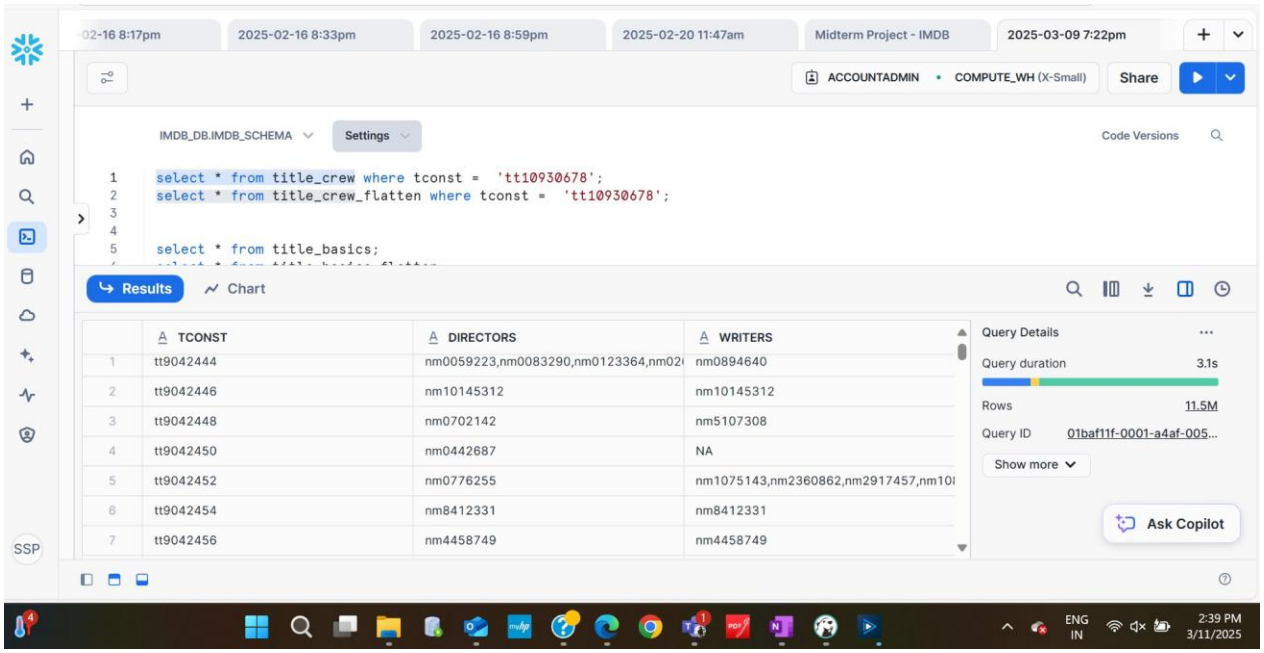
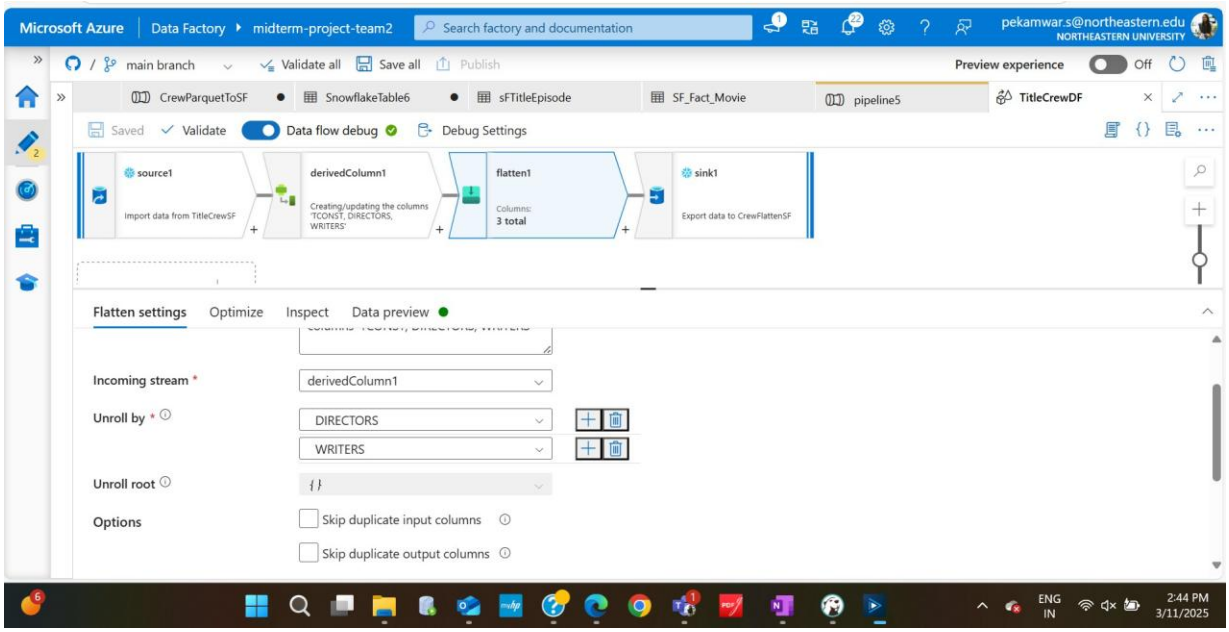
Data Flow pipelines and Load

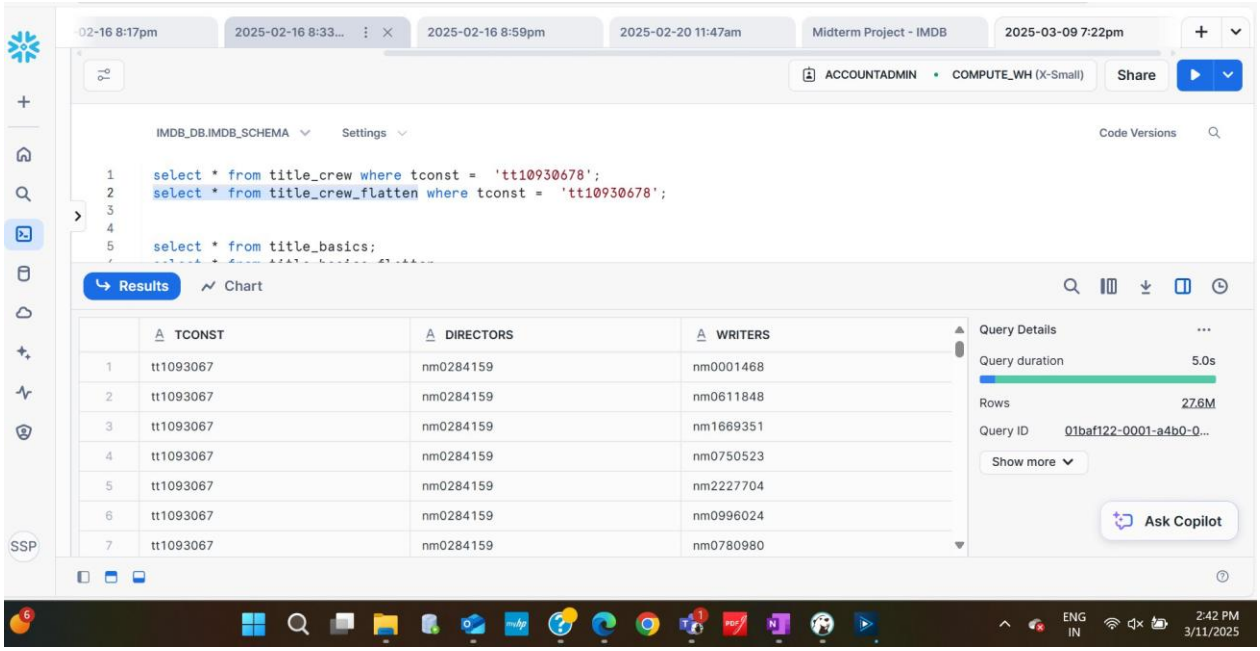
Stage Data Loading

1] title.crew dataset

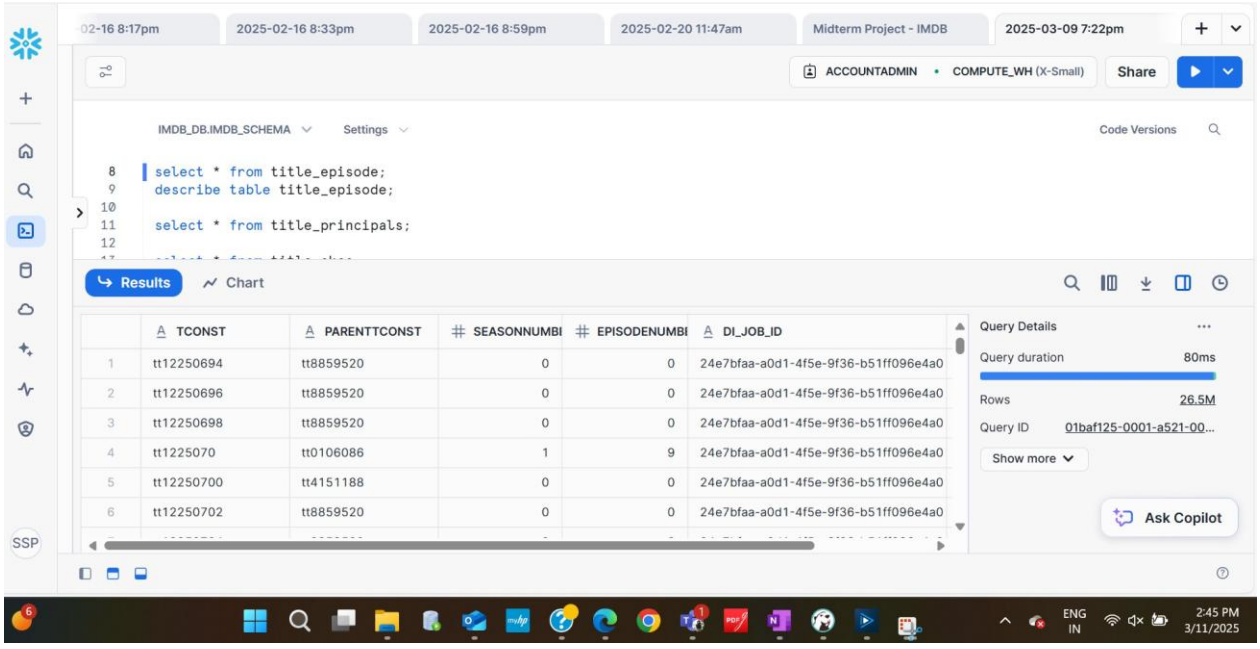


As the data in writers and directors column has values in the form of string, the data is flattened using adf pipeline

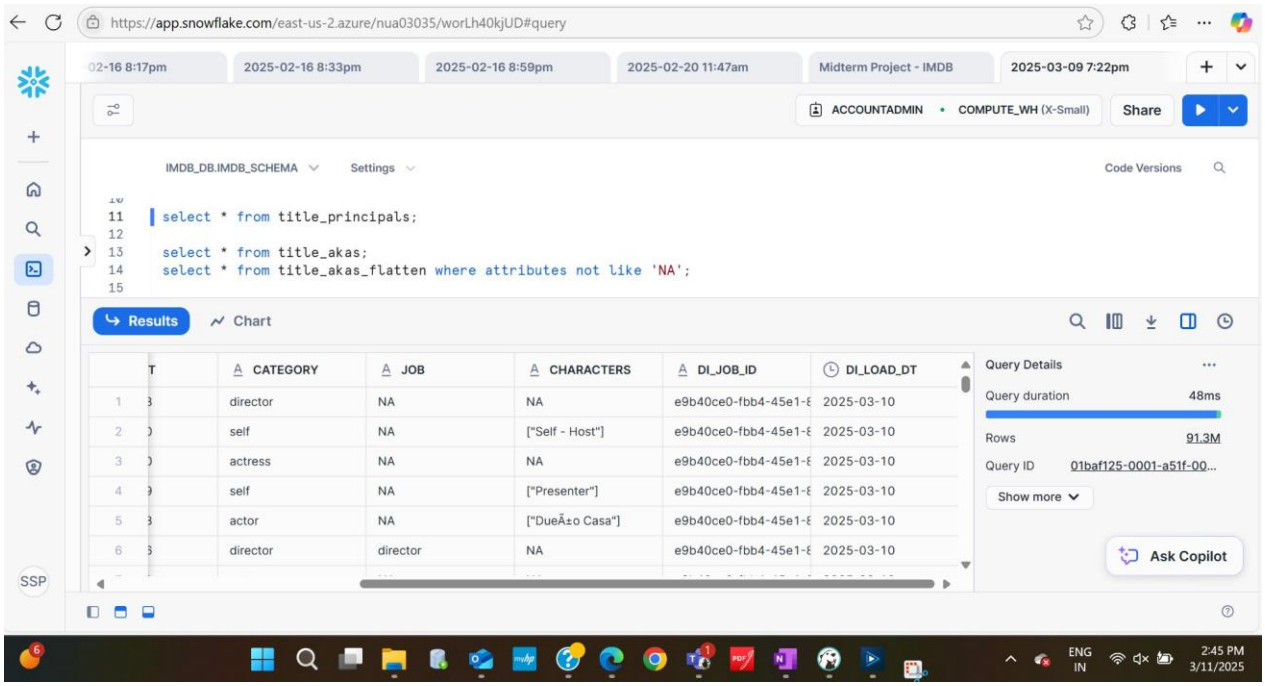




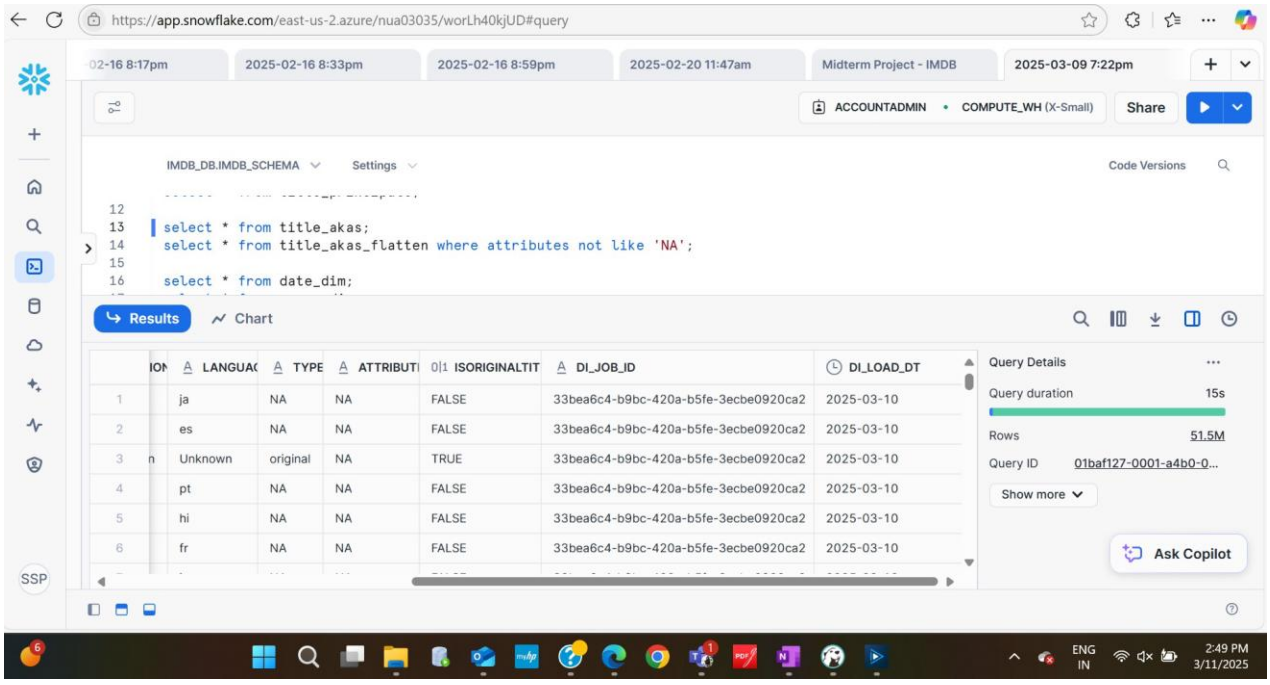
2] title.episode



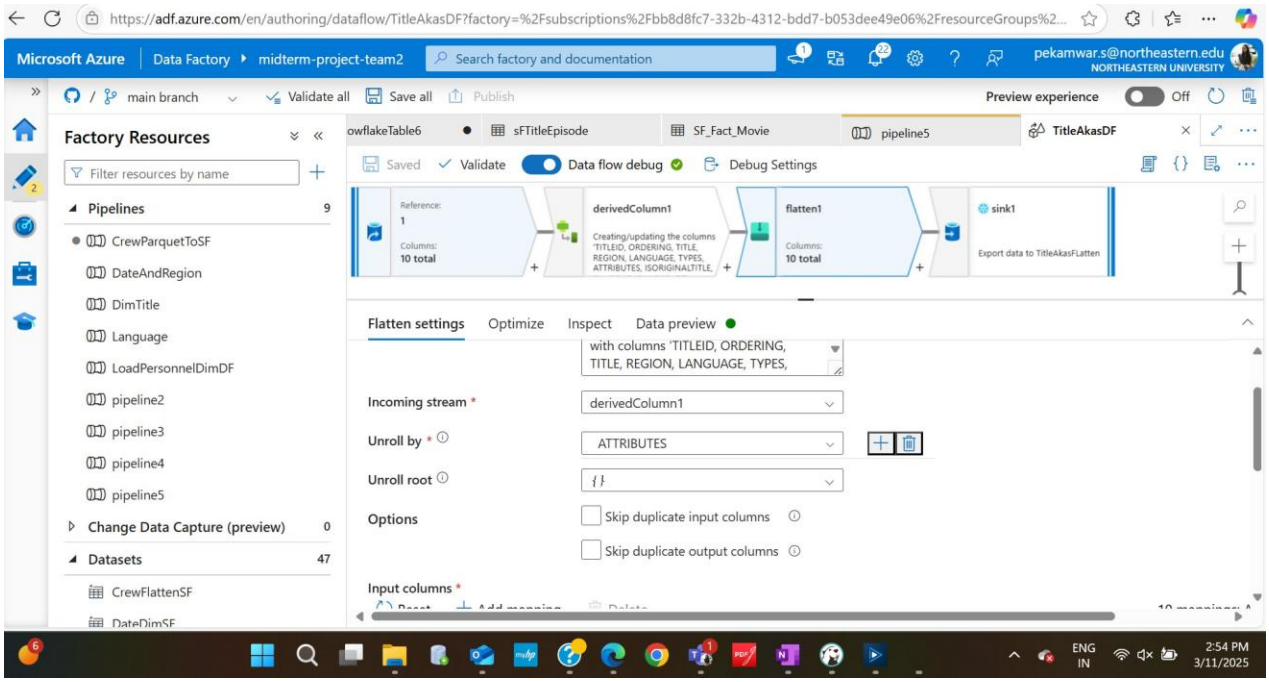
3] title.principals

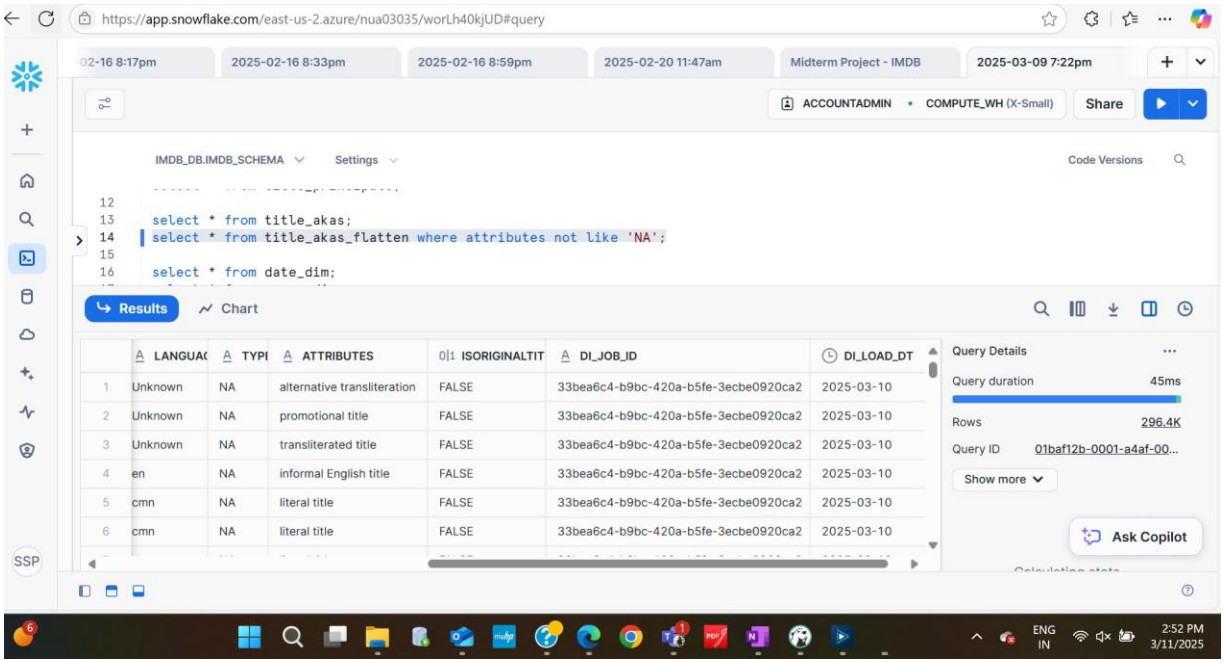


4] title.akas

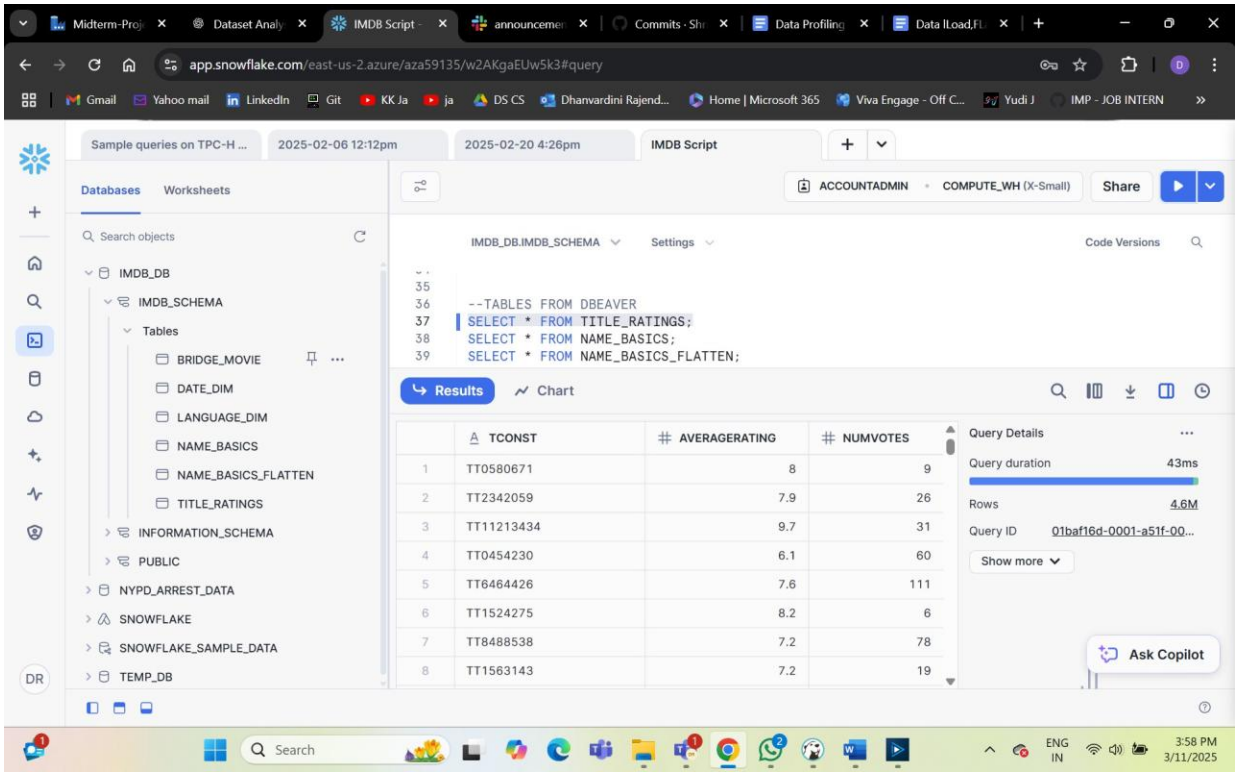


Here the attributes column is flattened and separated into different columns

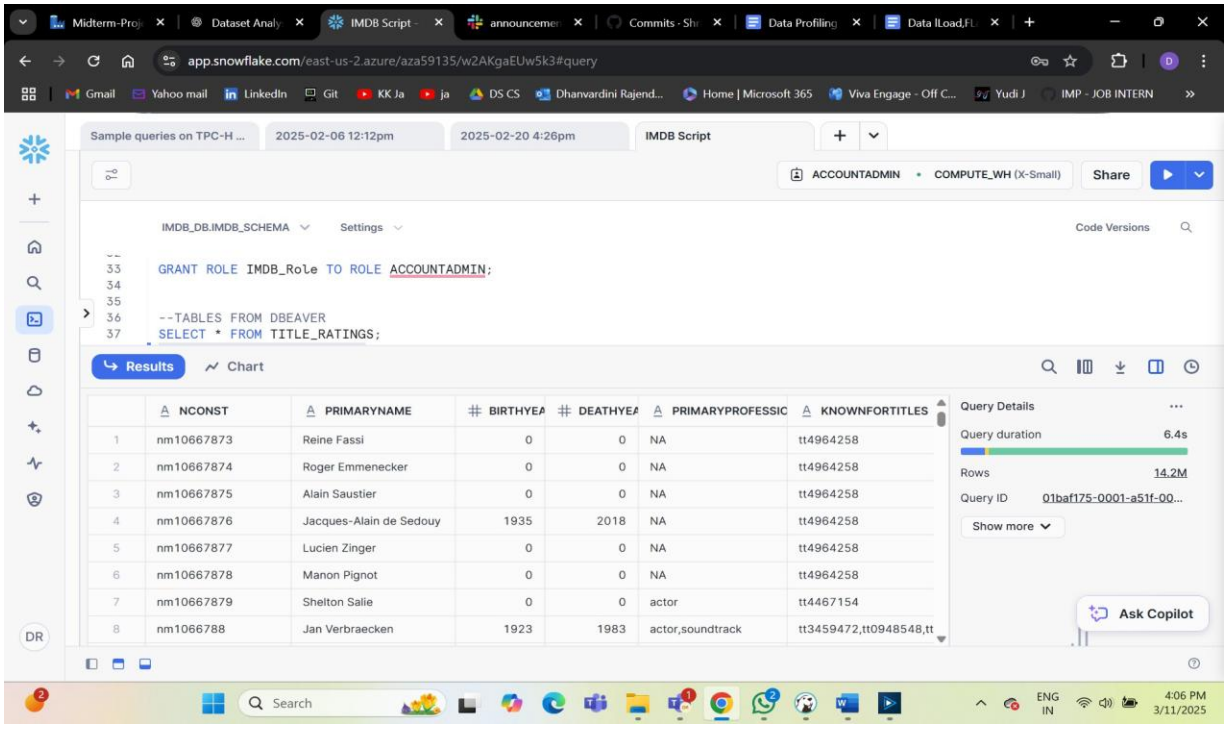

https://adf.azure.com/en/authoring/dataflow/TitleAkasDF?factory=%2Fsubscriptions%2Fbb8d8fc7-332b-4312-bdd7-b053dee49e06%2FresourceGroups%2...
Microsoft Azure
Data Factory
midterm-project-team2
Search factory and documentation
pekamwar.s@northeastern.edu
NORTHEASTERN UNIVERSITY
main branch
Validate all
Save all
Publish
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 9
CrewParquetToSF
DateAndRegion
DimTitle
Language
LoadPersonnelDimDF
pipeline2
pipeline3
pipeline4
pipeline5
Change Data Capture (preview) 0
Datasets 47
CrewFlattenSF
DateDimSF
owflakeTable6
sFTitleEpisode
SF_Fact_Movie
pipeline5
TitleAkasDF
Saved
Validate
Data flow debug
Debug Settings
Reference: 1
Columns: 10 total
derivedColumn1
Creating/updating the columns 'TITLEID, ORDERING, TITLE, REGION, LANGUAGE, TYPES, ATTRIBUTES, ISORIGINALTITLE,
flatten1
Columns: 10 total
sink1
Export data to TitleAkasFlatten
Flatten settings
Optimize
Inspect
Data preview
with columns 'TITLEID, ORDERING, TITLE, REGION, LANGUAGE, TYPES,
Incoming stream *
derivedColumn1
Unroll by *
ATTRIBUTES
Unroll root
{}
Options
Skip duplicate input columns
Skip duplicate output columns
Input columns *
ENG
IN
2:54 PM
3/11/2025

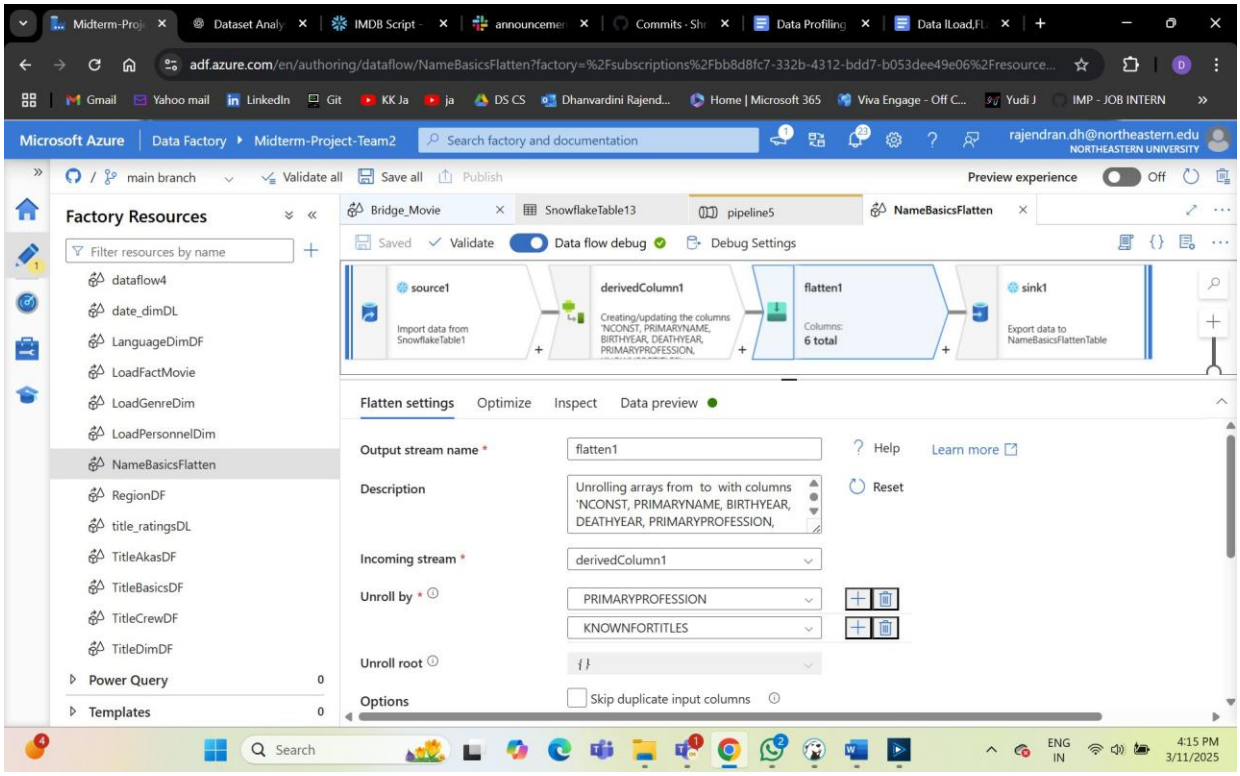

https://app.snowflake.com/east-us-2.azure/nua03035/worLh40kjUD#query
02-16 8:17pm
2025-02-16 8:33pm
2025-02-16 8:59pm
2025-02-20 11:47am
Midterm Project - IMDB
2025-03-09 7:22pm
ACCOUNTADMIN
COMPUTE_WH (X-Small)
Share
IMDB_DB.IMDB_SCHEMA
Settings
Code Versions
select * from title_akas;
select * from title_akas_flatten where attributes not like 'NA';
select * from date_dim;
Results
Chart
LANGUAG
TYPE
ATTRIBUTES
ISORIGINALTIT
DI_JOB_ID
DI_LOAD_DT
1 Unknown NA alternative transliteration FALSE 33bea6c4-b9bc-420a-b5fe-3ecbe0920ca2 2025-03-10
2 Unknown NA promotional title FALSE 33bea6c4-b9bc-420a-b5fe-3ecbe0920ca2 2025-03-10
3 Unknown NA transliterated title FALSE 33bea6c4-b9bc-420a-b5fe-3ecbe0920ca2 2025-03-10
4 en NA informal English title FALSE 33bea6c4-b9bc-420a-b5fe-3ecbe0920ca2 2025-03-10
5 cmn NA literal title FALSE 33bea6c4-b9bc-420a-b5fe-3ecbe0920ca2 2025-03-10
6 cmn NA literal title FALSE 33bea6c4-b9bc-420a-b5fe-3ecbe0920ca2 2025-03-10
Query Details
Query duration 45ms
Rows 296.4K
Query ID 01baf12b-0001-a4af-00...
Show more
Ask Copilot
ENG
IN
2:52 PM
3/11/2025

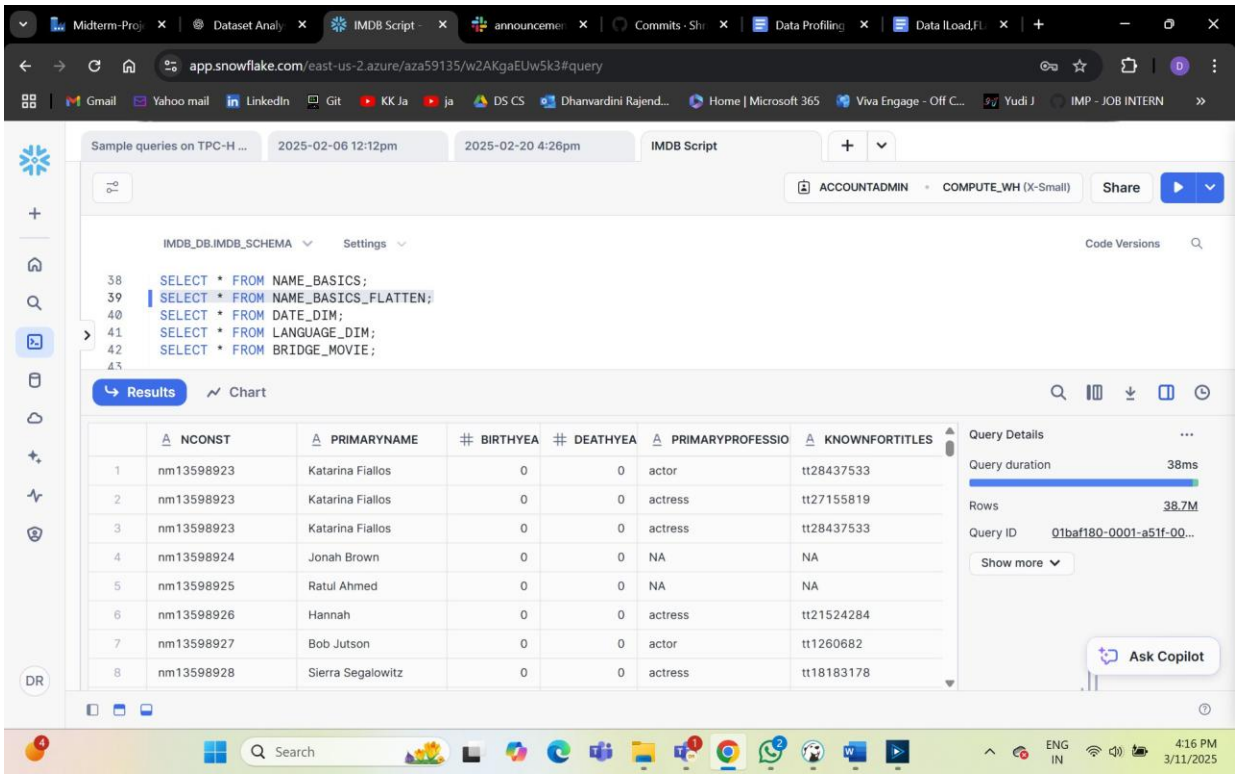
5] title.ratings



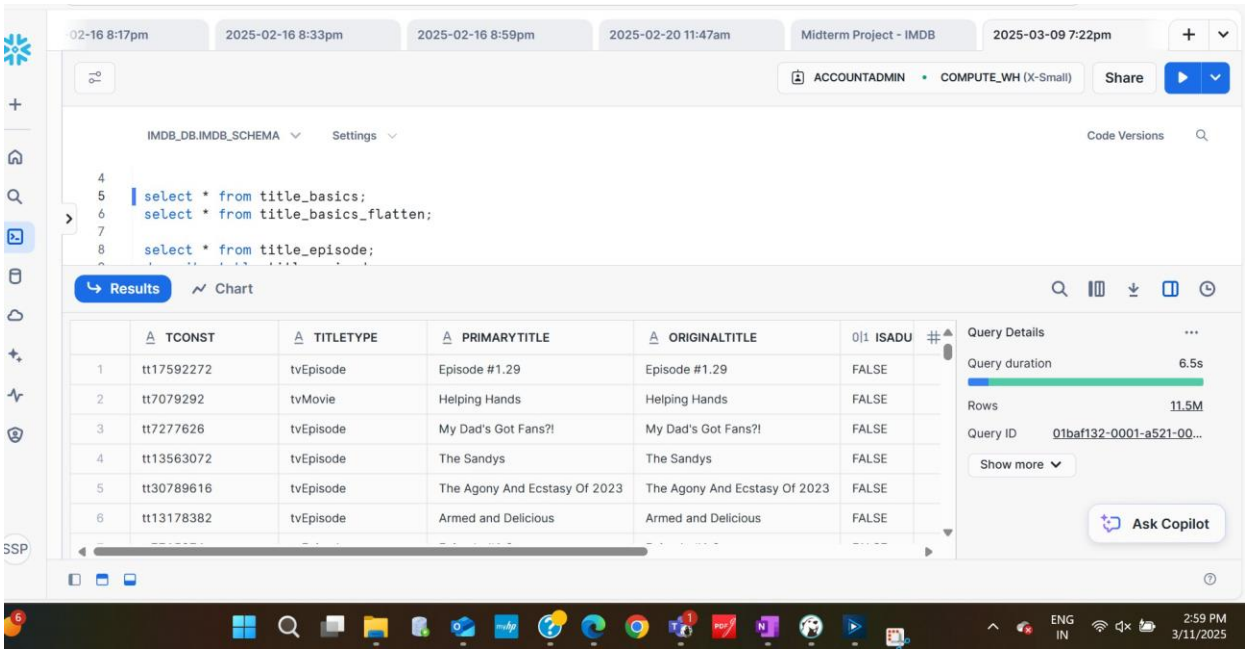
6] name.basics



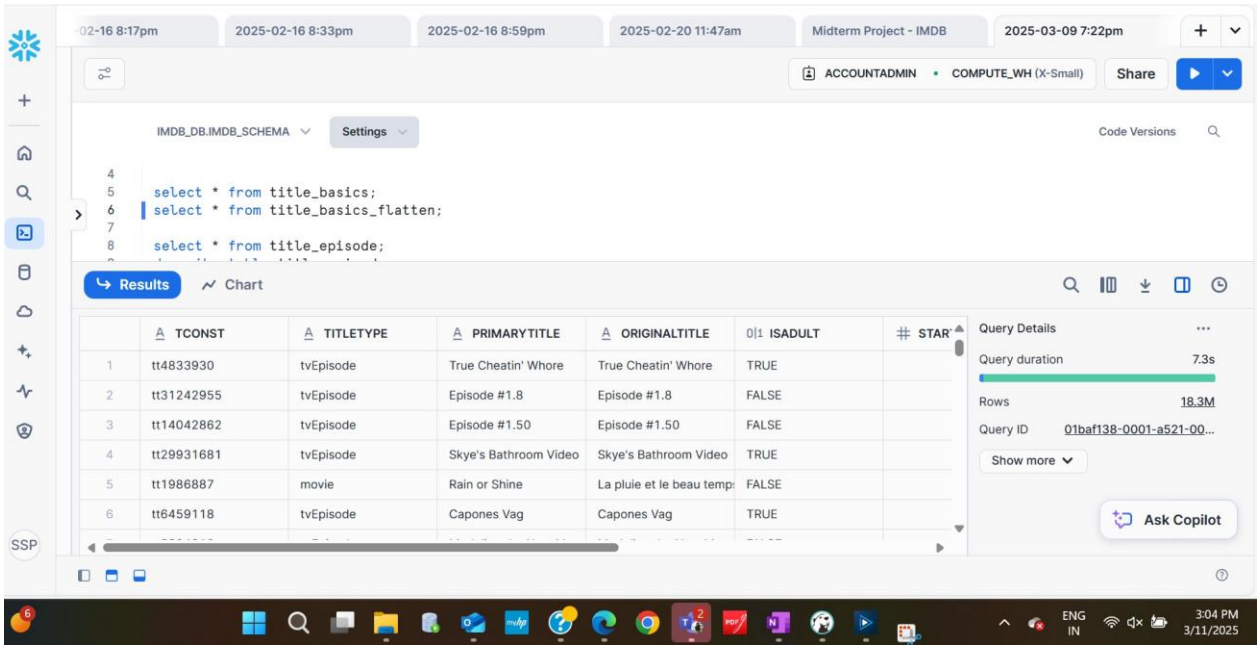

Microsoft Azure
Data Factory
Midterm-Project-Team2
Factory Resources
NameBasicsFlatten
source1
derivedColumn1
flatten1
sink1
Flatten settings
Optimize
Inspect
Data preview
Output stream name *
flatten1
Description
Unrolling arrays from to with columns 'NCONST, PRIMARYNAME, BIRTHYEAR, DEATHYEAR, PRIMARYPROFESSION,
Incoming stream *
derivedColumn1
Unroll by *
PRIMARYPROFESSION
KNOWNFORTITLES
Unroll root
Options
Skip duplicate input columns


IMDB Script
IMDB_DB.IMDB_SCHEMA
SELECT * FROM NAME_BASICS;
SELECT * FROM NAME_BASICS_FLATTEN;
SELECT * FROM DATE_DIM;
SELECT * FROM LANGUAGE_DIM;
SELECT * FROM BRIDGE_MOVIE;
Results
Chart
NCONST
PRIMARYNAME
BIRTHYEA
DEATHYEA
PRIMARYPROFESSIO
KNOWNFORTITLES
nm13598923 Katarina Fiallos 0 0 actor tt28437533
nm13598923 Katarina Fiallos 0 0 actress tt27155819
nm13598923 Katarina Fiallos 0 0 actress tt28437533
nm13598924 Jonah Brown 0 0 NA NA
nm13598925 Ratul Ahmed 0 0 NA NA
nm13598926 Hannah 0 0 actress tt21524284
nm13598927 Bob Jutson 0 0 actor tt1260682
nm13598928 Sierra Segalowitz 0 0 actress tt18183178
Query Details
Query duration 38ms
Rows 38.7M
Ask Copilot

7] title.basics

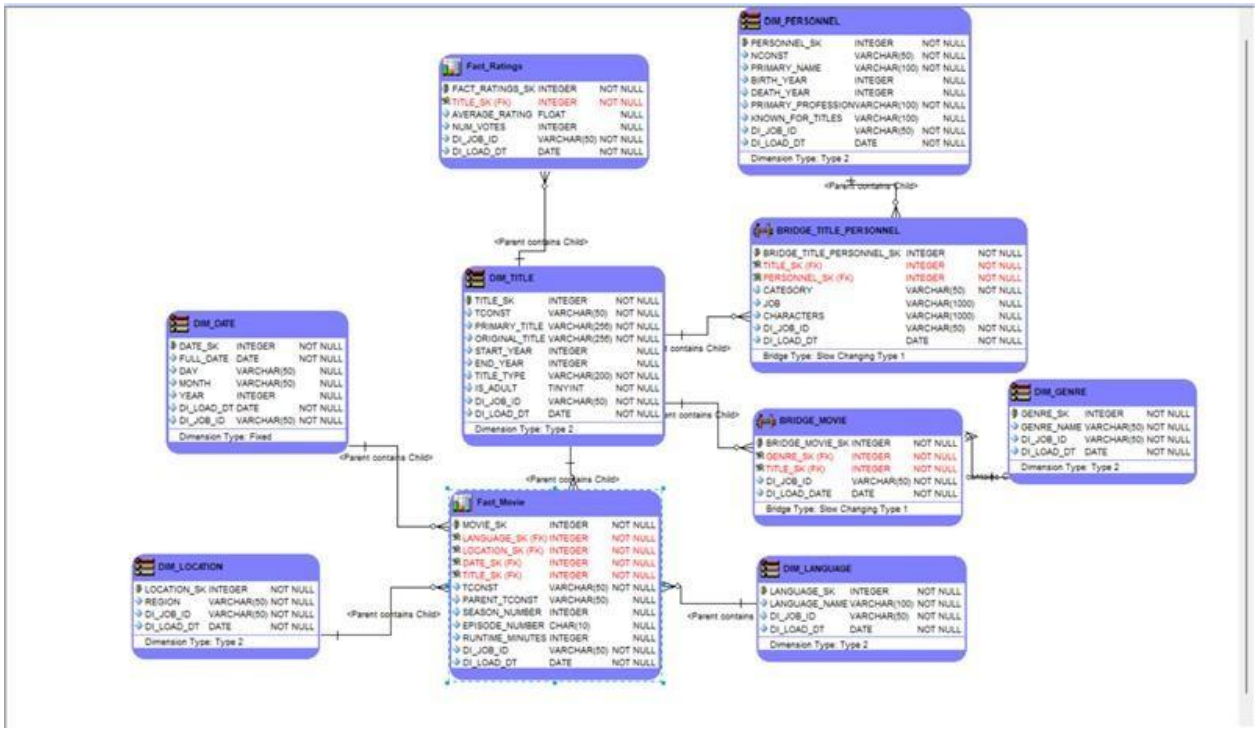


Here the flattening of columns on Genres, OriginalTitle and PrimaryTitle into a new table

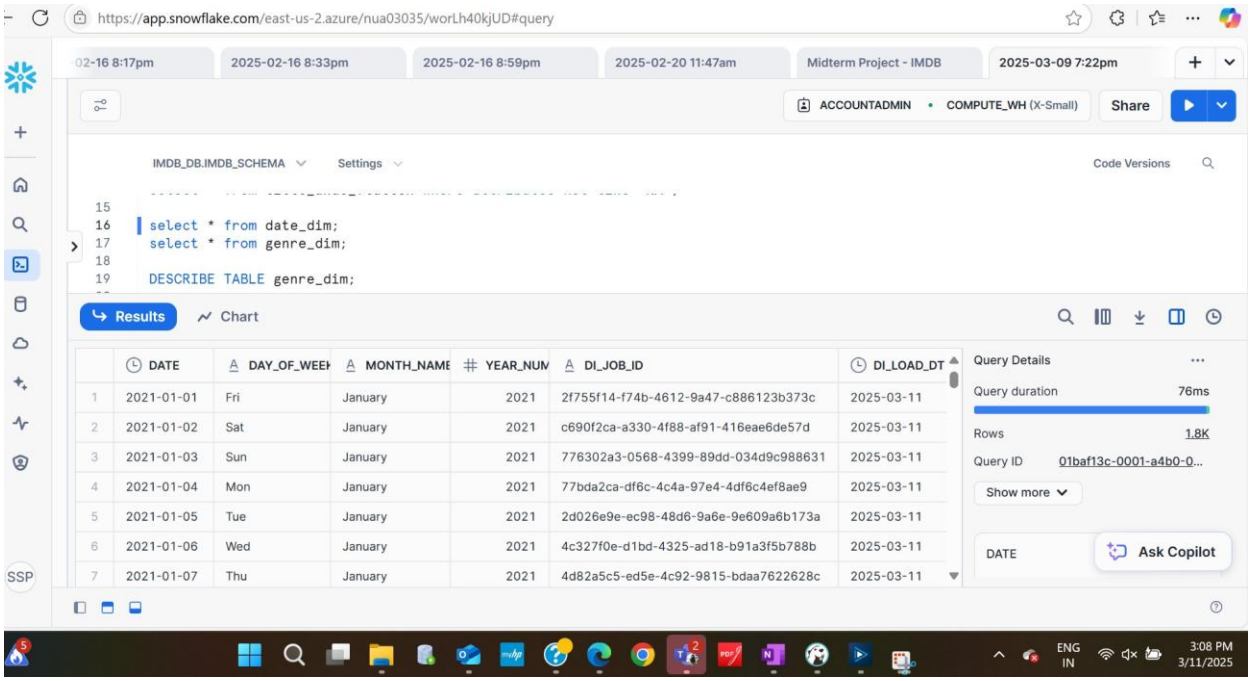


Loading Dims and Facts Tables :

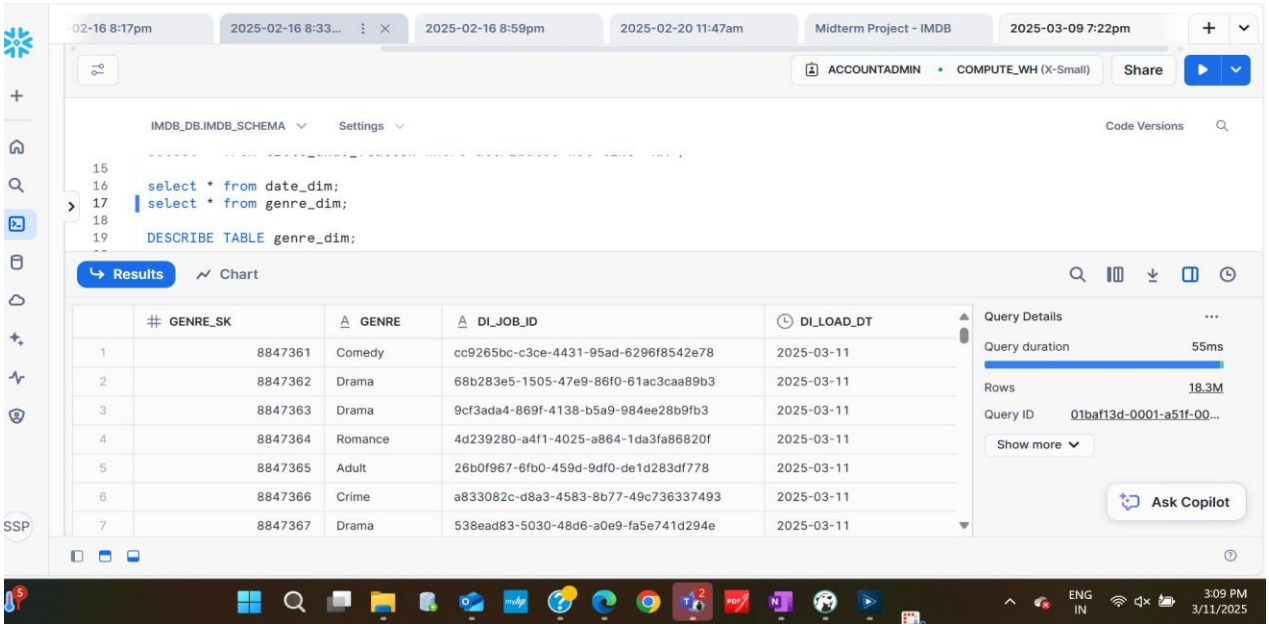
ER Studio Diagram :



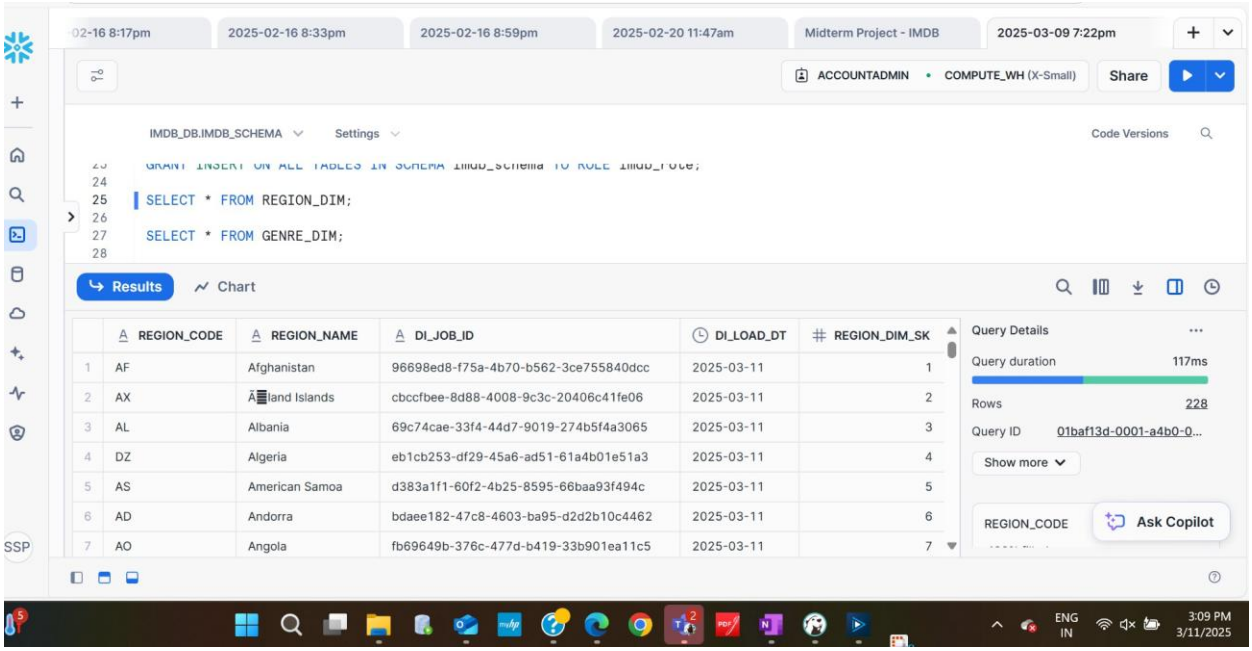
1] Date_Dim



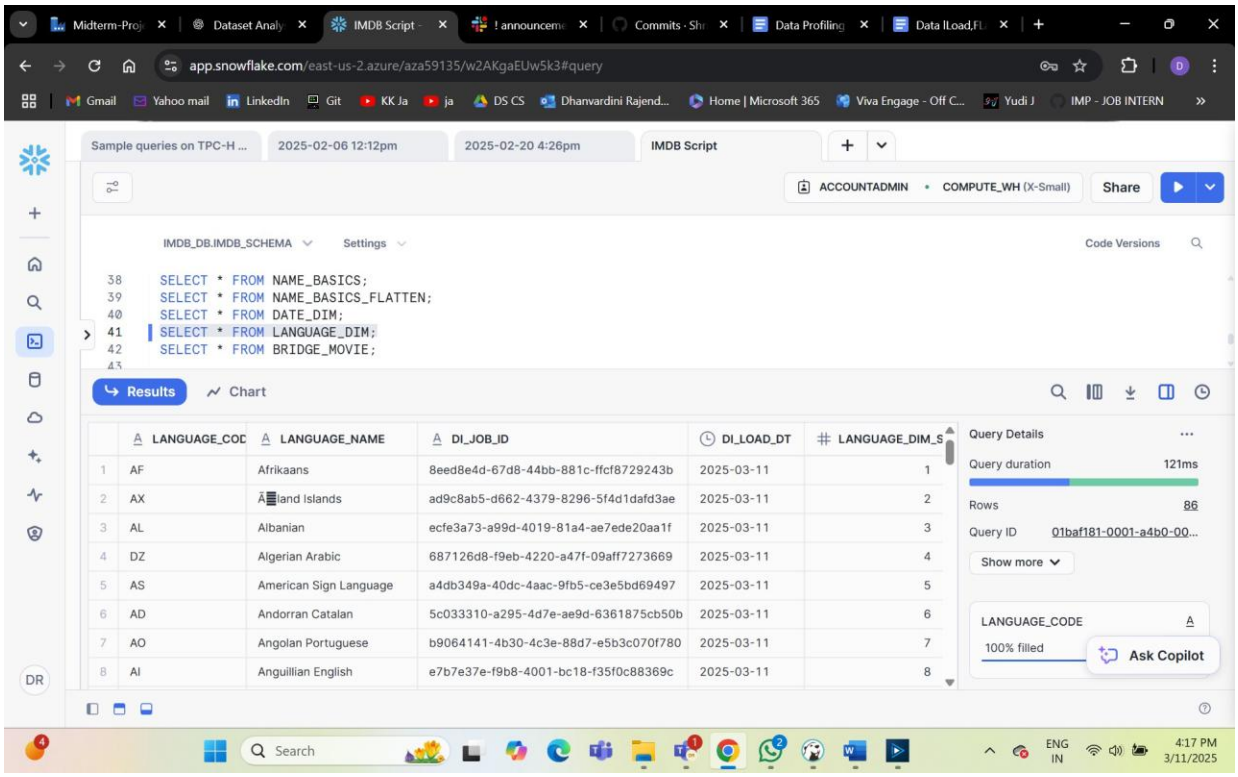
2] Genre _Dim



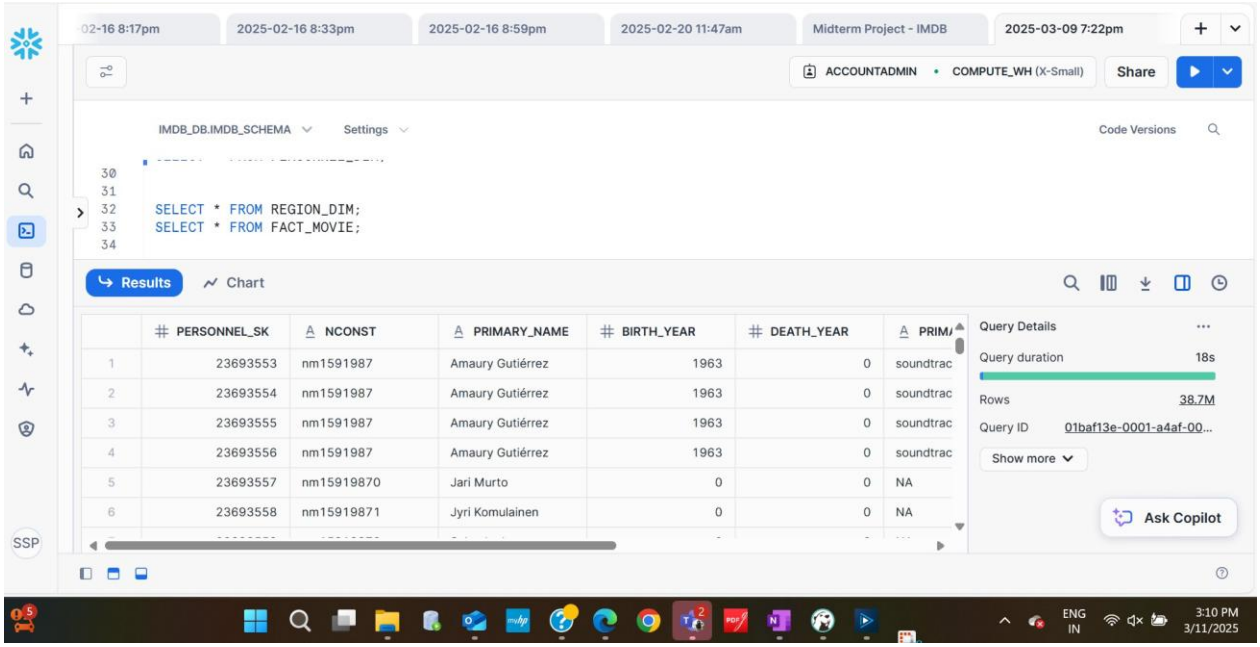
3] Region_Dim



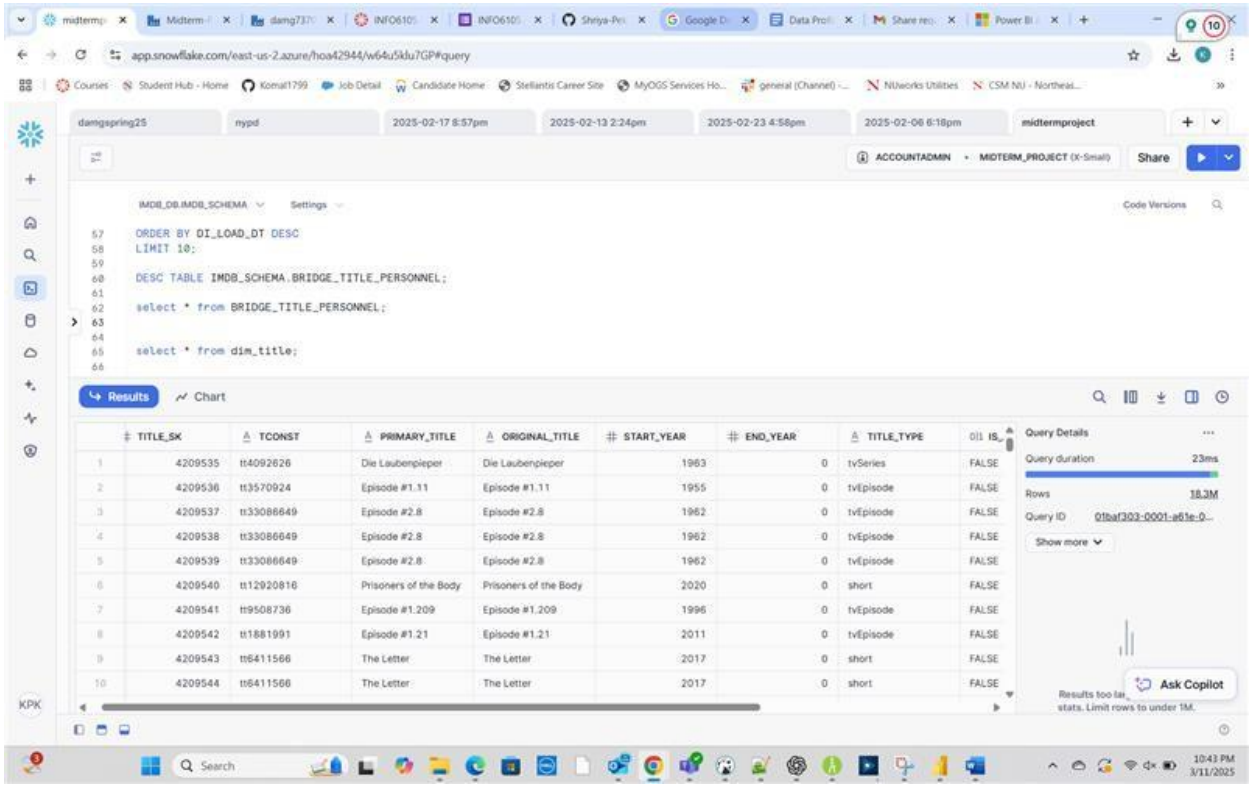
4] Language_Dim



5]Personnel_Dim

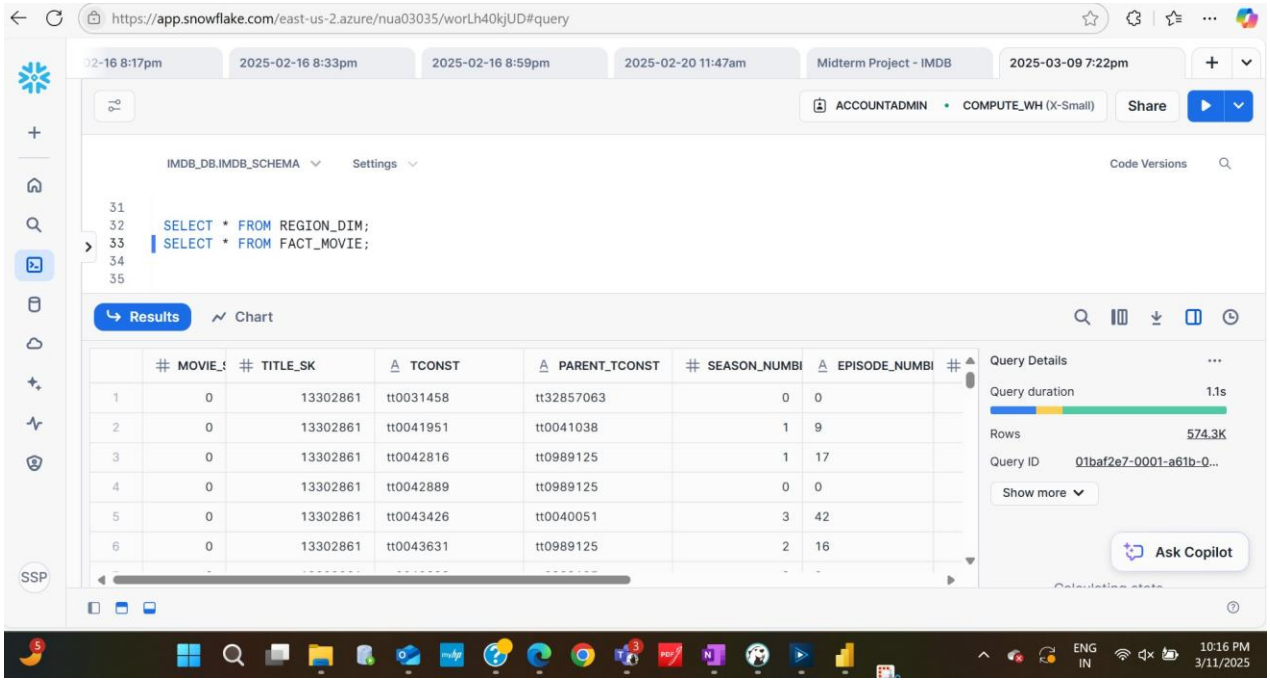


6] Title_Dim

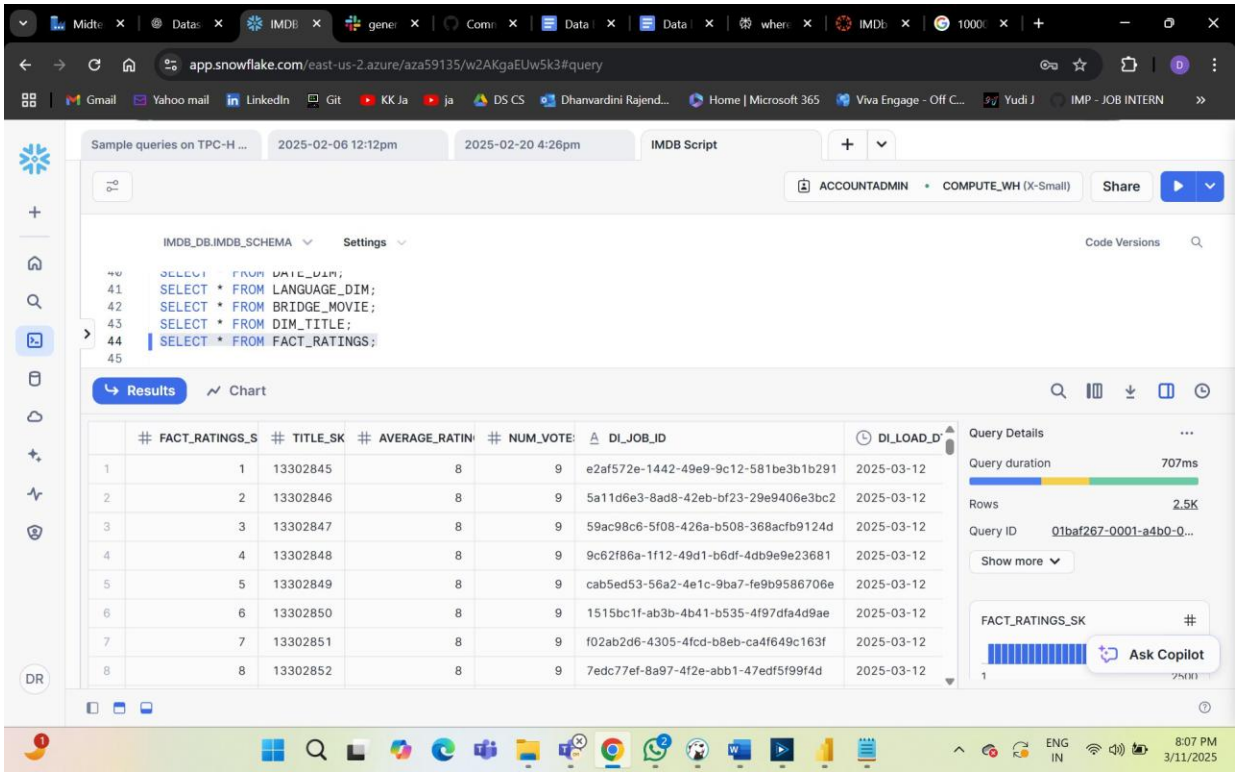


FACTS :

1] Fact _Movie



2] Fact_Ratings



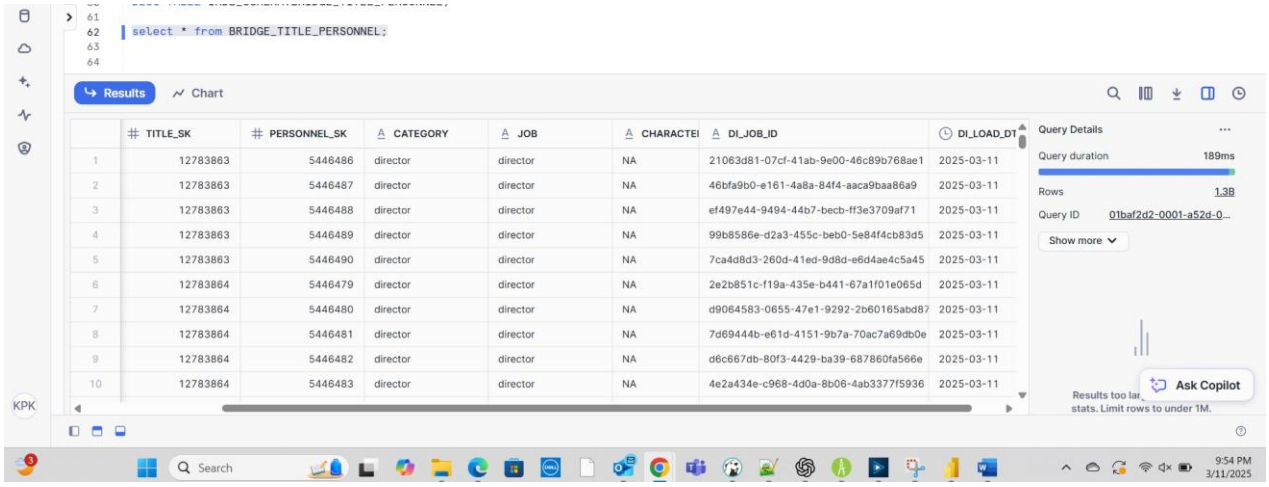
BRIDGE TABLES :

1] Bridge Movie

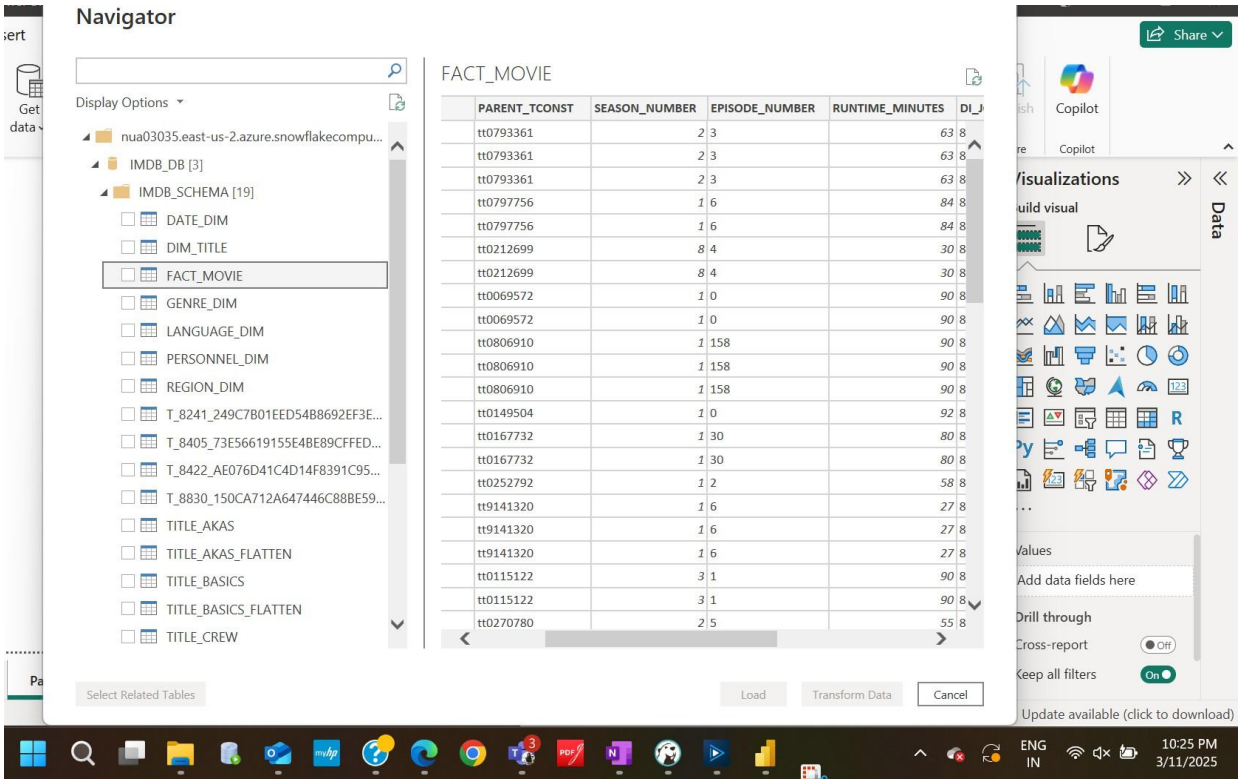
Dim_Title and Dim_Genre

2]Bridge Title Personnel

Dim_Title and Dim_Personnel

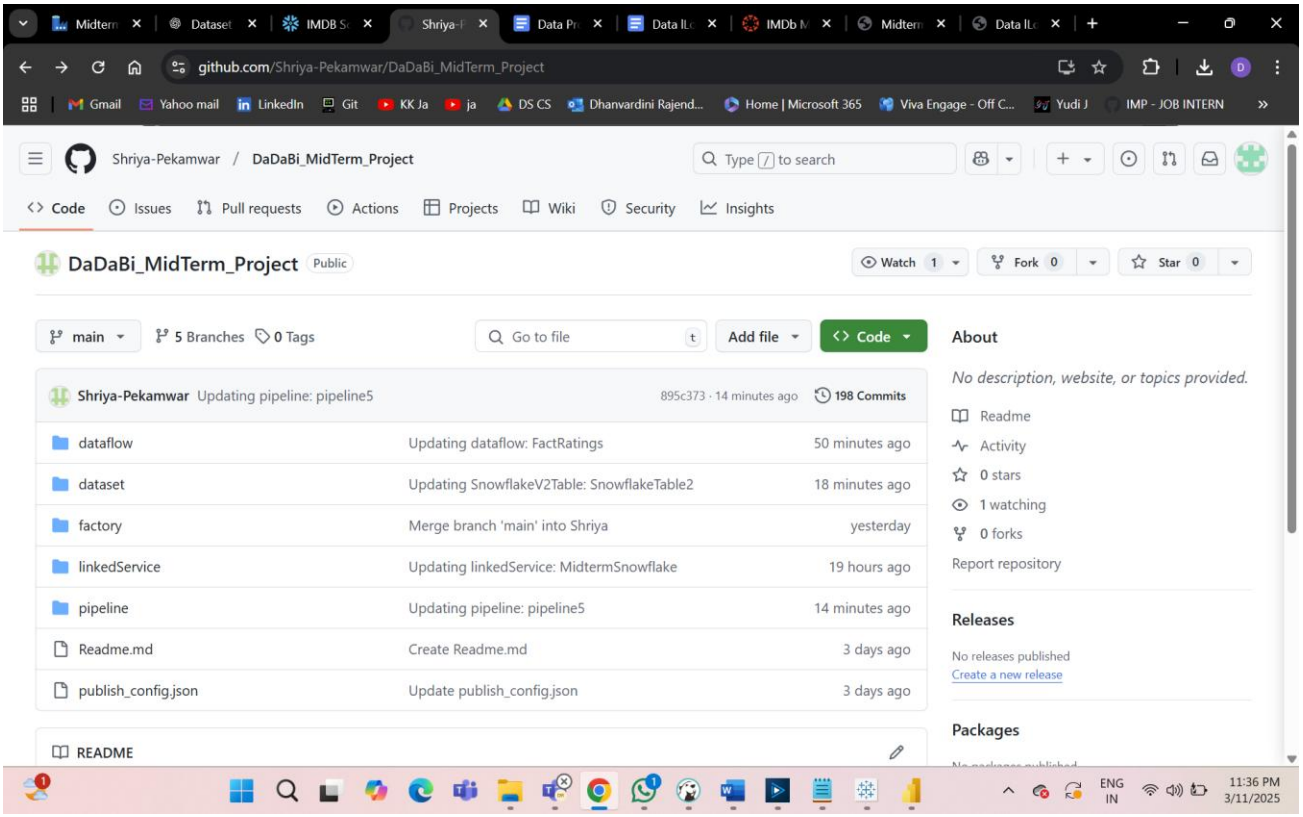


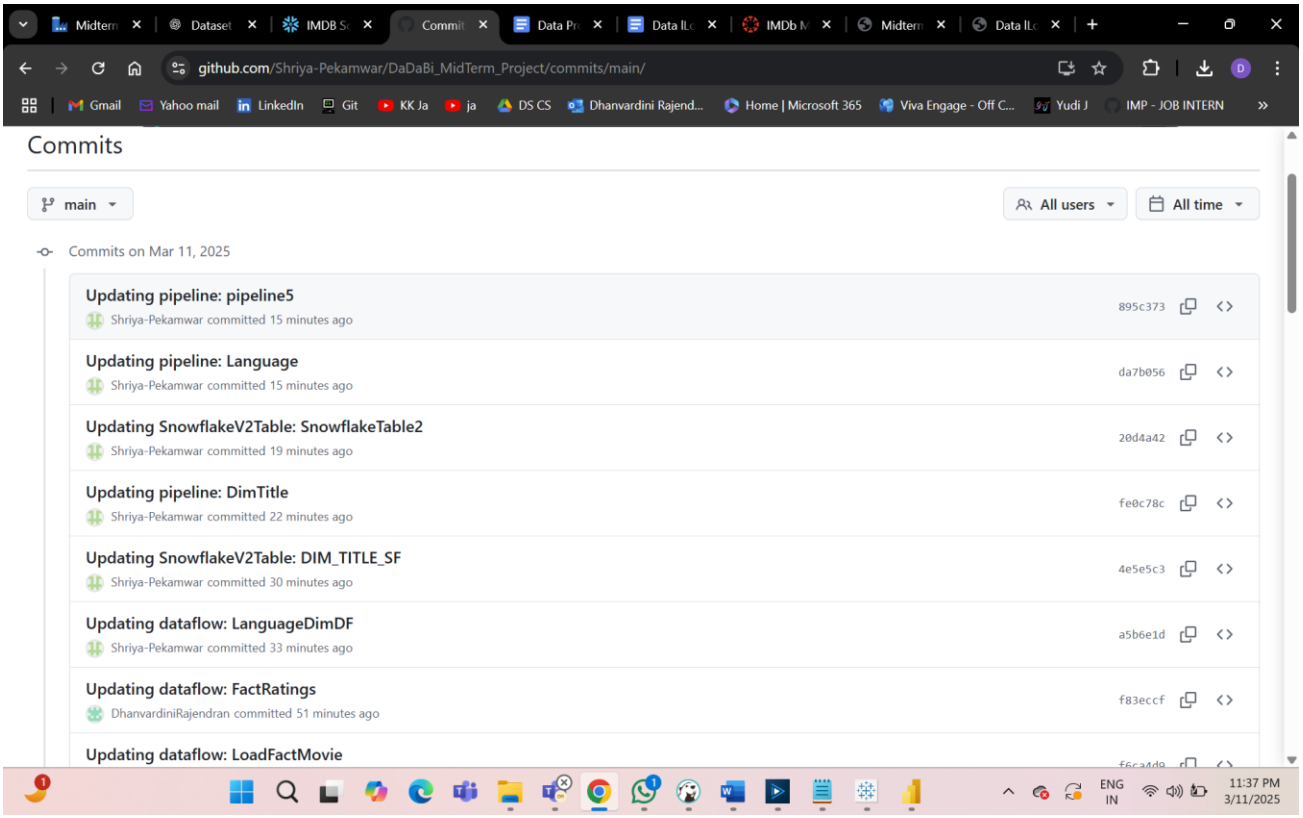
Loading Fact_Movie into Power BI



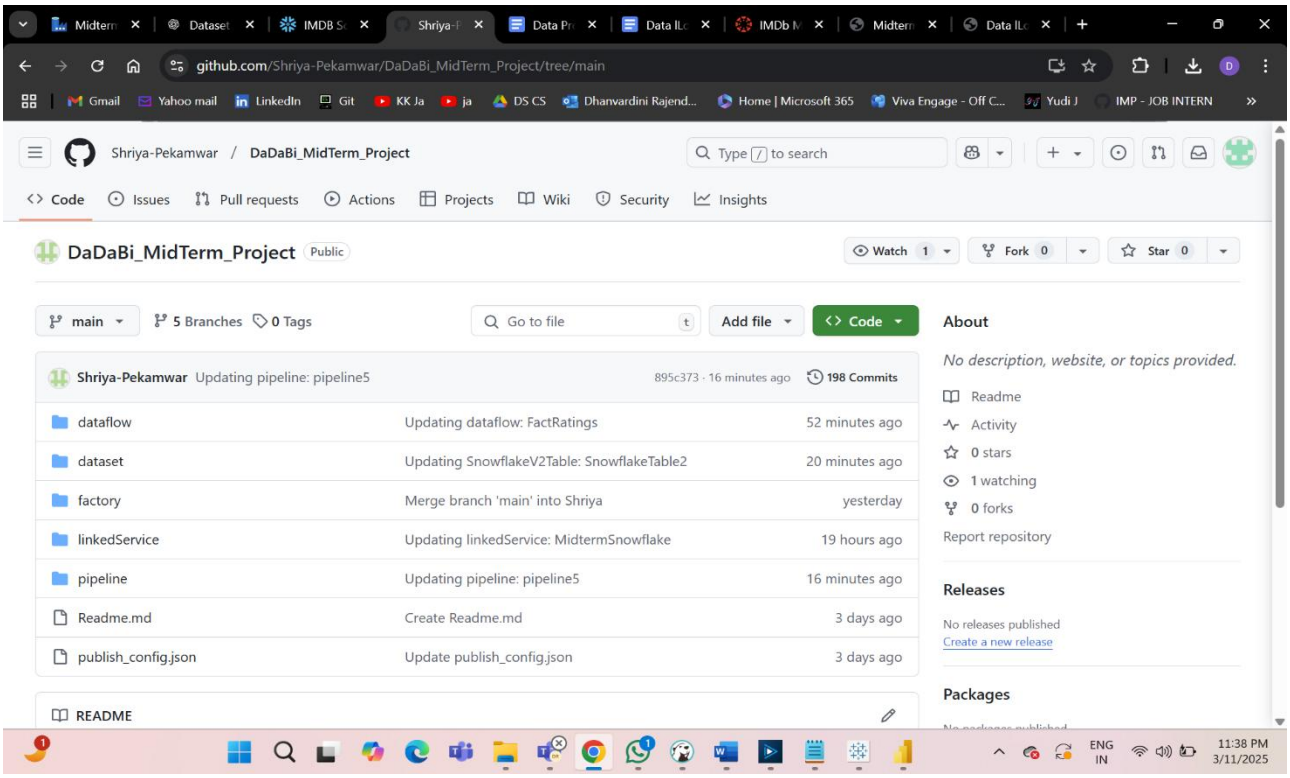
GIT SCREENSHOTS

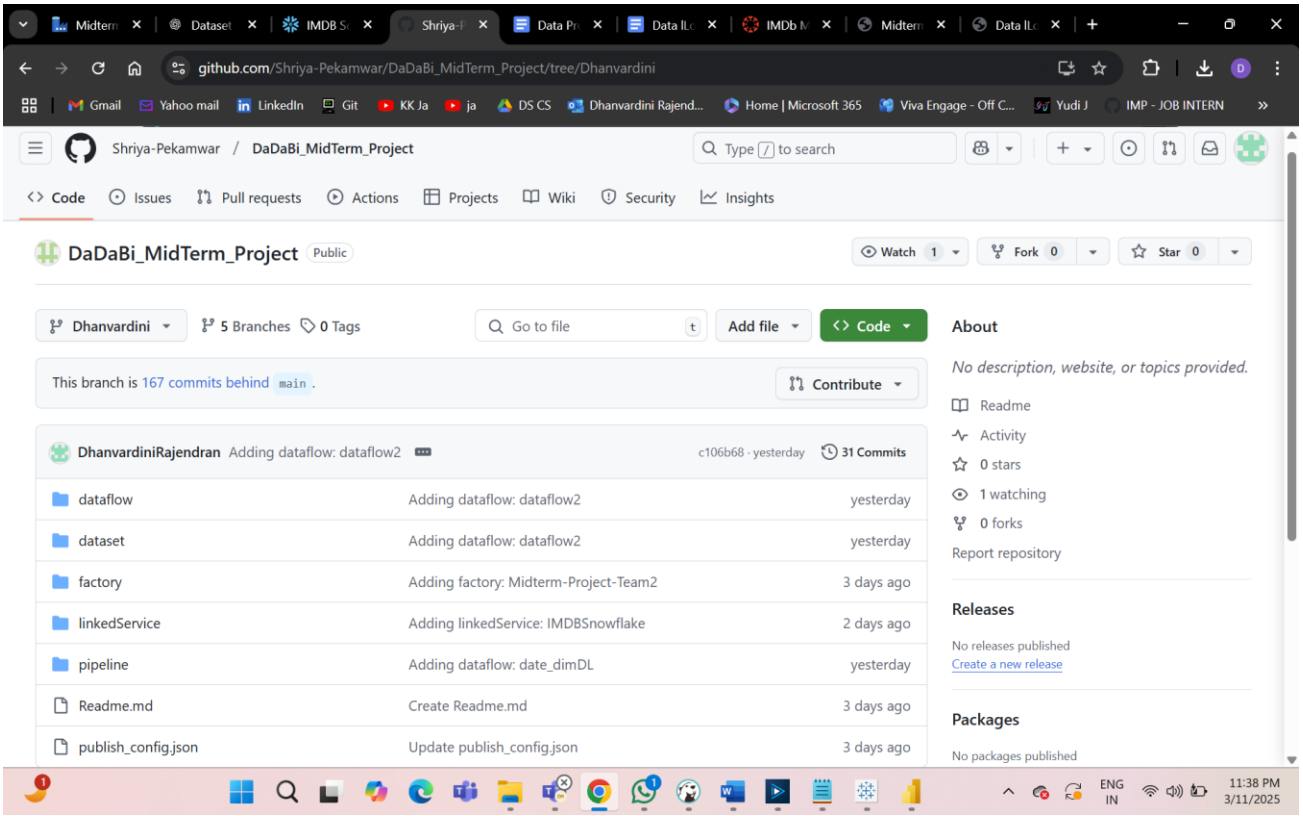
COMMITS:

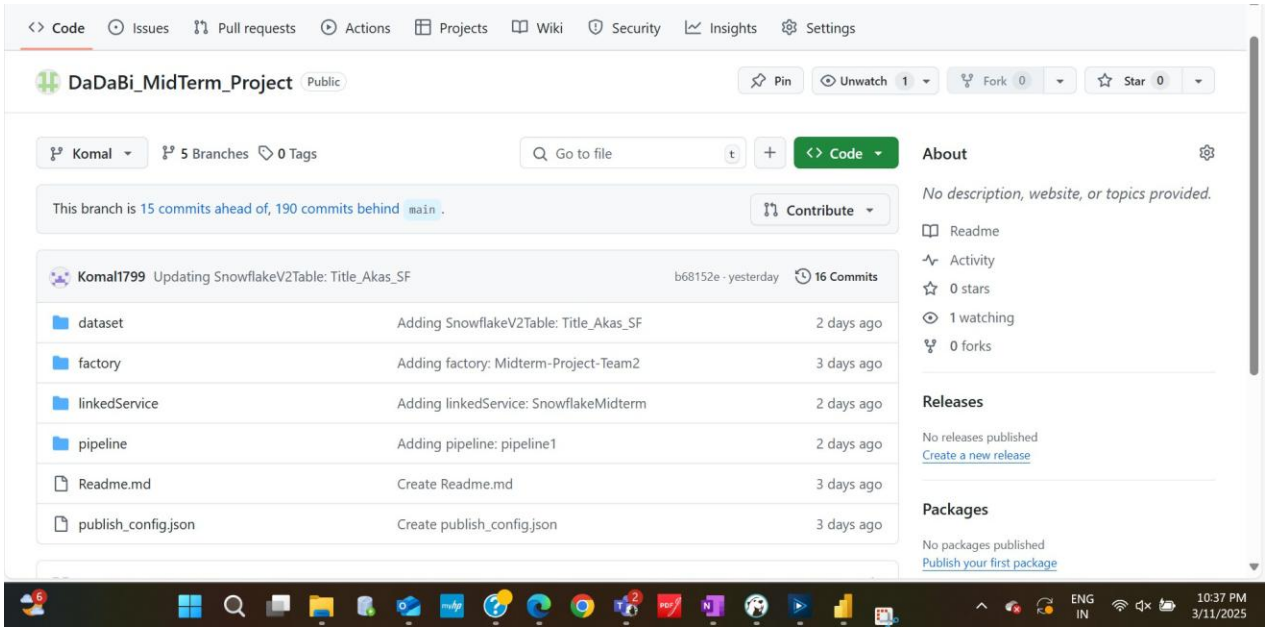

github.com/Shriya-Pekamwar/DaDaBi_MidTerm_Project
Shriya-Pekamwar / DaDaBi_MidTerm_Project
Code Issues Pull requests Actions Projects Wiki Security Insights
DaDaBi_MidTerm_Project Public
Watch 1 Fork 0 Star 0
main 5 Branches 0 Tags
Go to file Add file Code
Shriya-Pekamwar Updating pipeline: pipeline5 895c373 · 14 minutes ago 198 Commits
dataflow Updating dataflow: FactRatings 50 minutes ago
dataset Updating SnowflakeV2Table: SnowflakeTable2 18 minutes ago
factory Merge branch 'main' into Shriya yesterday
linkedService Updating linkedService: MidtermSnowflake 19 hours ago
pipeline Updating pipeline: pipeline5 14 minutes ago
Readme.md Create Readme.md 3 days ago
publish_config.json Update publish_config.json 3 days ago
README
About
No description, website, or topics provided.
Readme
Activity
0 stars
1 watching
0 forks
Report repository
Releases
No releases published
Create a new release
Packages

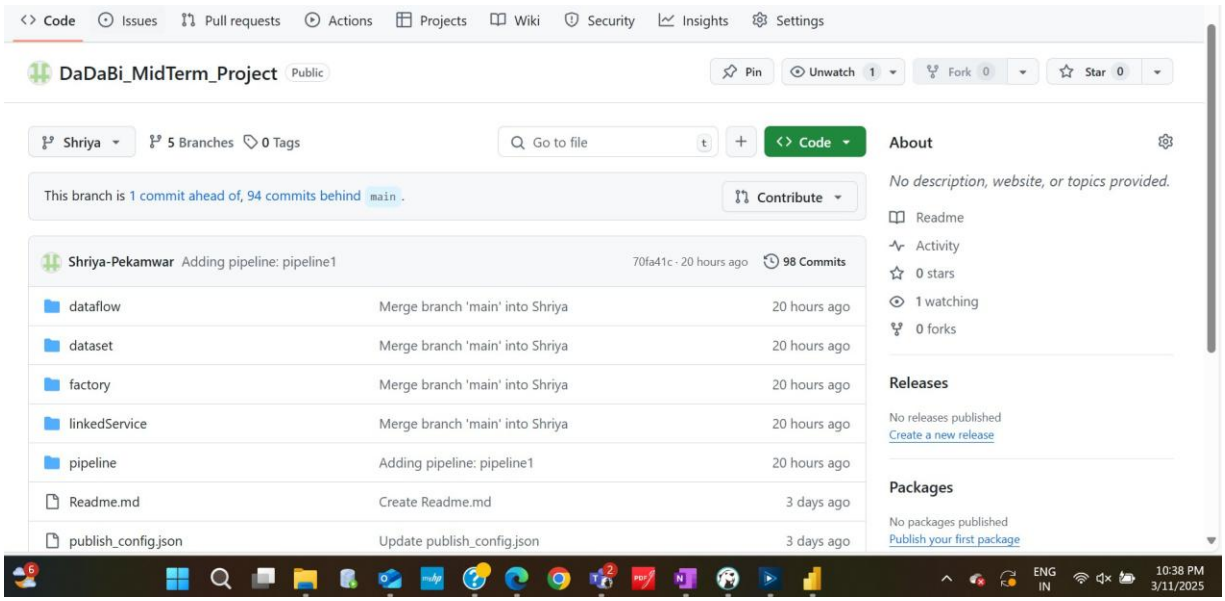

github.com/Shriya-Pekamwar/DaDaBi_MidTerm_Project/commits/main/
Commits
main
All users
All time
Commits on Mar 11, 2025
Updating pipeline: pipeline5
Shriya-Pekamwar committed 15 minutes ago
895c373
Updating pipeline: Language
Shriya-Pekamwar committed 15 minutes ago
da7b056
Updating SnowflakeV2Table: SnowflakeTable2
Shriya-Pekamwar committed 19 minutes ago
20d4a42
Updating pipeline: DimTitle
Shriya-Pekamwar committed 22 minutes ago
fe0c78c
Updating SnowflakeV2Table: DIM_TITLE_SF
Shriya-Pekamwar committed 30 minutes ago
4e5e5c3
Updating dataflow: LanguageDimDF
Shriya-Pekamwar committed 33 minutes ago
a5b6e1d
Updating dataflow: FactRatings
DhanvardiniRajendran committed 51 minutes ago
f83eccf
Updating dataflow: LoadFactMovie

Respective Git Contribution






Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
DaDaBi_MidTerm_Project
Public
Pin
Unwatch 1
Fork 0
Star 0
Komal
5 Branches
0 Tags
Go to file
t
Code
About
This branch is 15 commits ahead of, 190 commits behind main.
Contribute
No description, website, or topics provided.
Komal1799 Updating SnowflakeV2Table: Title_Akas_SF
b68152e · yesterday
16 Commits
Readme
Activity
0 stars
1 watching
0 forks
dataset
Adding SnowflakeV2Table: Title_Akas_SF
2 days ago
factory
Adding factory: Midterm-Project-Team2
3 days ago
linkedService
Adding linkedService: SnowflakeMidterm
2 days ago
pipeline
Adding pipeline: pipeline1
2 days ago
Readme.md
Create Readme.md
3 days ago
publish_config.json
Create publish_config.json
3 days ago
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package
ENG
IN
10:37 PM
3/11/2025


Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
DaDaBi_MidTerm_Project
Public
Pin
Unwatch 1
Fork 0
Star 0
Shriya
5 Branches
0 Tags
Go to file
t
Code
About
This branch is 1 commit ahead of, 94 commits behind main.
Contribute
No description, website, or topics provided.
Shriya-Pekamwar Adding pipeline: pipeline1
70fa41c · 20 hours ago
98 Commits
Readme
Activity
0 stars
1 watching
0 forks
dataflow
Merge branch 'main' into Shriya
20 hours ago
dataset
Merge branch 'main' into Shriya
20 hours ago
factory
Merge branch 'main' into Shriya
20 hours ago
linkedService
Merge branch 'main' into Shriya
20 hours ago
pipeline
Adding pipeline: pipeline1
20 hours ago
Readme.md
Create Readme.md
3 days ago
publish_config.json
Update publish_config.json
3 days ago
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package
ENG
IN
10:38 PM
3/11/2025