Bank Loans ELT Data Pipeline

Team members:
Ahmed Mohsen - Hanin Baher - Mahmoud Ashraf
Facilitator:
Rawan Ehab

Table of Contents

Table of Contents

01

Introduction

Introduction

- Big data is increasingly used for financial risk analysis.
- The goal: Analyze loan default risk using a full ELT pipeline.
- Pipeline spans from raw data ingestion to visualization.

- In the U.S., the average default rate for personal loans ranges from 2–6%, but this can spike above 10% during economic downturns.
- According to the World Bank, non-performing loans (NPLs) globally account for over $1.4 trillion, highlighting the critical need for predictive risk management.

02

Problem Statment

The Problem

Loan providers struggle with fragmented data, making it hard to compute risk metrics like DTI and LTV.

Manual analysis is slow, and without centralized dashboards, real-time monitoring, and data-driven decisions are limited.

03

Key Tasks

Key tasks

Cluster setup — Using Docker and Postgres for the source database.

Data Ingestion — Extract raw loan data from Postgres into HDFS using Sqoop.

Data Transformation — By using PySpark (Zeppelin) in:

- Cleaning Data
- Dimensional modeling (fact + dimension tables)

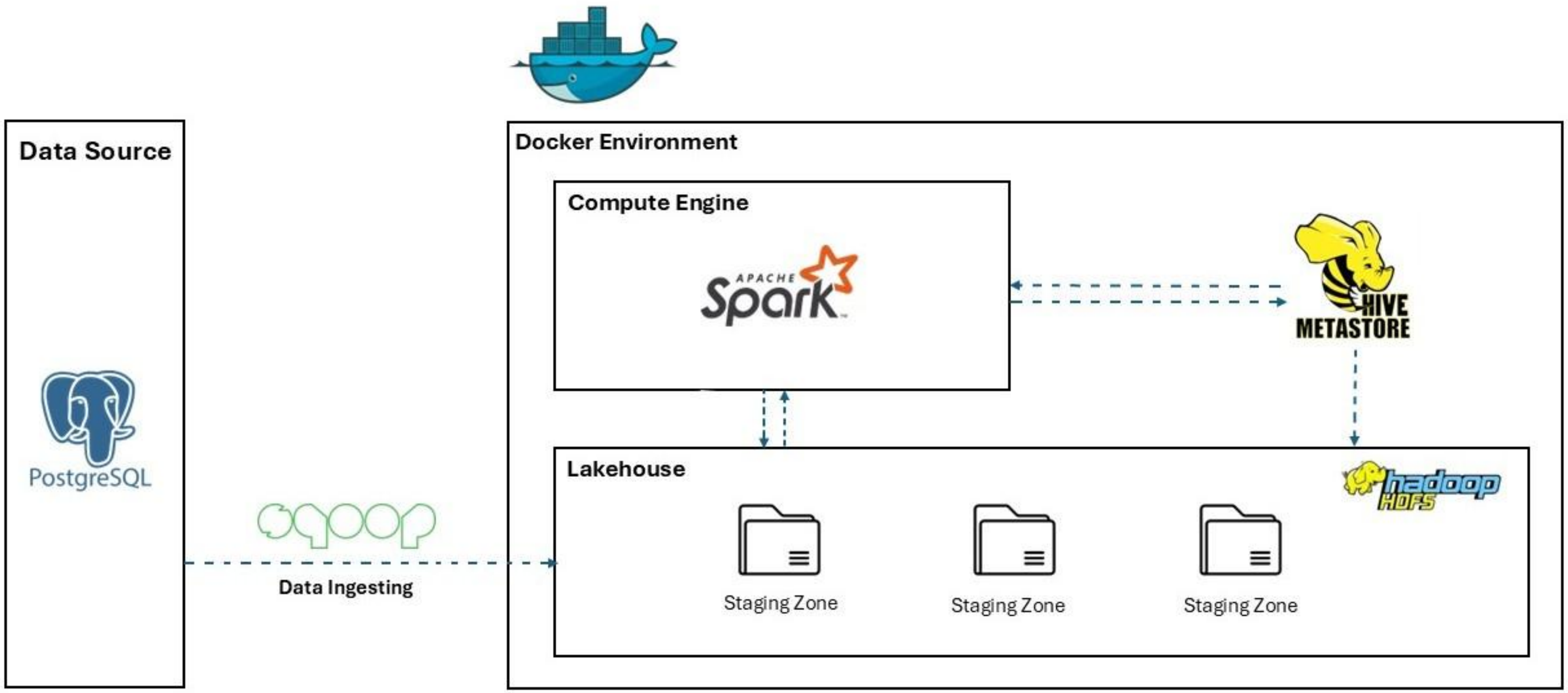
Data Warehouse Loading — Into Hive with Parquet storage.

Visualization & Analytics — Using Power BI dashboards.

04

Methodology

Methodology



05

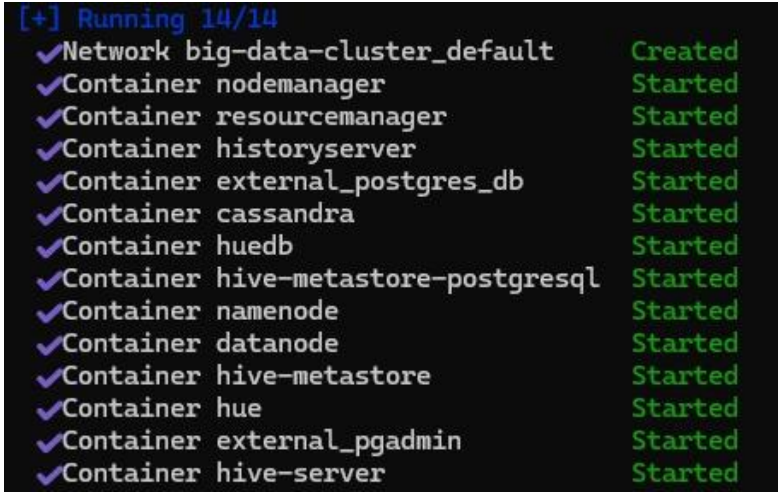
Our pipeline

Cluster Setup

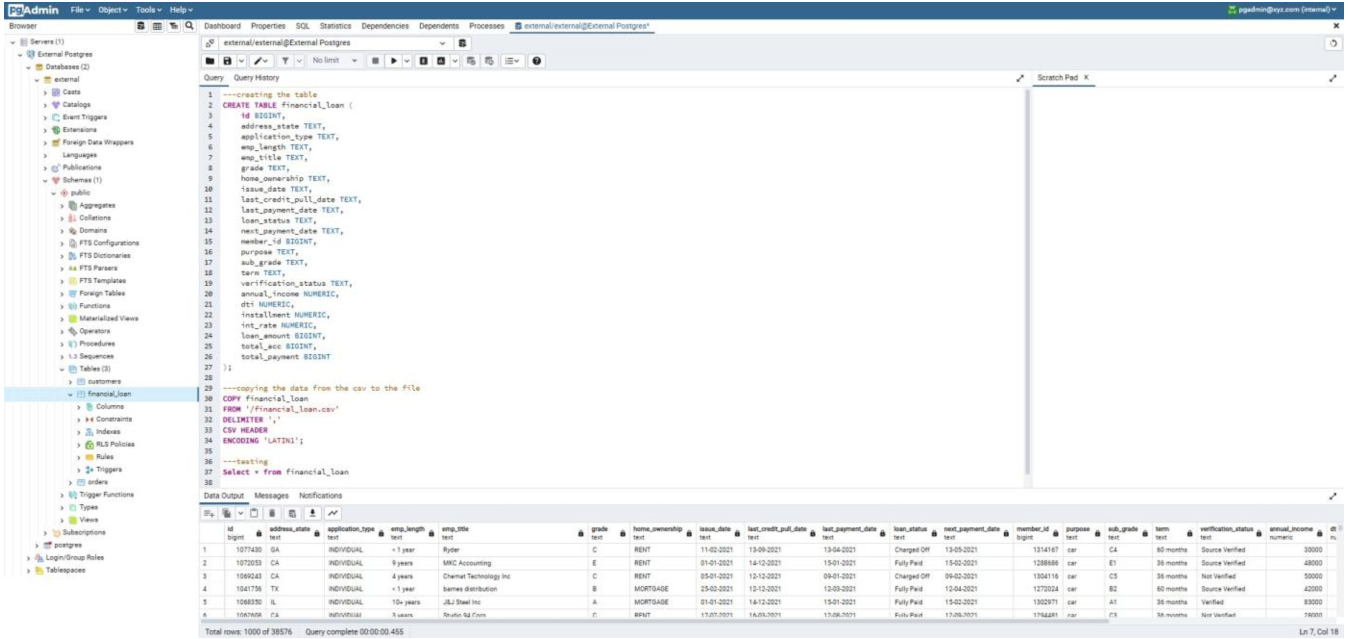
- Docker Compose used for multi-container cluster
- Services included: Postgres, Hive, Sqoop, Zeppelin, HDFS, Hue, Power BI connection

URL references:

- PgAdmin (http://localhost:5000)
- Zeppelin (http://localhost:8082)
- Hue (http://localhost:8888)
- Verified services via docker ps.



Data Creation (Postgres)



- CSV dataset (financial_loan.csv) uploaded into Postgres via pgAdmin.
- Created table financial_loan with 20+ attributes (borrower details, loan amount, payment dates, etc.).
- Used COPY command for efficient data import.

Verified with SQL queries (SELECT * LIMIT 10).

Data Ingesting (Sqoop)

Sqoop imports data from Postgres → HDFS as Parquet files

First:

Opening Sqoop inside the Hive container:

docker exec -it hive-server bash

Second:

Running Sqoop import command

Path: /staging_zone/financial_loan

```
sqoop import \
--connect jdbc:postgresql://external_postgres_db/postgres \
--username external \
--password external \
--table financial_loan \
--target-dir /staging_zone/financial_loan \
--as-parquetfile \
--m 1
```

Data Storage (HDFS)

Raw financial loan data is stored in HDFS.

Benefits:

- High availability & fault tolerance.
- Scalability for millions of records.
- Parallel access for Spark transformations.

```
root@90b03649b1d5:/# hdfs dfs -ls /staging_zone/financial_loan
Found 3 items
drwxr-xr-x   - root supergroup          0 2025-10-01 00:42 /staging_zone/financial_loan/.metadata
drwxr-xr-x   - root supergroup          0 2025-10-01 00:42 /staging_zone/financial_loan/.signals
-rw-r--r--   3 root supergroup    1698990 2025-10-01 00:42 /staging_zone/financial_loan/983a2c69-23b0-482e-81c8-035f5d8e1366.parquet
```

Data Transformations & Modeling (Spark)

Data Cleaning

Null handling in Emp_title cloumn, emp_length and term columns formatting, date type conversions.

Dimensional Modeling

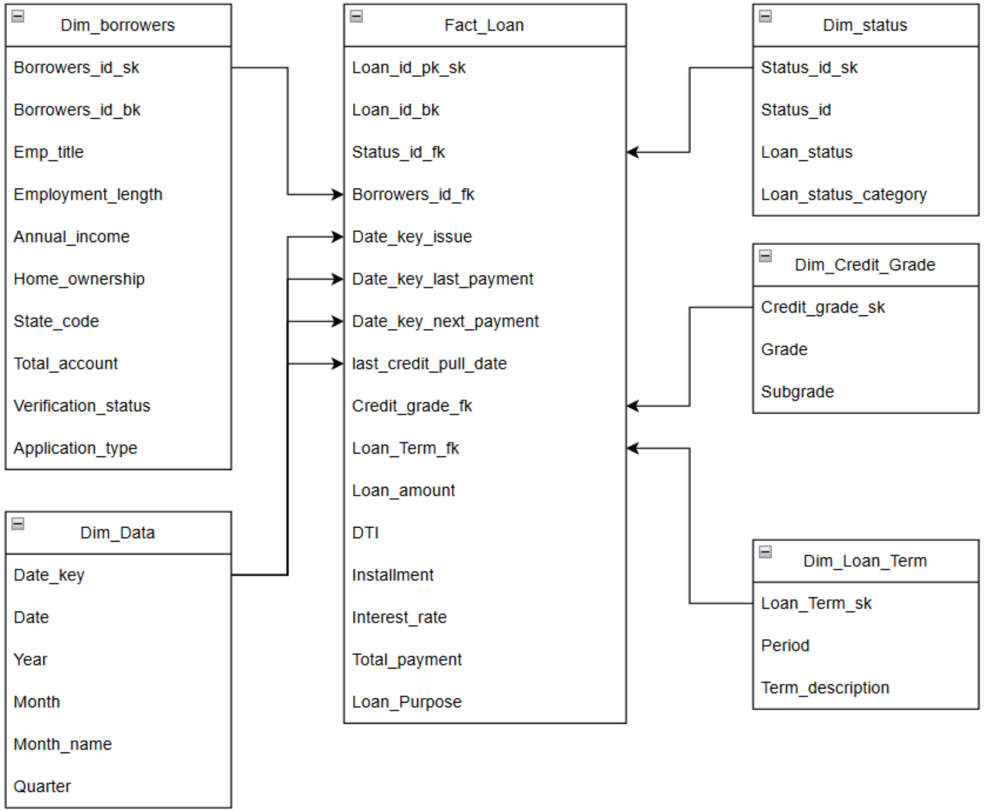
Fact Table:

- fact_loan (central metrics, risks, loan status).

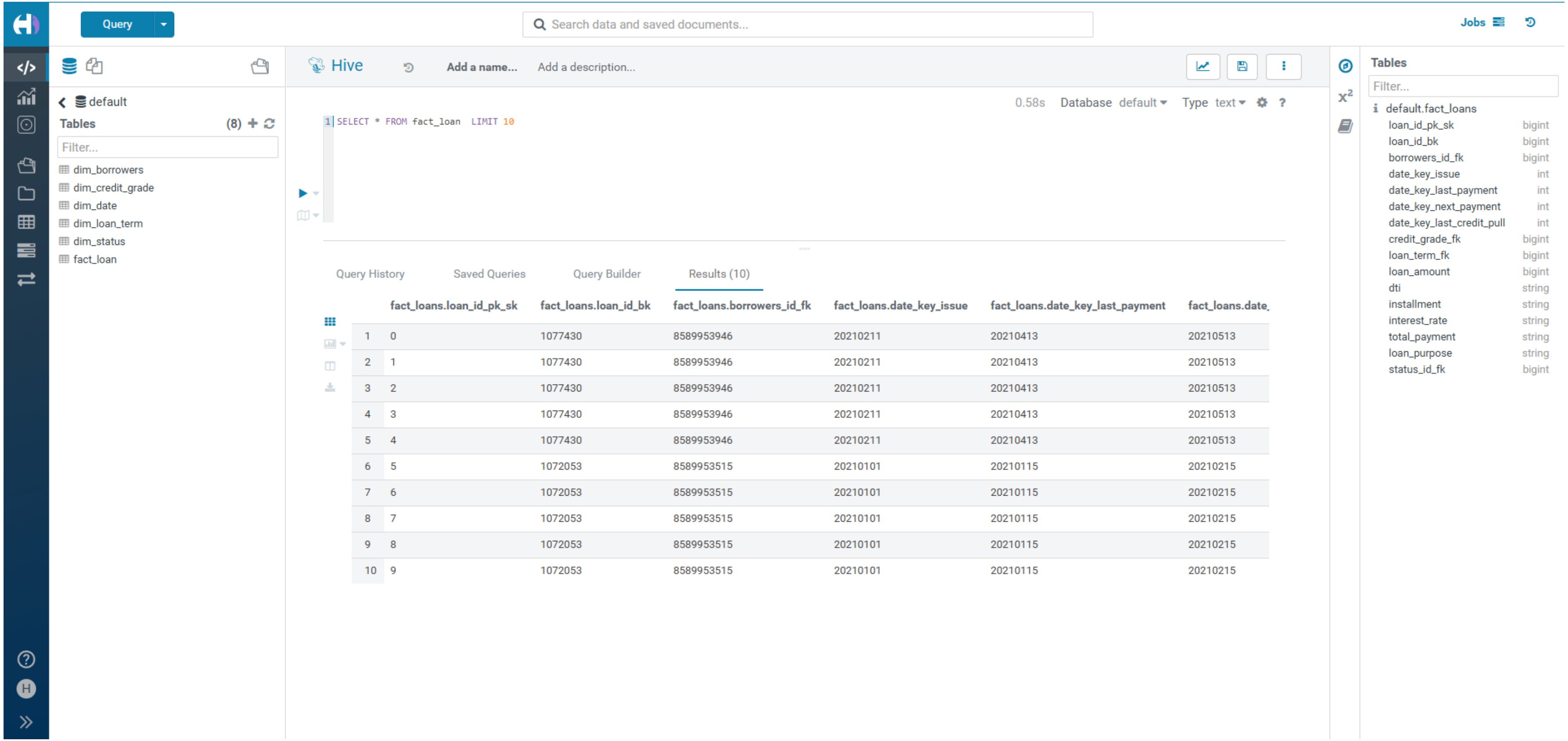
Dimensions:

- dim_borrowers (borrower details) - dim_status (loan status categories)
- dim_credit_grade (risk grades) - dim_loan_term (term duration)
- dim_date (date dimension).

Data Transformations & Modeling (Spark)



Data Warehouse (Hive)



06

Results & Insights

Power BI Connection

Connected Power BI to Hive via ODBC.

- Host: localhost
- Port: 10000
- Authentication: Username & Password
- Database: your Hive DB
- Test connection → Save

Power BI Dashboards

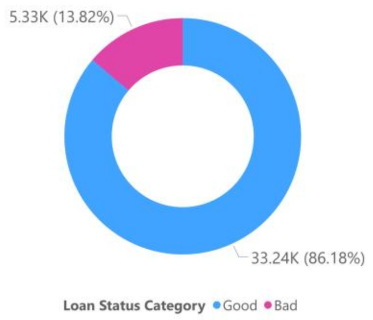
Summary

1/1/2021 – 12/31/2021

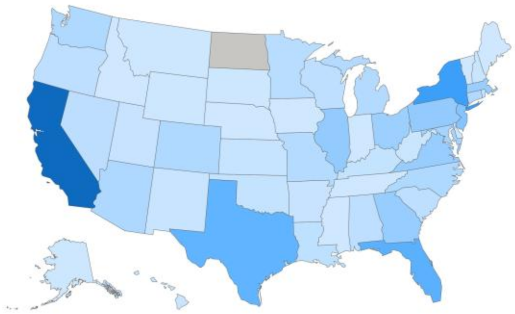
Loan Purpose: All

Total Loans	Total Funded Amount	Total Amount Received	Avg Interest Rate	Avg DTI
38.58K	436M	473M	12.0%	13.3%

Good vs Bad loans

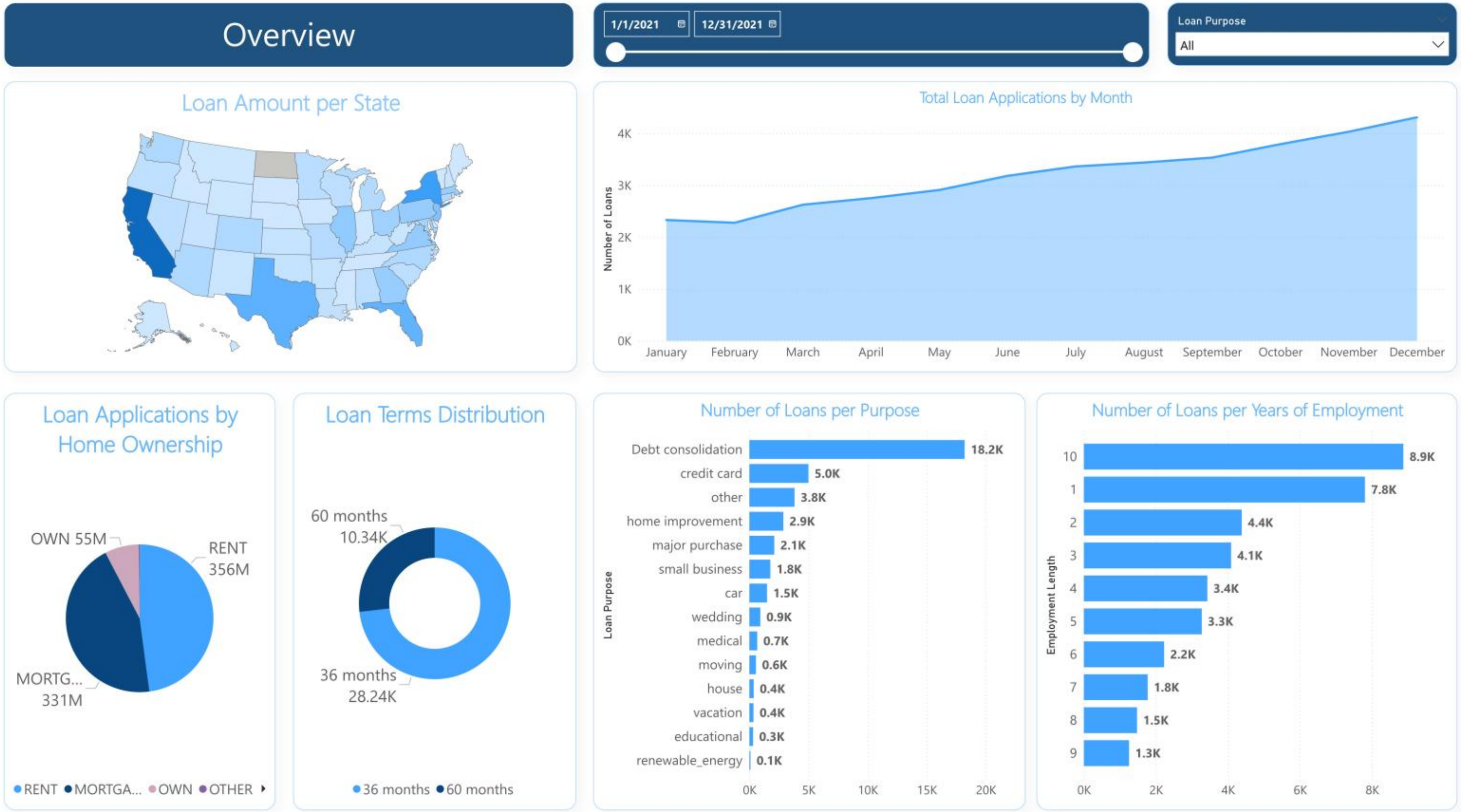


Number of Loans per State



Loan Status	Total Loans	Total Amount Received	Total Funded Amount	AVG Interest Rate	AVG DTI
Current	1098	24M	19M	15.1%	14.7%
Charged Off	5333	37M	66M	13.9%	14.0%
Fully Paid	32145	412M	351M	11.6%	13.2%
Total	**38576**	**473M**	**436M**	**12.0%**	**13.3%**

Power BI Dashboards



Power BI Dashboards

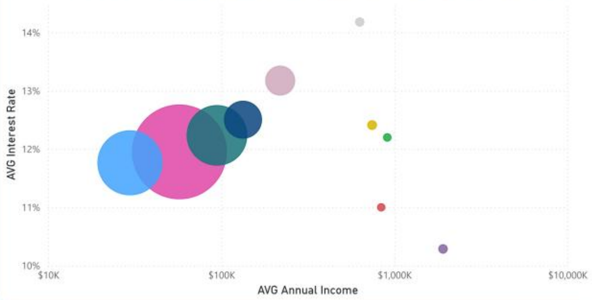
Details

1/1/2021 12/31/2021

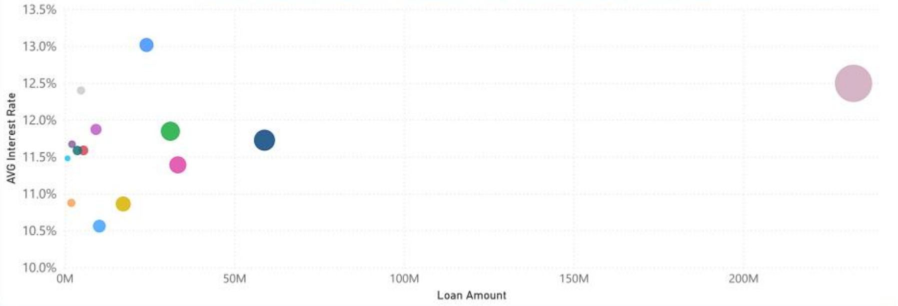
Loan Purpose

All

Relationship Between AVG Annual Income & AVG Interest Rate per Income Group



Relationship Between Loan Amount & AVG Interest Rate per Loan Purpose



Loan_ID	Issue Date	Loan Amount	Loan Purpose	Interest Rate	Installment	Paid So Far	Home Ownership	AVG DTI	Loan Term
59006	9/9/2021	3000	credit card	14.0%	102.92	3705	MORTGAGE	15.0%	36
61390	2/10/2021	4000	credit card	8.0%	125.13	4452	MORTGAGE	17.0%	36
65426	8/9/2021	4000	car	11.0%	131.22	2755	MORTGAGE	11.0%	36
65640	5/8/2021	5000	home improvement	11.0%	87.19	3154	MORTGAGE	17.0%	36
66749	12/8/2021	10625	Debt consolidation	13.0%	360.43	12975	MORTGAGE	22.0%	36
66964	6/8/2021	7500	Debt consolidation	13.0%	253.58	9129	MORTGAGE	9.0%	36
67503	10/9/2021	10000	Debt consolidation	9.0%	316.11	11280	MORTGAGE	15.0%	36
68163	2/10/2021	3000	small business	7.0%	92.82	3342	MORTGAGE	7.0%	36
68817	3/8/2021	10000	major purchase	11.0%	327.53	11709	MORTGAGE	13.0%	36

07

Conclusion

Conclusion

- End-to-end Big Data ELT pipeline successfully built.
- Automated data ingestion, transformation, and warehousing.
- Power BI dashboards turned raw data into actionable insights.
- Framework can be extended to predictive loan default models.

08

Future Work

Future Work

ML	Integrate ML models for loan default prediction.
Orchestration	Automate pipeline orchestration with Apache Airflow.
Multiple Sources	Add external data sources (credit bureau reports, customer profiles).
Streaming	Enable real-time streaming ingestion (Kafka + Spark Streaming).

Thanks

Do you have any questions?