RESULTS/INSIGHTS

FROM THE DATA THE FOLLOWING INSIGHTS HAVE BEEN DERIVED

- Trends in Credit Card Usage Seems to be Very bad as the number of active customers are gradually decreasing over the Months
- Age and tenure shows that older age groups tend to have longer tenures. This could imply that older customers are more loyal and stable in debt payment.

Challenges Faced

1. Dataset Authenticity: Determining the genuineness of the dataset raised concerns about the validity of the analysis.

2. Performance Issues: The pipeline experienced longer running times, especially during data movement, leading to delays in data availability.

3. Budget Constraints: A strict budget limit of 100 USD for cloud services necessitated careful resource management to maintain essential operations.

FUTURE SCOPE

1)Upgrading the Architecture to support real-time Processing

2)Integrating more data sources such as user feedback ,social media trends and real time monitoring tools

3)Automation of the Dashboard in Azure

4)The Transformed data could be sent to following teams for further improvent…They are:

i)Data Analysts: Uses the pipeline's output for reporting and insights.

ii)Data Scientists: For Building models that can be deployed for predictive insights, such as credit scoring, fraud detection, or customer segmentation.

iii)CI/CD :For handling the pipeline's deployment and updates.

iv)DevOps :For Managing the infrastructure, scaling, and monitoring of the pipeline environment.

THANK YOU