Individual Project 1 : NYPD Arrest Data

DIMENSIONAL MODELING:

1. Identify the Grain

The **grain** defines the level of detail in the fact table.

Grain of the Fact Table (FACT_ARRESTS):
Each record in the fact table represents **a single arrest event** with a unique ARREST_KEY. This ensures that the data is stored at the most granular level—each arrest occurrence.

2. Define Facts and Dimensions

Fact Table: FACT_ARRESTS

This table records every arrest event, which serves as the primary measure for analysis.

Column Name	Data Type	Description
ARREST_KEY	BIGINT	Unique identifier for each arrest (Primary Key)
ARREST_DATE	DATE	Date of the arrest
PRECINCT_ID	INT	Foreign key referencing DIM_PRECINCT
BOROUGH_ID	INT	Foreign key referencing DIM_BOROUGH
LOCATION_ID	INT	Foreign key referencing DIM_LOCATION
OFFENSE_ID	INT	Foreign key referencing DIM_OFFENSE
LAW_ID	VARCHAR(20)	Foreign key referencing DIM_LAW
PERP_ID	INT	Foreign key referencing DIM_PERPETRATOR
JURISDICTION_CODE	INT	Code representing the jurisdiction of arrest

EVENT-BASED MEASURE:

1. ARREST_DATE 2. JURISDICTION_CODE

Dimension Tables:

1. DIM_PRECINCT

Represents police precincts in NYC.

Column Name	Data Type	Description
PRECINCT_ID	INT (PK)	Unique precinct ID

Column Name	Data Type	Description
ARREST_PRECINCT	INT	Precinct number

2. DIM_BOROUGH

Stores NYC borough details.

Column Name	Data Type	Description
BOROUGH_ID	INT (PK)	Unique borough ID
ARREST_BORO	VARCHAR(5)	Borough name abbreviation (M, B, Q, S, K)

3. DIM_LOCATION

Stores the geographic data. **SCD Type 2** ensures history tracking for location changes.

Column Name	Data Type	Description
LOCATION_ID	INT (PK)	Unique location ID
X_COORD_CD	FLOAT	X Coordinate
Y_COORD_CD	FLOAT	Y Coordinate
Latitude	FLOAT	Latitude
Longitude	FLOAT	Longitude

4. DIM_OFFENSE

Stores details about offenses.

Column Name	Data Type	Description
OFFENSE_ID	INT (PK)	Unique offense ID
PD_CD	INT	Internal offense code
PD_DESC	VARCHAR(255)	Internal offense description
KY_CD	INT	Official offense code
OFNS_DESC	VARCHAR(255)	Official offense description

5. DIM_LAW

Stores law-related details. **SCD Type 1** ensures only current values are stored.

Column Name	Data Type	Description
LAW_ID	VARCHAR(20) (PK)	Unique law identifier
LAW_CODE	VARCHAR(20)	Law code
LAW_CAT_CD	VARCHAR(5)	Category of law (Felony, Misdemeanor, Violation)

6. DIM_PERPETRATOR

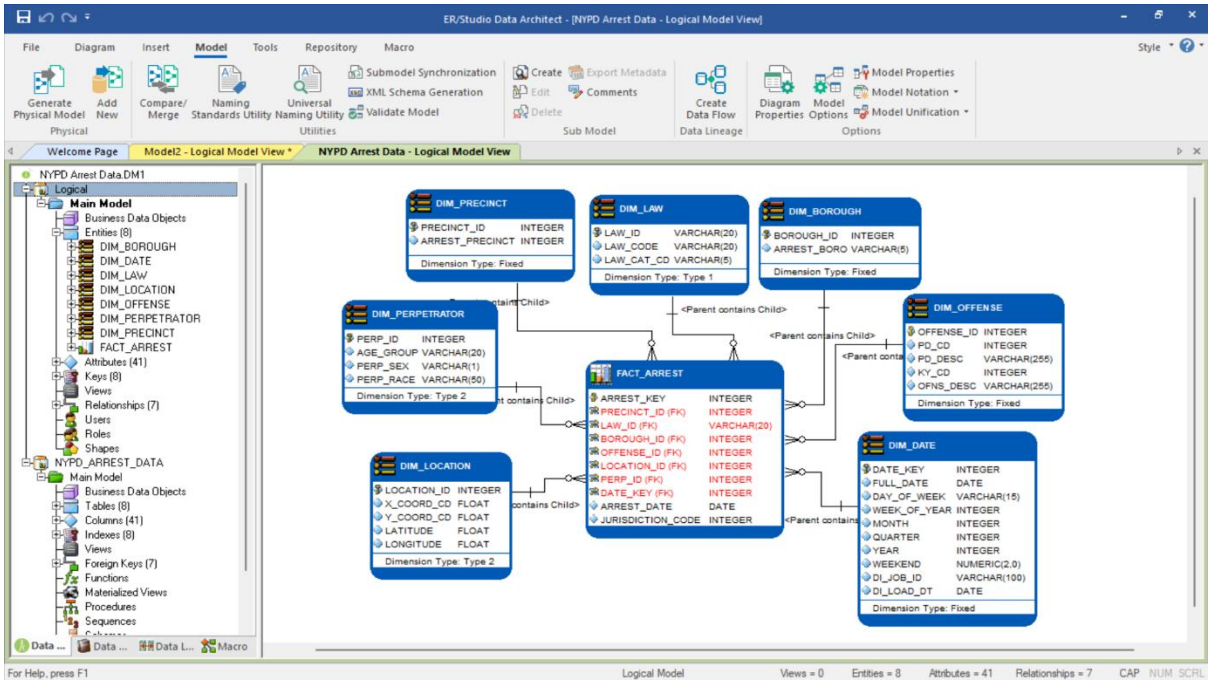
Stores perpetrator demographics. **SCD Type 2** ensures history tracking.

Column Name	Data Type	Description
PERP_ID	INT (PK)	Unique perpetrator ID
AGE_GROUP	VARCHAR(20)	Age group (e.g., <18, 18-24, 25-44, etc.)
PERP_SEX	VARCHAR(1)	Gender (M/F)
PERP_RACE	VARCHAR(50)	Race classification

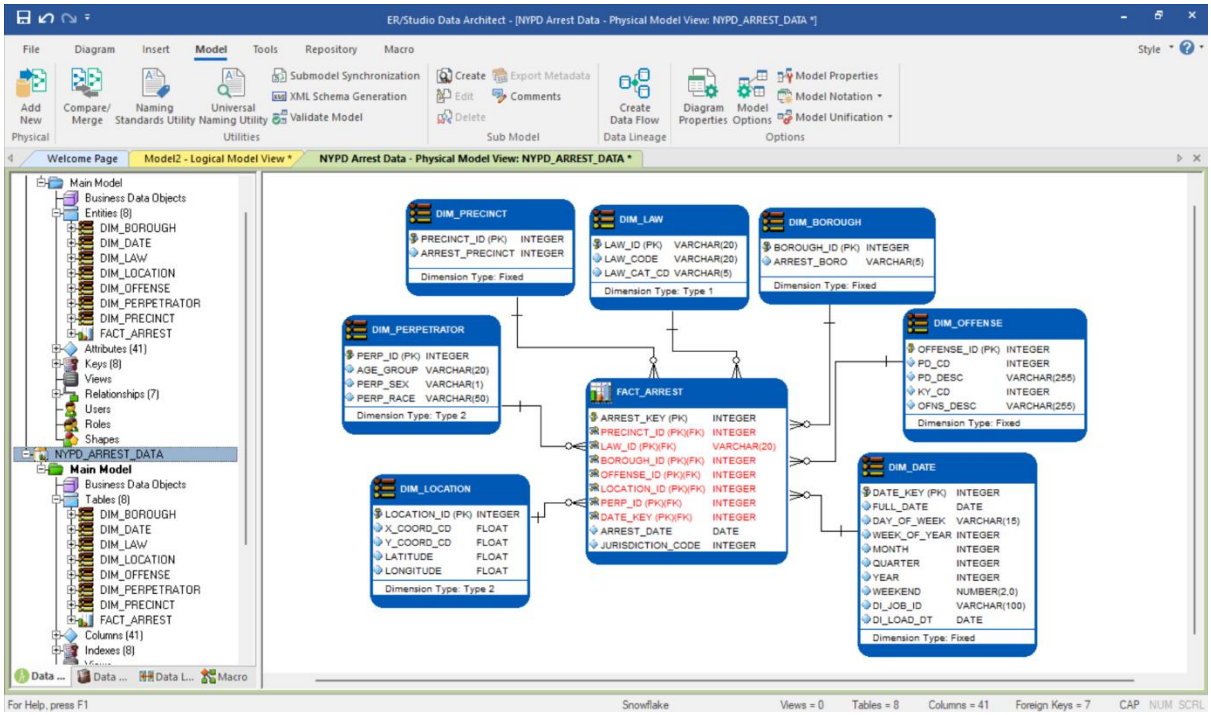
3. Business Requirements

- **How many arrests occurred on any specific day, week, month, quarter, or year?**
  Query FACT_ARRESTS using ARREST_DATE.
- **What are the peak days and months for arrests?**
  Aggregate FACT_ARRESTS by ARREST_DATE.
- **What are the top 5 most frequently occurring crimes?**
  Join FACT_ARRESTS with DIM_OFFENSE and count occurrences of OFNS_DESC.
- **Which crimes have increased or decreased the most over time?**
  Compare FACT_ARRESTS year-over-year using ARREST_DATE.
- **Are there specific precincts with higher felony arrests than misdemeanors?**
  Join FACT_ARRESTS with DIM_LAW and group by PRECINCT_ID.
- **Which borough has the highest number of arrests?**
  Aggregate FACT_ARRESTS by BOROUGH_ID.
- **What is the distribution of arrestees by age, race, and gender?**
  Join FACT_ARRESTS with DIM_PERPETRATOR and group by AGE_GROUP, PERP_SEX, PERP_RACE.
- **Can we predict high-crime areas based on past arrest data?**
  Use FACT_ARRESTS with DIM_LOCATION to identify hotspots.

LOGICAL MODEL:



PHYSICAL MODEL:

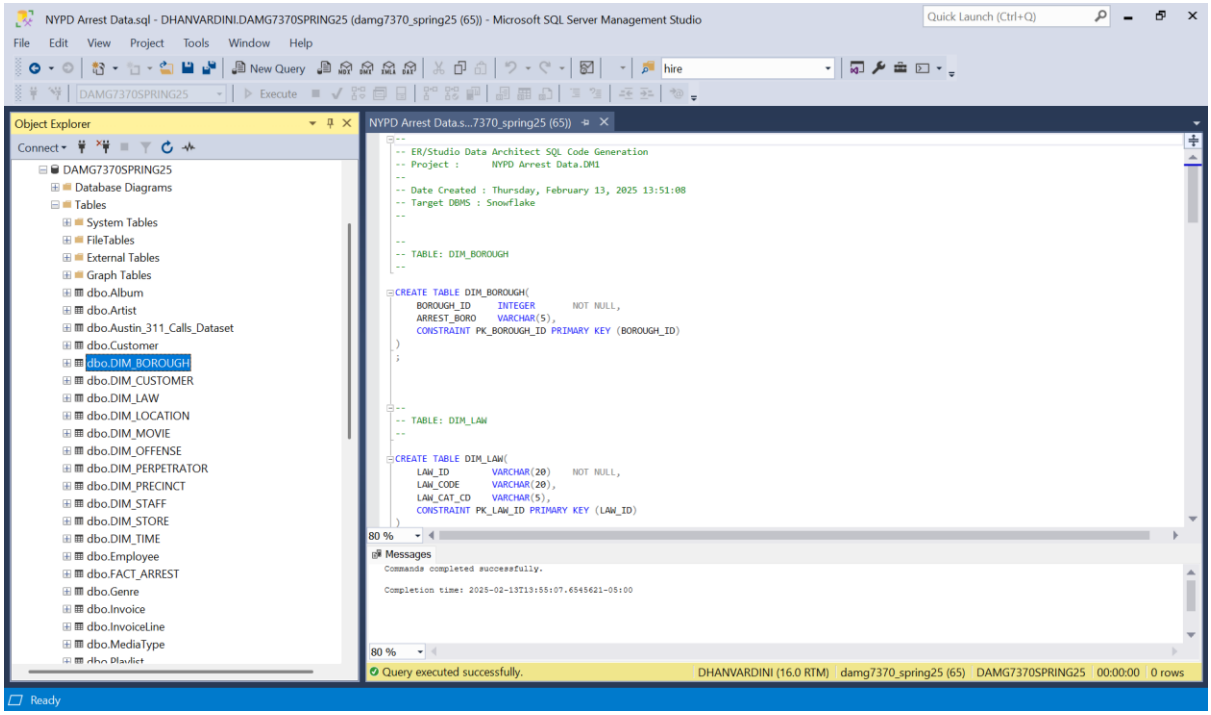


ER STUDIO FILE & SQL SCRIPT:

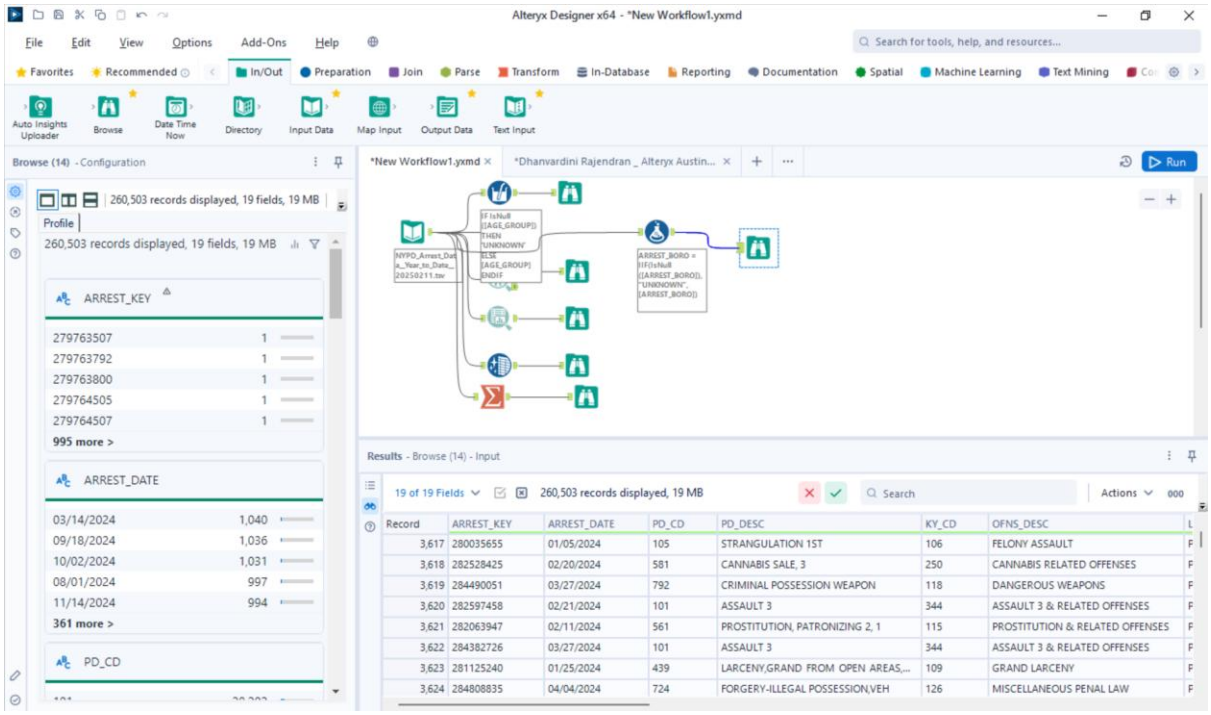
NYPD Arrest Data.DM1

NYPD_ARREST_DATA.sql

TABLES CREATED IN SSMS:



ALTERYX:



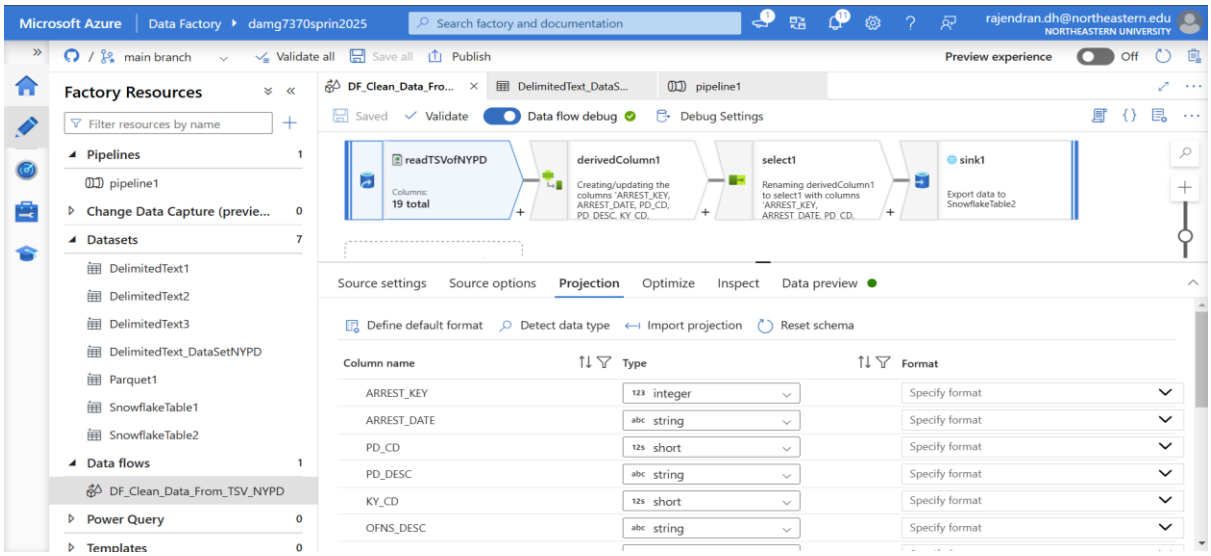
Observations:

1. **Missing Values:**
    - PD_CD: 8 missing values.

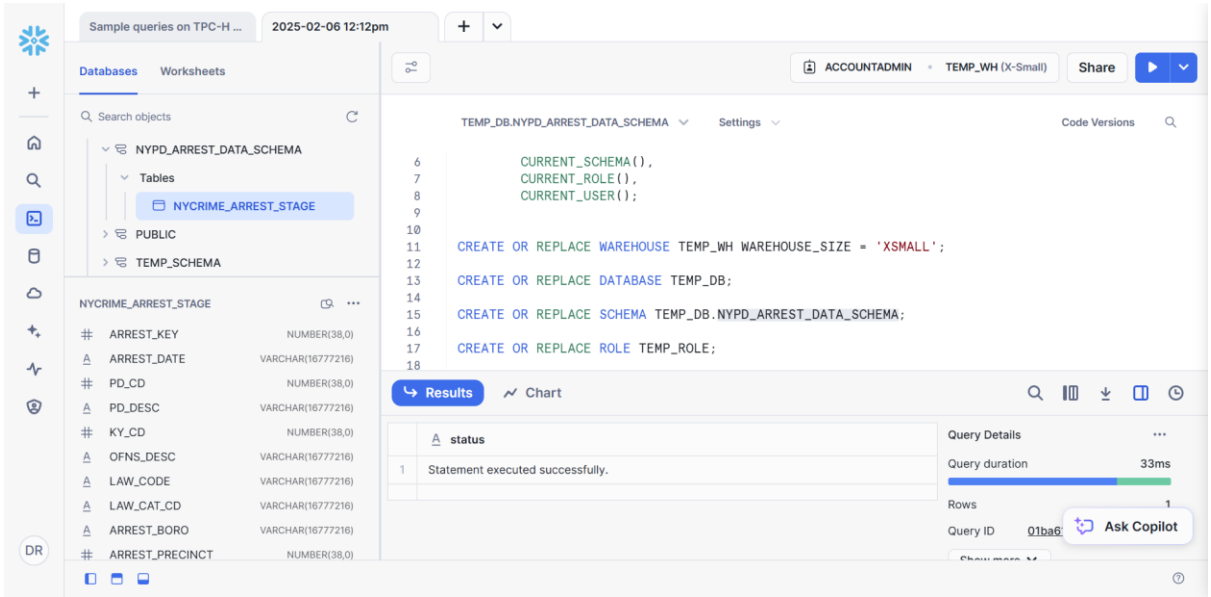
- KY_CD: 32 missing values.
- LAW_CAT_CD: 1,390 missing values.
- Latitude, Longitude, and New Georeferenced Column: 4 missing values.

2. **Inconsistencies:**
   - ARREST_DATE is stored as an object (string) instead of a date format.
   - LAW_CAT_CD is missing for some records, which might indicate incomplete categorization.
   - Some AGE_GROUP categories might be inconsistent or need standardization.
   - ARREST_BORO uses single-letter codes that might need mapping for clarity.

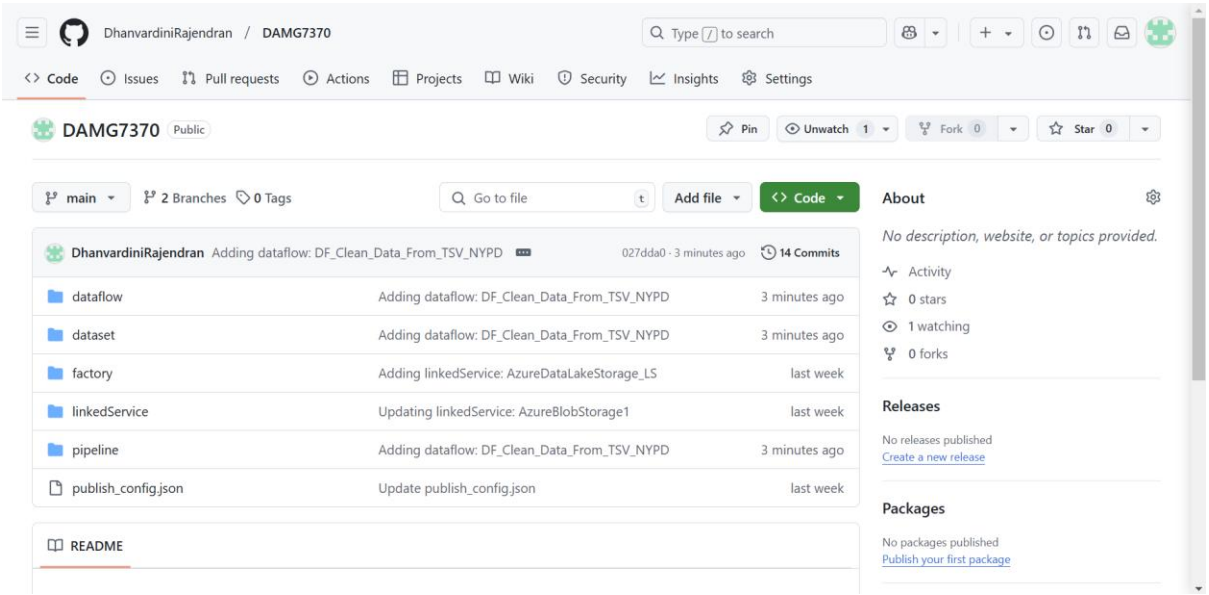
ADF PIPELINE:



SNOWFLAKE:



GITHUB:


DhanvardiniRajendran / DAMG7370
Type / to search
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
DAMG7370
Public
Pin
Unwatch 1
Fork 0
Star 0
main
2 Branches
0 Tags
Go to file
t
Add file
Code
DhanvardiniRajendran Adding dataflow: DF_Clean_Data_From_TSV_NYPD
027dda0 · 3 minutes ago
14 Commits
dataflow
Adding dataflow: DF_Clean_Data_From_TSV_NYPD
3 minutes ago
dataset
Adding dataflow: DF_Clean_Data_From_TSV_NYPD
3 minutes ago
factory
Adding linkedService: AzureDataLakeStorage_LS
last week
linkedService
Updating linkedService: AzureBlobStorage1
last week
pipeline
Adding dataflow: DF_Clean_Data_From_TSV_NYPD
3 minutes ago
publish_config.json
Update publish_config.json
last week
README
About
No description, website, or topics provided.
Activity
0 stars
1 watching
0 forks
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package