Project Report on ETL Pipeline (ADF)

Project Overview

The project aims to create an efficient data engineering pipeline to process loan data from GitHub, transform it for analytics, and visualize insights using Power BI. The pipeline involves Azure Data Factory (ADF) for ETL (Extract, Transform, Load) operations, Azure Synapse Analytics for staging and transformation, and Power BI for reporting and visualization.

Tools and Technologies

- **Data Source:** GitHub (Loan Dataset)
- **ETL Platform:** Azure Data Factory (ADF)
- **Data Storage:** Azure Blob Storage and Azure Synapse Analytics
- **Visualization Tool:** Power BI
- **Data Transformation:** Data Flow activity in ADF
- **Validation:** Script activity in ADF using SQL scripts

Key Layers

1. **Raw Data Layer:** Raw data extracted from GitHub and stored in Azure Blob Storage.
2. **Staging Layer:** Data loaded into Azure Synapse Analytics for validation and temporary storage.
3. **Processed Data Layer:** Transformed and validated data stored in Synapse's final table for analytics.
4. **Presentation Layer:** Data fetched into Power BI for generating dashboards and reports.

Data and File Formats

- **Source Format:** CSV files from GitHub
- **Intermediate Format:** Parquet (for transformation in ADF)
- **Final Table Schema:** Synapse table LoanProcessData with detailed loan attributes.

Storage Location

- **Raw Data:** Azure Blob Storage
- **Transformed Data:** Azure Synapse Analytics (staging and processed tables)
- **Visualization Data:** Synapse Analytics integrated with Power BI

Dataset Understanding:

1. Unique Identifiers

- **id**: Critical.

  Reason: Identifies each loan application. If missing, the row loses uniqueness, making it impossible to track specific loans.

  Action: Check for null or duplicate values.

- **member_id**: Critical.

  Reason: Identifies the borrower. Without this, customer-level analysis cannot be performed.

  Action: Check for null values.

---

2. Loan Attributes

- **loan_amount**: Critical.

  Reason: Core financial figure. Null values affect financial metrics such as total disbursed loans.

  Action: Check for null or zero values.

- **int_rate (Interest Rate)**: Important.

  Reason: Used in calculating the cost of borrowing. Null values would mislead analysis on loan profitability.

- **installment**: Important.

  Reason: Monthly repayment amount. Null values impact cash flow predictions.

- **term**: Important.

  Reason: Represents loan duration (e.g., 36 months). Null values would prevent understanding loan periods.

---

3. Borrower Information

- **application_type**: Important.

  Reason: Indicates individual vs. joint application. Helps in customer segmentation.

- **emp_length (Employment Length)**: Important.

  Reason: Used to evaluate borrower stability. Null or missing values reduce reliability of risk assessments.

- **emp_title (Employment Title)**: Less important (depends on analysis goal).

  Reason: Provides job information. Not critical for financial calculations but useful in customer profiling.

- **annual_income**: Critical.

  Reason: Key determinant for loan approval and debt-to-income (DTI) ratio calculations.

- **dti (Debt-to-Income)**: Important.

  Reason: Measures borrower's debt relative to income. Null values affect risk evaluation.

---

4. Loan Status and Dates

- **loan_status**: Critical.

  Reason: Determines whether the loan is current, defaulted, or paid off. Null values prevent tracking performance.

- **issue_date**: Critical.

  Reason: Date loan was issued. Needed for time-based analysis.

- **last_payment_date**: Important.

  Reason: Helps analyze payment history. Missing values are acceptable if the loan is new.

- **next_payment_date**: Important.

  Reason: Relevant for forecasting future payments. Null values are acceptable if the loan is paid off.

- **last_credit_pull_date**: Less critical.

  Reason: Date of last credit check.

5. Credit Information

- **total_acc (Total Accounts)**: Important.

  Reason: Number of credit accounts affects creditworthiness.

- **total_payment**: Critical.

  Reason: Represents the total payments made. Affects revenue and performance calculations.

6. Risk Assessment

- **grade** and **sub_grade**: Important.

  Reason: Reflect credit risk levels.

- **Good vs Bad Loan**: Critical.

  Reason: Directly indicates loan performance. Null values undermine model training for risk prediction.

7. Other Attributes

- **home_ownership**: Less critical.

  Reason: Provides collateral information.

- **verification_status**: Important.

  Reason: Indicates if income was verified.

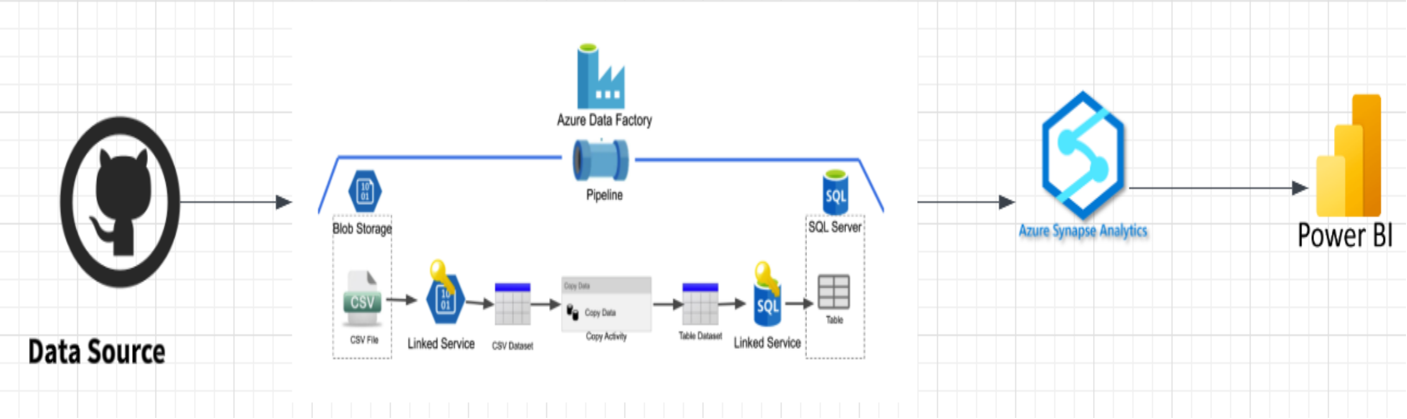
- **purpose**: Important.

  Reason: Describes the loan purpose. Useful for segmentation.

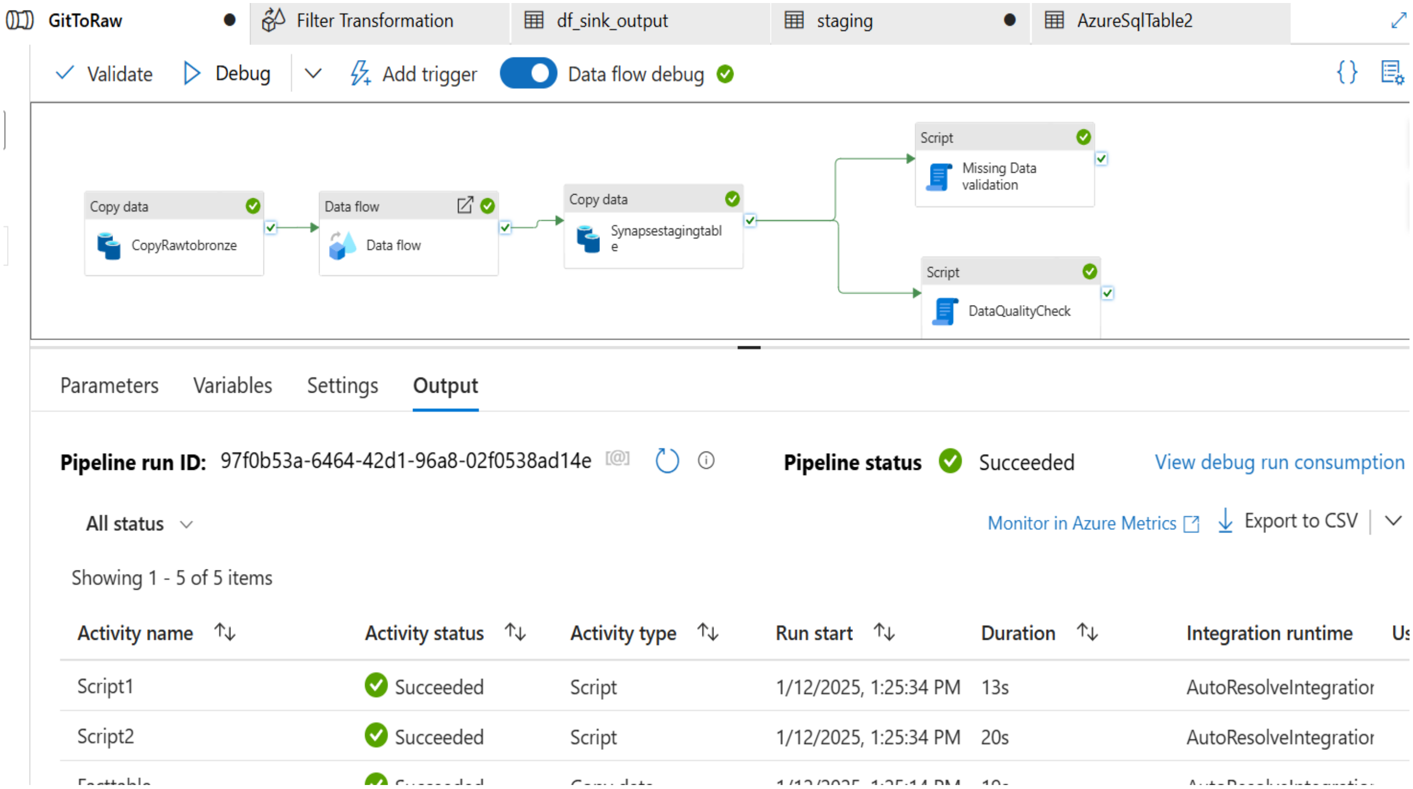
- **address_state**: Less critical.

  Reason: Geographic information for loan distribution.

Architectural Diagram:



Data Engineering Pipeline Design:



Pipeline run ID: 97f0b53a-6464-42d1-96a8-02f0538ad14e **Pipeline status** ✓ Succeeded

View debug run consumption

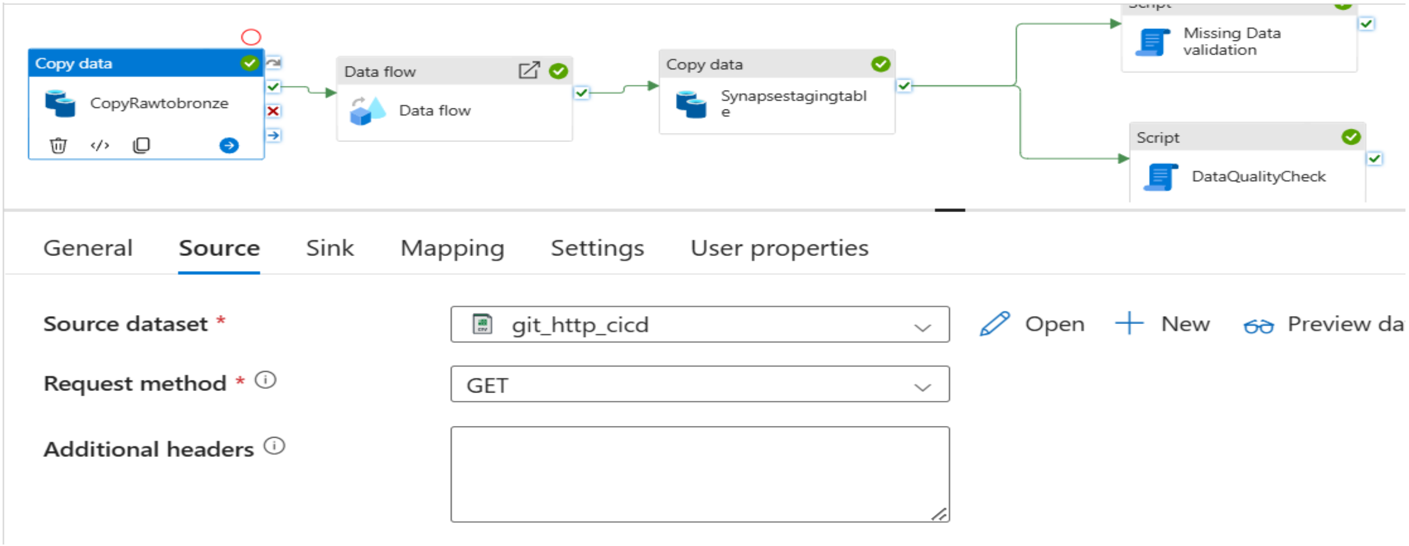
All status

Monitor in Azure Metrics Export to CSV

Showing 1 - 5 of 5 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	Us
Script1	✓ Succeeded	Script	1/12/2025, 1:25:34 PM	13s	AutoResolveIntegratior	
Script2	✓ Succeeded	Script	1/12/2025, 1:25:34 PM	20s	AutoResolveIntegratior	
Facttable	✓ Succeeded	Copy data	1/12/2025, 1:25:14 PM	10s	AutoResolveIntegration	

1. **Copy Data Activity:**
   - **Objective:** Extract raw data from GitHub and load it into Azure Blob Storage.
   - **Configuration:**
     - Source: GitHub dataset.
     - Sink: Azure Blob Storage (Raw Data Container).

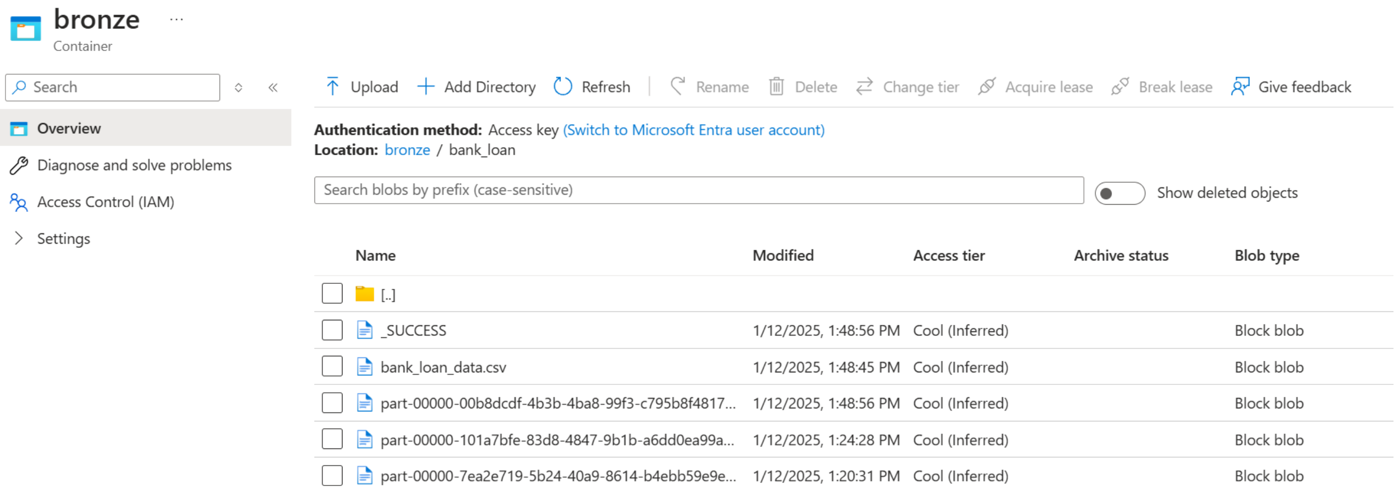


Source dataset * git_http_cicd

Request method * GET

Additional headers

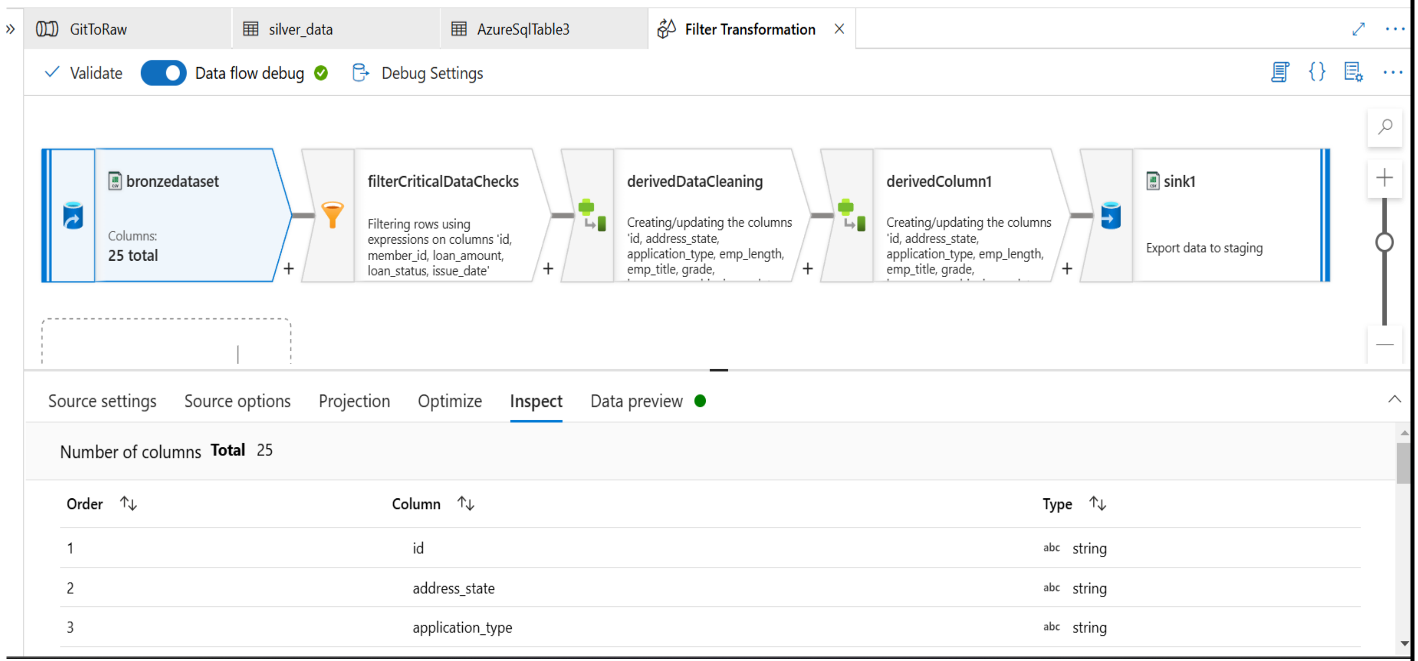
Storage of Data to bronze:



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: bronze / bank_loan

Name	Modified	Access tier	Archive status	Blob type
[..]				
_SUCCESS	1/12/2025, 1:48:56 PM	Cool (Inferred)		Block blob
bank_loan_data.csv	1/12/2025, 1:48:45 PM	Cool (Inferred)		Block blob
part-00000-00b8dcdf-4b3b-4ba8-99f3-c795b8f4817...	1/12/2025, 1:48:56 PM	Cool (Inferred)		Block blob
part-00000-101a7bfe-83d8-4847-9b1b-a6dd0ea99a...	1/12/2025, 1:24:28 PM	Cool (Inferred)		Block blob
part-00000-7ea2e719-5b24-40a9-8614-b4ebb59e9e...	1/12/2025, 1:20:31 PM	Cool (Inferred)		Block blob

2. Data Flow Activity:


GitToRaw
silver_data
AzureSqlTable3
Filter Transformation
Validate
Data flow debug
Debug Settings
bronzedataset
Columns:
25 total
filterCriticalDataChecks
Filtering rows using expressions on columns 'id, member_id, loan_amount, loan_status, issue_date'
derivedDataCleaning
Creating/updating the columns 'id, address_state, application_type, emp_length, emp_title, grade,
derivedColumn1
Creating/updating the columns 'id, address_state, application_type, emp_length, emp_title, grade,
sink1
Export data to staging
Source settings
Source options
Projection
Optimize
Inspect
Data preview

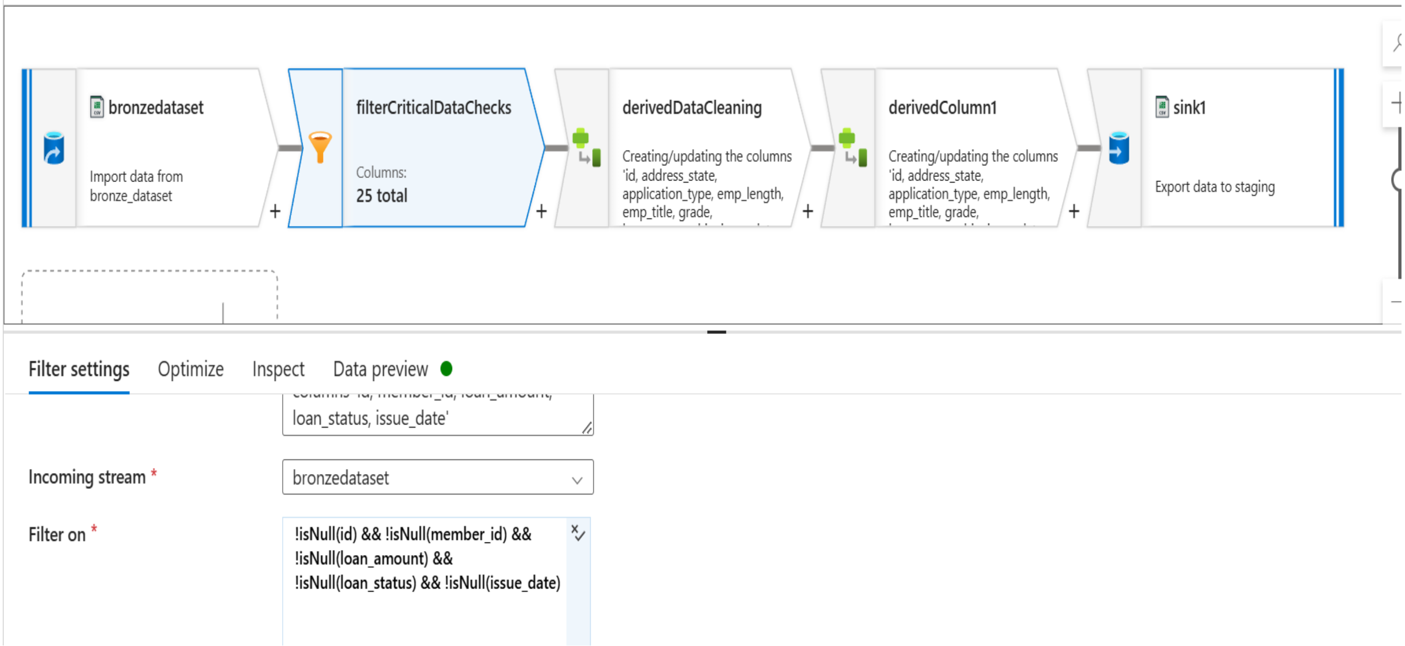
Number of columns **Total** 25

Order	Column	Type
1	id	abc string
2	address_state	abc string
3	application_type	abc string

- **Objective:** Transform data and load it into the Synapse final table.
- **Transformation Operations:**
  - Data cleaning (null handling and type conversion).

3. Summary of Critical Columns to Check for Null/Empty:

Column Name	Why It’s Critical
id	**Unique identifier for loans**
member_id	**Unique identifier for borrowers**
loan_amount	**Core financial metric**
annual_income	**Determines borrower’s financial capability**
loan_status	**Tracks loan performance**
issue_date	**Needed for time-based analysis**
Good vs Bad Loan	**Indicates loan risk and performance**
total_payment	**Represents total payments made**


bronzedataset
Import data from bronze_dataset
filterCriticalDataChecks
Columns: 25 total
derivedDataCleaning
Creating/updating the columns 'id, address_state, application_type, emp_length, emp_title, grade,
derivedColumn1
Creating/updating the columns 'id, address_state, application_type, emp_length, emp_title, grade,
sink1
Export data to staging
Filter settings
Optimize
Inspect
Data preview

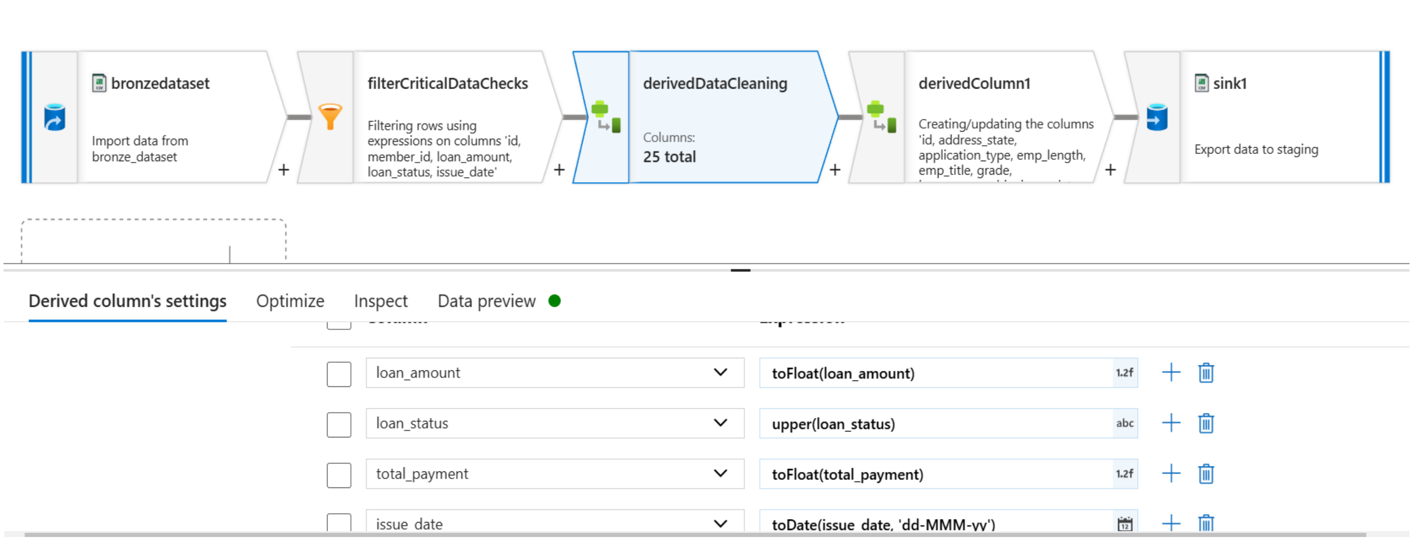
columns 'id, member_id, loan_amount, loan_status, issue_date'

Incoming stream * bronzedataset

Filter on * !isNull(id) && !isNull(member_id) && !isNull(loan_amount) && !isNull(loan_status) && !isNull(issue_date)

Columns Important for Data Quality:

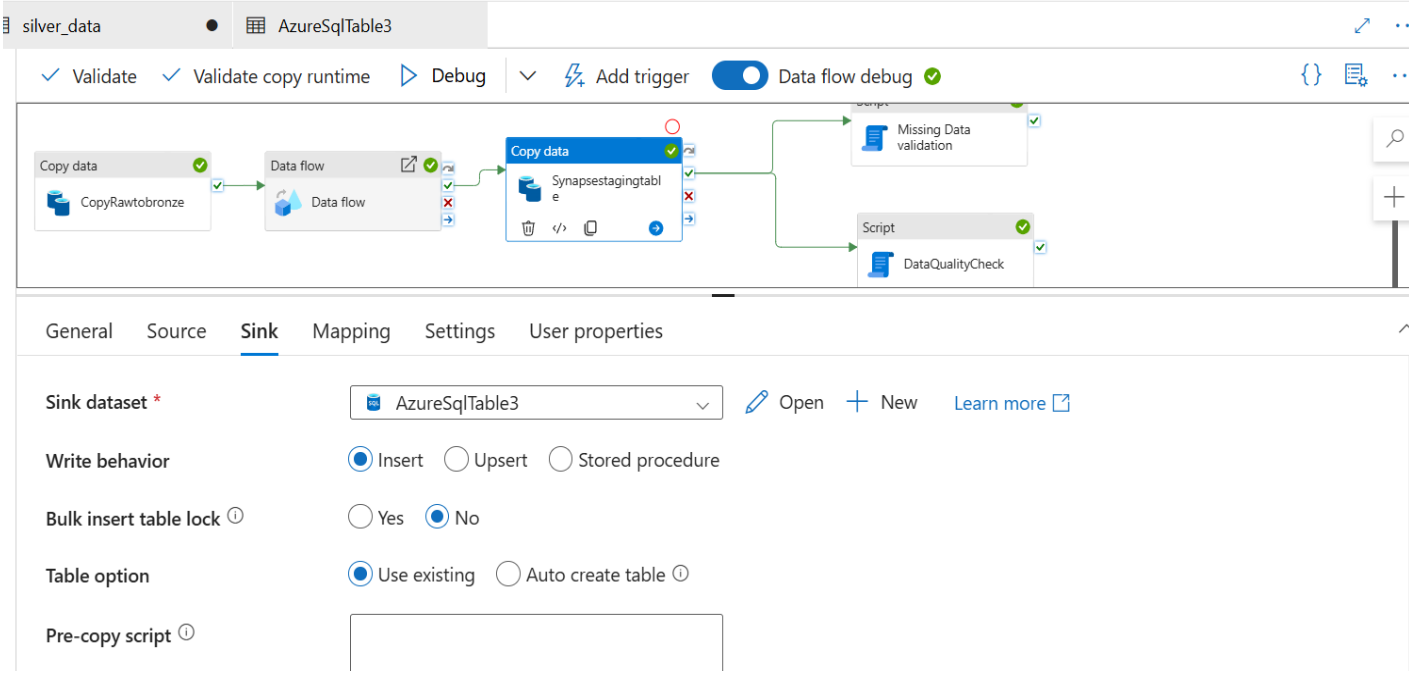
Column Name	Why It’s Important
emp_length	Employment stability
dti	Debt-to-income ratio
int_rate	Loan cost
term	Loan duration
grade, sub_grade	Credit risk assessment


bronzedataset
Import data from bronze_dataset
filterCriticalDataChecks
Filtering rows using expressions on columns 'id, member_id, loan_amount, loan_status, issue_date'
derivedDataCleaning
Columns:
25 total
derivedColumn1
Creating/updating the columns 'id, address_state, application_type, emp_length, emp_title, grade,
sink1
Export data to staging
Derived column's settings
Optimize
Inspect
Data preview

	Column	Expression
□	loan_amount	toFloat(loan_amount)
□	loan_status	upper(loan_status)
□	total_payment	toFloat(total_payment)
□	issue_date	toDate(issue_date, 'dd-MMM-yy')

- **Output:** Transformed data written to LoanProcessData table.

4. **Copy Data Activity:**
    - **Objective:** Load clean data from Blob Storage to Synapse staging table.
    - **Configuration:**
        - Source: Azure Blob Storage.
        - Sink: Synapse staging table.



Sink dataset * AzureSqlTable3 Open + New Learn more

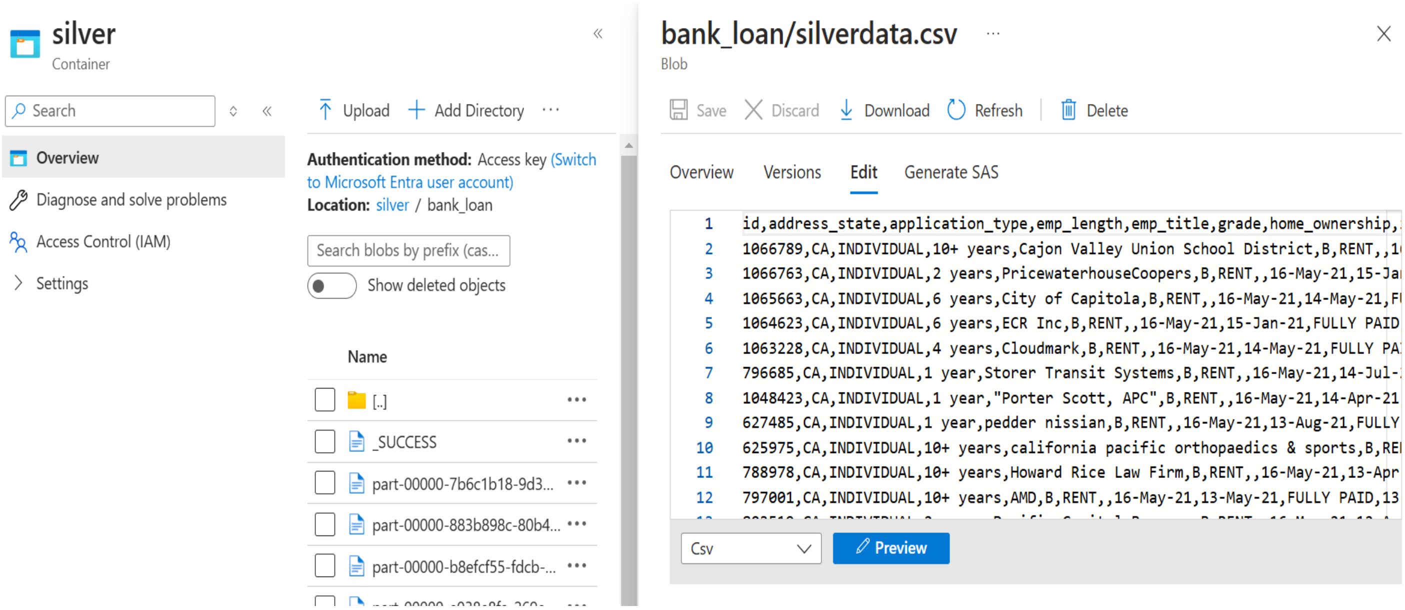
Write behavior ● Insert ○ Upsert ○ Stored procedure

Bulk insert table lock ○ Yes ● No

Table option ● Use existing ○ Auto create table

Pre-copy script

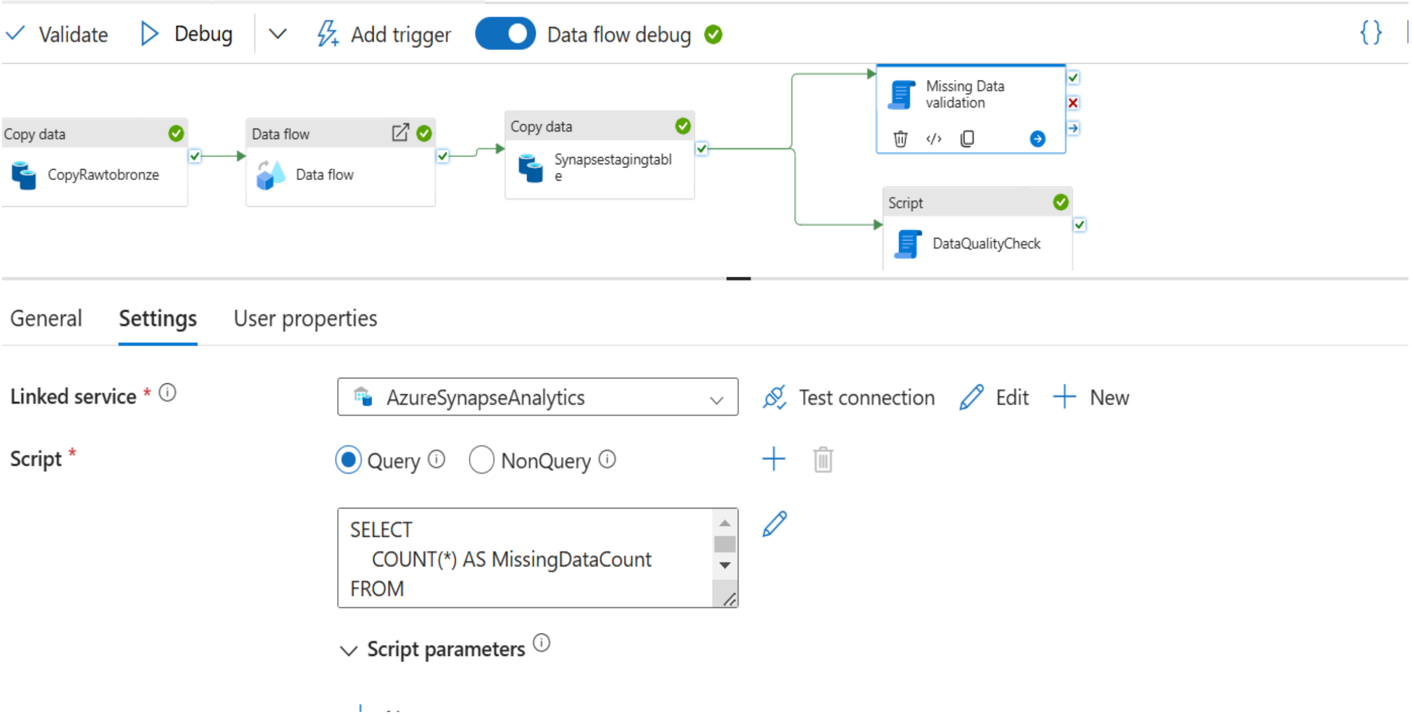
Storage of clean silver Data:



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: silver / bank_loan

```
id,address_state,application_type,emp_length,emp_title,grade,home_ownership,
1066789,CA,INDIVIDUAL,10+ years,Cajon Valley Union School District,B,RENT,,1
1066763,CA,INDIVIDUAL,2 years,PricewaterhouseCoopers,B,RENT,,16-May-21,15-Ja
1065663,CA,INDIVIDUAL,6 years,City of Capitola,B,RENT,,16-May-21,14-May-21,F
1064623,CA,INDIVIDUAL,6 years,ECR Inc,B,RENT,,16-May-21,15-Jan-21,FULLY PAID
1063228,CA,INDIVIDUAL,4 years,Cloudmark,B,RENT,,16-May-21,14-May-21,FULLY PA
796685,CA,INDIVIDUAL,1 year,Storer Transit Systems,B,RENT,,16-May-21,14-Jul-
1048423,CA,INDIVIDUAL,1 year,"Porter Scott, APC",B,RENT,,16-May-21,14-Apr-21
627485,CA,INDIVIDUAL,1 year,pedder nissian,B,RENT,,16-May-21,13-Aug-21,FULLY
625975,CA,INDIVIDUAL,10+ years,california pacific orthopaedics & sports,B,RE
788978,CA,INDIVIDUAL,10+ years,Howard Rice Law Firm,B,RENT,,16-May-21,13-Apr
797001,CA,INDIVIDUAL,10+ years,AMD,B,RENT,,16-May-21,13-May-21,FULLY PAID,13
```

5. **Script Activity:**
   - **Objective:** Validate transformed data.
   - **Validation Checks:**
     - Duplicate records.
     - Null values in critical columns.

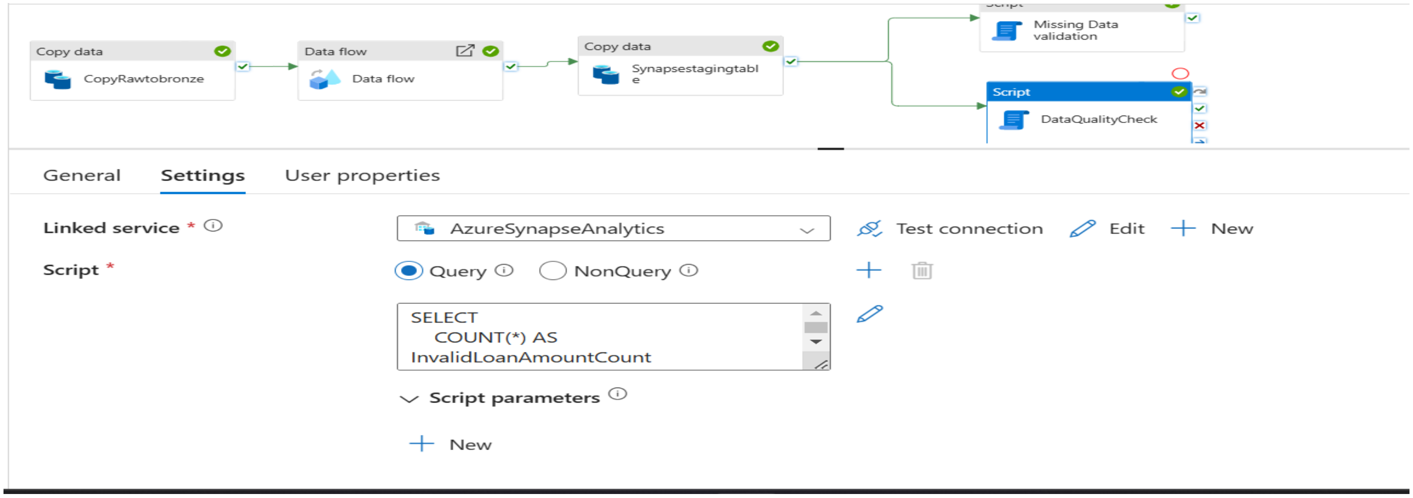


Linked service * AzureSynapseAnalytics

Script * ● Query ○ NonQuery

```
SELECT
    COUNT(*) AS MissingDataCount
FROM
```

Synapse Data Quality Check:



Linked service * AzureSynapseAnalytics

Script * ● Query ○ NonQuery

```
SELECT
    COUNT(*) AS
InvalidLoanAmountCount
```

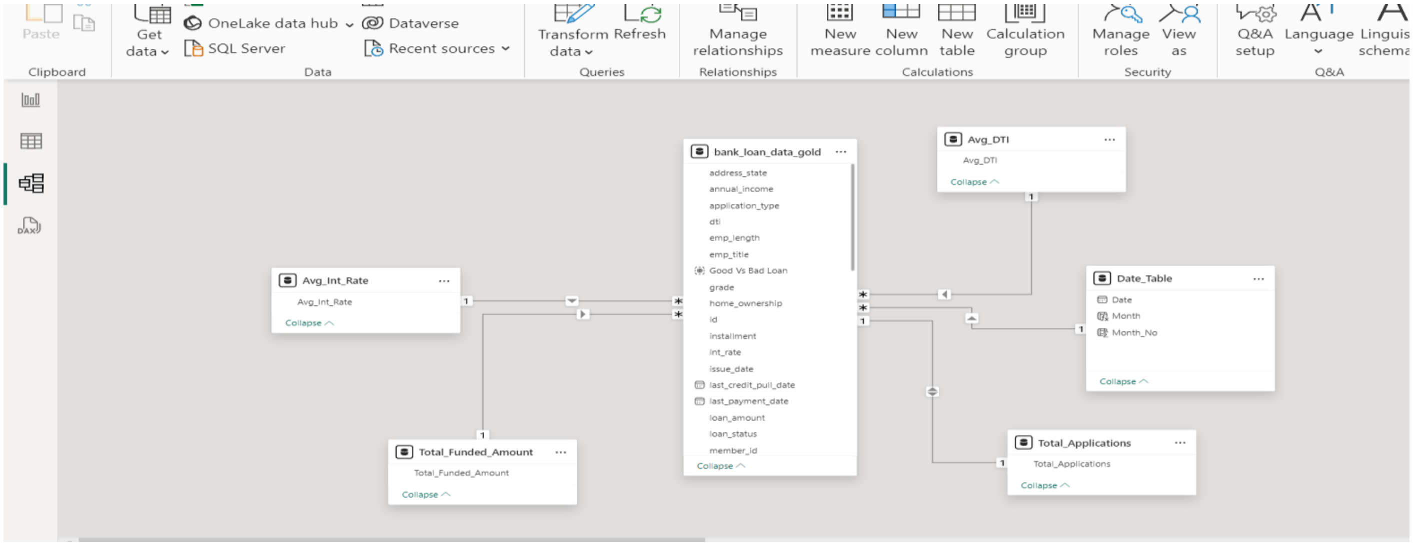
6. **Power BI Integration:**
   - **Objective:** Visualize data from Synapse final table.
   - **Configuration:**
     - Data fetched using DirectQuery for real-time insights.
     - Dashboards include metrics like loan status distribution, average interest rates, and bad loan percentages.

Integration with Power BI

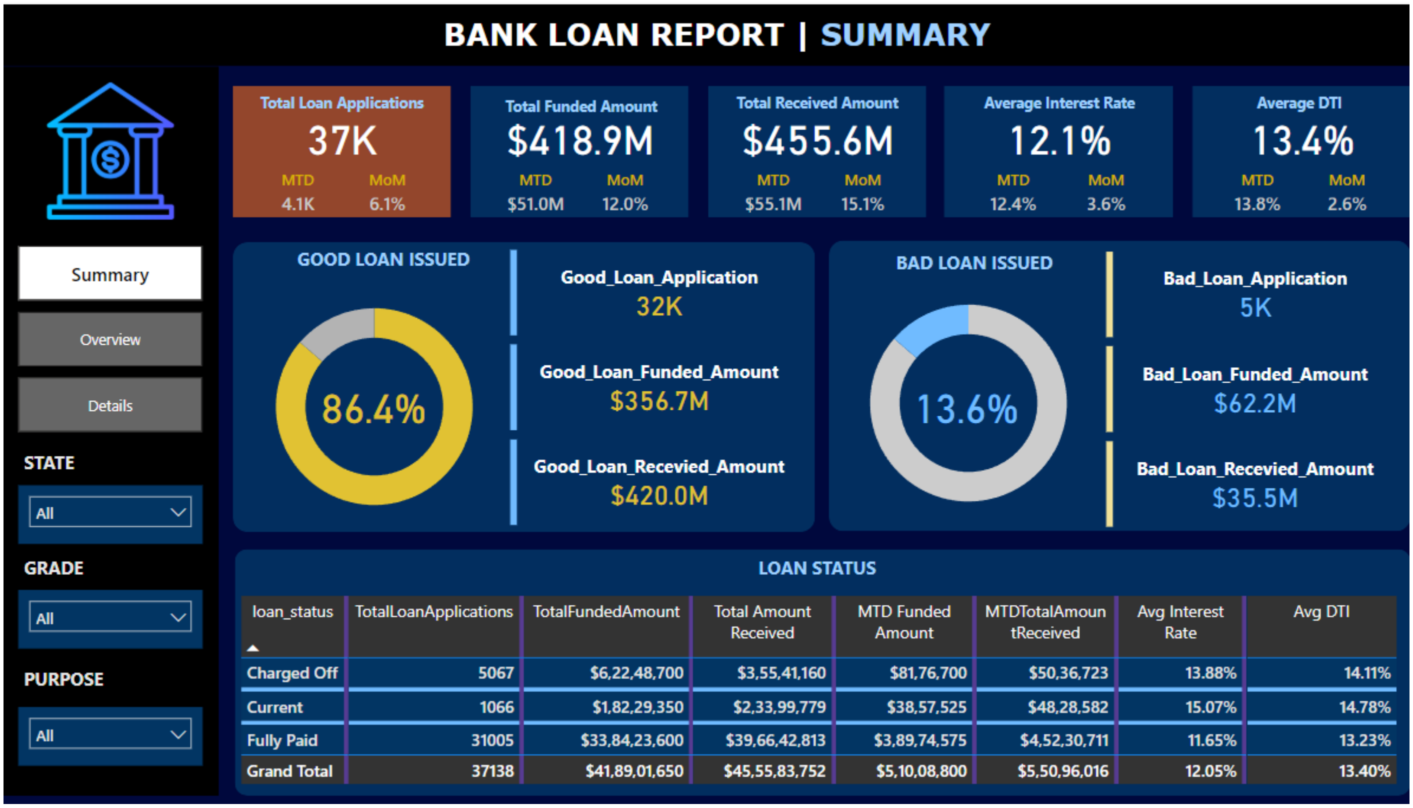
Power BI connects directly to the Synapse table "LoanOProcessData" for real-time reporting. The dashboards provide actionable insights such as:

- Loan status analysis ("Good Vs Bad Loan" split).
- Top states by loan amount.
- Average interest rates by grade and sub-grade.

DataModelling:



Dashboard:



BANK LOAN REPORT | SUMMARY

Total Loan Applications	Total Funded Amount	Total Received Amount	Average Interest Rate	Average DTI
37K	$418.9M	$455.6M	12.1%	13.4%
MTD 4.1K	MTD $51.0M	MTD $55.1M	MTD 12.4%	MTD 13.8%
MoM 6.1%	MoM 12.0%	MoM 15.1%	MoM 3.6%	MoM 2.6%

GOOD LOAN ISSUED

86.4%

Good_Loan_Application 32K

Good_Loan_Funded_Amount $356.7M

Good_Loan_Recevied_Amount $420.0M

BAD LOAN ISSUED

13.6%

Bad_Loan_Application 5K

Bad_Loan_Funded_Amount $62.2M

Bad_Loan_Recevied_Amount $35.5M

LOAN STATUS

loan_status	TotalLoanApplications	TotalFundedAmount	Total Amount Received	MTD Funded Amount	MTDTotalAmountReceived	Avg Interest Rate	Avg DTI
Charged Off	5067	$6,22,48,700	$3,55,41,160	$81,76,700	$50,36,723	13.88%	14.11%
Current	1066	$1,82,29,350	$2,33,99,779	$38,57,525	$48,28,582	15.07%	14.78%
Fully Paid	31005	$33,84,23,600	$39,66,42,813	$3,89,74,575	$4,52,30,711	11.65%	13.23%
Grand Total	37138	$41,89,01,650	$45,55,83,752	$5,10,08,800	$5,50,96,016	12.05%	13.40%

Conclusion

The project successfully implemented a robust ETL pipeline to process loan data from GitHub and provide valuable business insights through Power BI dashboards. The integration of ADF and Synapse Analytics ensures scalability and real-time data processing capabilities.

Future Enhancements

1. **Automated Scheduling:** Implementing triggers in ADF to automate the data pipeline execution.
2. **Real-Time Data Updates:** Integrating Event Hub for streaming real-time data changes.
3. **Advanced Analytics:** Using Synapse ML for predictive modeling of loan defaults.
4. **Data Governance:** Implementing data masking and role-based access control (RBAC) for sensitive information.
5. **Dashboard Enhancements:** Adding drill-through and cross-filtering capabilities in Power BI dashboards for deeper insights.