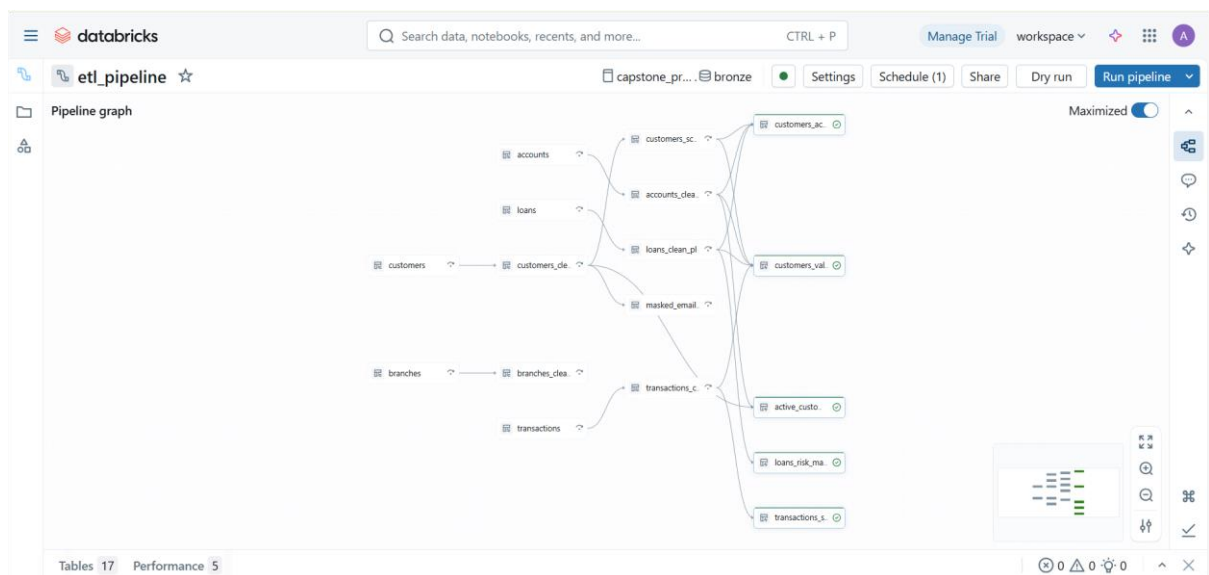


Capstone Project

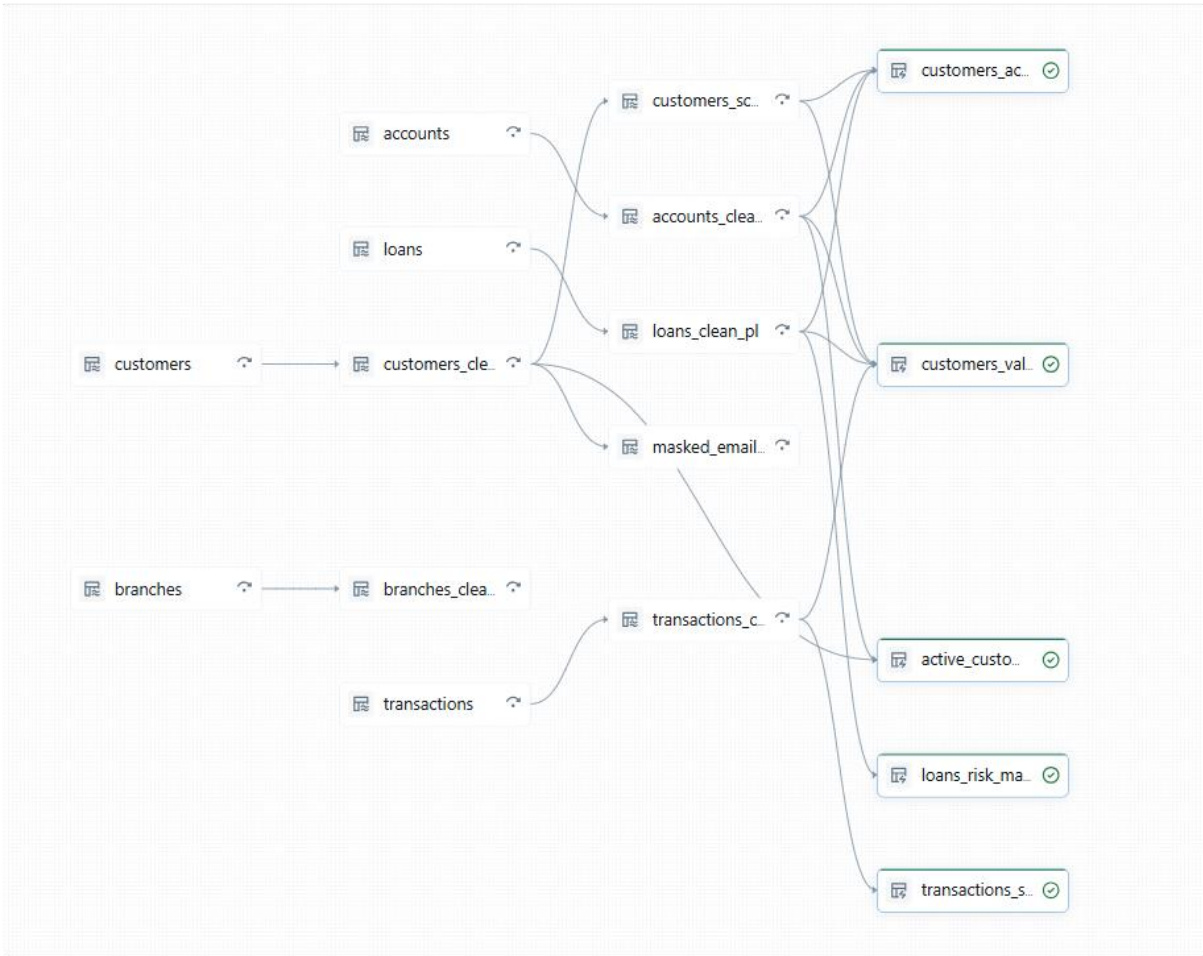
Task Completed

1. Creating Structure Streaming tables
2. Clean Tables (handled Null, Duplicates)
3. Casting (On Date)
4. Masking (Email)
5. Applying SCD
6. Materialized Views
7. Dashboards
8. Dashboard Scheduling
9. Genie Feature
10. Creating a Job for overall ETL process, Dashboards with Scheduling
11. Databricks Version Control
12. Commit to git hub
13. Result in GitHub

Over all pipeline:



All Structure Streaming tables, clean Tables, Casting, Masking, SCD & Materialize Views:



Masking:

▶ 4 days ago (2s) 9

%sql
select * from capstone_project.silver.masked_email_customers
> [See performance \(1\)](#) Optimize

Table +

	customer_id	first_name	last_name	email	phone	date_of_birth	
1	CUST_004387	FirstName4387	LastName4387	xxxx@gmail.com	910397967801	1947-01-12	F
2	CUST_004879	FirstName4879	LastName4879	xxxx@yahoo.com	919238175657	1971-06-06	M
3	CUST_004457	FirstName4457	LastName4457	xxxx@hotmail.com	918624646409	1996-12-13	M
4	CUST_004729	FirstName4729	LastName4729	xxxx@gmail.com	917327502551	1974-01-15	F
5	CUST_002454	FirstName2454	LastName2454	xxxx@gmail.com	913908767254	1983-12-10	M
6	CUST_004780	FirstName4780	LastName4780	xxxx@outlook.co...	911638547799	1969-09-13	F
7	CUST_004434	FirstName4434	LastName4434	xxxx@gmail.com	910034921753	1955-12-13	M
8	CUST_000393	FirstName393	LastName393	xxxx@outlook.co...	917813675009	1957-03-23	M
9	CUST_003509	FirstName3509	LastName3509	xxxx@yahoo.com	913138643490	1973-04-21	F
10	CUST_002536	FirstName2536	LastName2536	xxxx@outlook.co...	910804298602	1942-04-28	M
11	CUST_004299	FirstName4299	LastName4299	xxxx@outlook.co...	913598089496	1941-02-27	F

SCD:

1 minute ago (22s)

8

SQL

✖

+

🔍

🔗

⋮

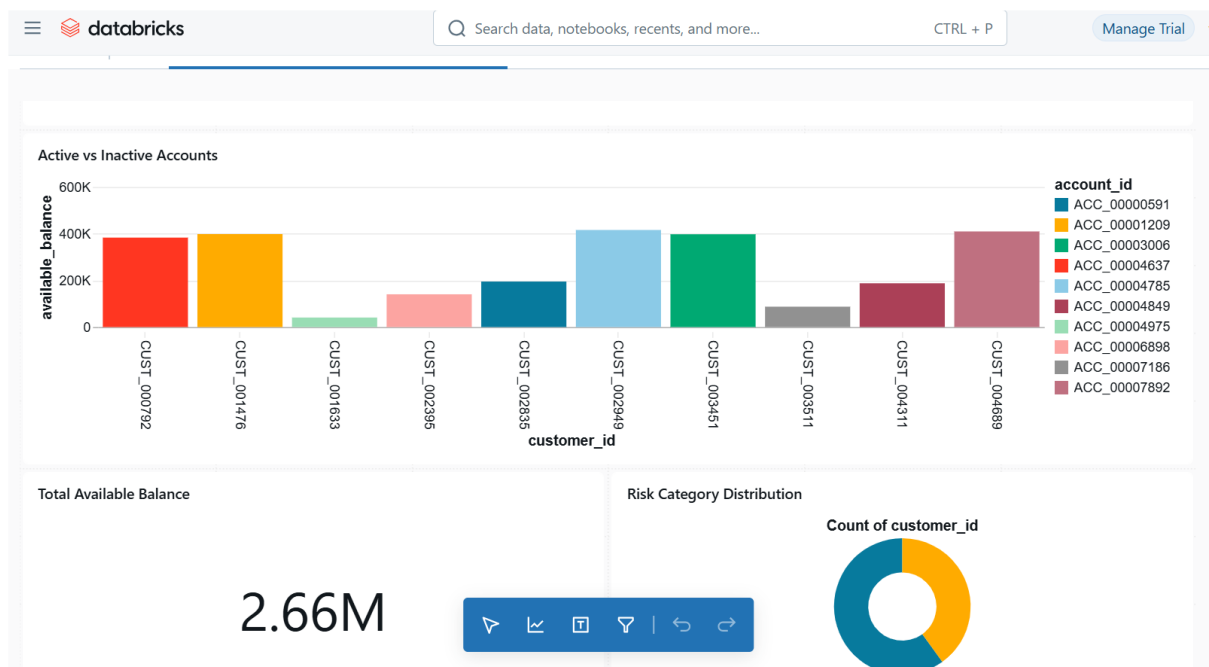
```
%sql
select * from capstone_project.silver.customers_SCD2
```

> [See performance \(1\)](#) Optimize

▸ `_sqldf: pyspark.sql.connect.dataframe.DataFrame = [customer_id: string, first_name: string ... 18 more fields]`

	kyc_status	credit_score	risk_category	is_active	_START_AT	_END_AT
4	verified	619	Medium	true	2025-09-28	null
5	rejected	400	Low	true	2025-09-28	null
6	verified	377	High	true	2025-09-28	null
7	verified	528	Low	true	2025-09-28	null
8	rejected	427	High	true	2025-09-28	null
9	rejected	544	Medium	true	2025-09-28	null
10	rejected	714	High	true	2025-09-28	null
11	rejected	523	Medium	true	2025-09-28	null
12	rejected	576	High	true	2025-09-28	null

Dashboards



Customers Information Dashboards 2025

View Published

38m ago

Data



customer references in accounts, loans, tra...

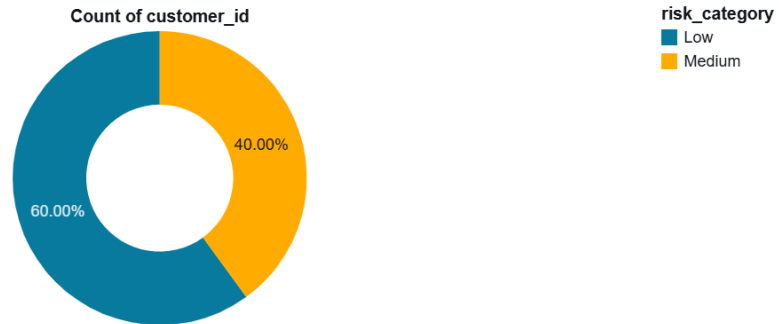
Dashboard with filters



customer_id

All

Risk Category Distribution



Dashboard Scheduling

databricks

Search data, notebooks, recents, and more...

CTRL + P

Manage Trial workspace

Customers Information Dashboards 20...

customer references in accounts, loans, transactions

Dashboard with filters

Customer wise different fields Da

Active vs Inactive Accounts

available_balance

customer_id	available_balance
CUST_000792	400K
CUST_001476	400K
CUST_001633	50K
CUST_002395	150K

Add schedule

Settings Subscribers

Name

New schedule

Schedule

Every Day at 09 : 00

☐ Show cron syntax

Timezone

Asia/Calcutta

Apply filters

Dashboard filters applied to schedule, email content & email links

☒ Use default values

☐ Use filter selections

Data updated based on the values of filters when schedule is created

Genie Feature

New

Workspace

Recents

Catalog

Jobs & Pipelines

Compute

Marketplace

SQL

SQL Editor

Queries

Dashboards

Genie

Alerts

Query History

SQL Warehouses

New Space

The SQL Warehouse Serverless Starter Warehouse is running.

was this correct?

What are the counts of active versus inactive customers?

Analysis complete customers_accounts_loans

is_active: IS NOT NULL

1 row

Hide code

```
1 SELECT
2   `is_active`,
3   COUNT(DISTINCT `customer_id`) AS `customer_count`
4 FROM
5   `capstone_project`.`gold`.`customers_accounts_loans`
6 WHERE
7   `is_active` IS NOT NULL
```

What is the distribution of available balance among active and inactive

Data

Instructions

Settings

Link existing or upload new data to start discovering insights.


Name	Type
customers_accounts_loans	Table
capstone_project.gold	

```
1 SELECT
2   `is_active`,
3   COUNT(DISTINCT `customer_id`) AS `customer_count`
4 FROM
5   `capstone_project`.`gold`.`customers_accounts_loans`
6 WHERE
7   `is_active` IS NOT NULL
8 GROUP BY
9   `is_active`
```

▶ Run


	is_active	customer_count
1	true	2261

Creating a Job for overall ETL process, Dashboards with Scheduling

 databricks

Search data, notebooks, recents, and more...

Jobs & Pipelines >

 **Capstone ETL Job** ☆ Lakeflow Jobs UI: ON [Send feedback](#)

Runs **Tasks**

Capstone_ETL_Process
Pipeline: etl_pipeline

Capstone_Dashboards
Customers Information Dashboards 2...

+ Add task

Task name* ⓘ
Capstone_Dashboards

Type*
Dashboard

Dashboard*
Customers Information Dashboards 2025-10-01 23:14:10

SQL warehouse ⓘ
Serverless Starter Warehouse (S)

Cancel Save task

With Scheduling:

Schedules & Triggers

Trigger Status

- Active
- Paused

Trigger type

Scheduled

Schedule type

Simple

Advanced

Periodic ⓘ

Every 1 Day

Cancel Save

Capstone ETL Job Lakeflow Jobs UI: ON [Send feedback](#) [Run now](#)

Runs **Tasks**

Capstone_ETL_Process
Pipeline: etl_pipeline

Capstone_Dashboards
Customers Information Dashboards 2...

+ Add task

Task name* Capstone_Dashboards

Type* Dashboard

Dashboard* Customers Information Dashboards 2025-10-01 23:14:10

SQL warehouse Serverless Starter Warehouse (S)

Cancel

Save task

Run as ashwini04.yadav10@gmail.com

Description Add description

Lineage No lineage information for this job. [Learn more](#)

Performance optimized ☒

Schedules & Triggers
Every day, next run at Oct 02, 2025, 11:28 AM
[Edit trigger](#) [Pause](#) [Delete](#)

Job parameters
No job parameters are defined for this job

Databricks Version Control:

etl_pipeline capstone_pr... bronze Connect Settings Schedule Share Dry run [Run pipeline](#)

bronze_streaming_tables.sql silver_transformations.sql gold_views.sql

Oct 01, 2025, 11:59 PM (Previous version)

Oct 02, 2025, 12:01 AM

Lakeflow Pipelines Editor: ON [Last edit was 51 minutes ago](#)

1 -- Valid customer references in accounts, loans, transactions

2

3 -- Valid customer references in accounts, loans, transactions

4

5

6 CREATE MATERIALIZED VIEW gold.customers_valid AS

7 SELECT

8 c.customer_id,

9 c.first_name,

10 c.last_name,

11 a.account_id,

12 t.transaction_id,

13 l.loan_id

14 FROM silver.customers_scd2 c

1 -- Valid customer references in accounts, loans, transactions

2

3 -- Just View

4

5 CREATE MATERIALIZED VIEW gold.customers_valid AS

6 SELECT

7 c.customer_id,

8 c.first_name,

9 c.last_name,

10 a.account_id,

11 t.transaction_id,

12 l.loan_id

13 FROM silver.customers_scd2 c

Version history

[Clear history](#)

Oct 02, 2025, 12:01 AM
ashwini04.yadav10@gmail.com
[Save now](#)

Oct 01, 2025, 11:59 PM
ashwini04.yadav10@gmail.com

Oct 01, 2025, 11:43 PM
ashwini04.yadav10@gmail.com

Oct 01, 2025, 11:41 PM
ashwini04.yadav10@gmail.com

Oct 01, 2025, 11:38 PM
ashwini04.yadav10@gmail.com

Oct 01, 2025, 11:36 PM
ashwini04.yadav10@gmail.com

Oct 01, 2025, 11:33 PM
ashwini04.yadav10@gmail.com

Oct 01, 2025, 11:27 PM
ashwini04.yadav10@gmail.com

Oct 01, 2025, 11:20 PM
ashwini04.yadav10@gmail.com

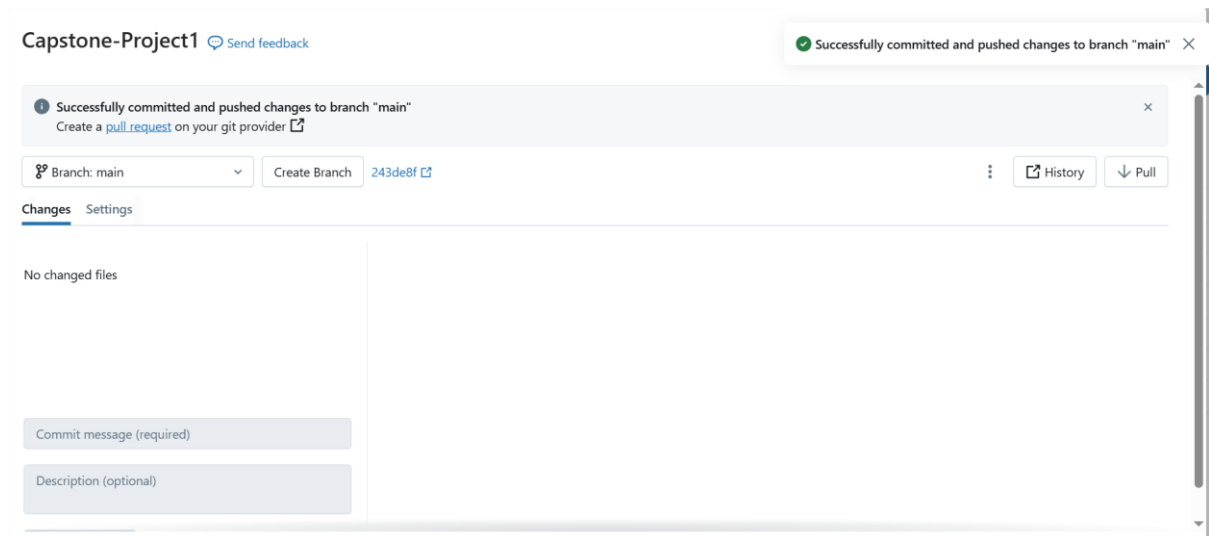
Sep 29, 2025, 04:11 PM

Tables 14 Performance 2

Oct 01, 2025, 11:42 PM 18s Refresh all [Query performance](#)

Commit to git hub:

All code committed to my git hub account.



Now see the committed files in to my git hub account:

