VARANA NAVADIYA-002308207

Individual Project 1: NYPD Arrest Data

In this project, Alteryx and Azure Data Factory (ADF) will be used to create a Dimensional Model for NYPD arrest data analysis and to develop a Data Staging Pipeline. Supporting analytical inquiries about NYC crime trends is the main objective to assist the public and city officials in comprehending arrest trends.

-Business Requirements:

1. How many arrests occurred on any specific day, week, month, quarter, or year?

   Relevant - The DIM_DATE table provides DAY_NUM, WEEK_NUM, MONTH_NUM, QTR_NUM, and YEAR_NUM for time-based analysis.

   QUERY:

   ```
   SELECT YEAR_NUM, MONTH_NUM, WEEK_NUM, DAY_NUM, COUNT(*) AS total_arrests

   FROM CRIME_FACT CF

   JOIN DIM_DATE DD ON CF.DATE_SK = DD.DATE_SK

   GROUP BY YEAR_NUM, MONTH_NUM, WEEK_NUM, DAY_NUM

   ORDER BY YEAR_NUM DESC, MONTH_NUM DESC, WEEK_NUM DESC, DAY_NUM DESC;
   ```

2. What are the peak days and months for arrests?

   Relevant - DIM_DATE allows us to identify the most common days and months.

   QUERY:

   ```
   SELECT MONTH_NUM, COUNT(*) AS total_arrests

   FROM CRIME_FACT CF

   JOIN DIM_DATE DD ON CF.DATE_SK = DD.DATE_SK

   GROUP BY MONTH_NUM

   ORDER BY total_arrests DESC
   ```

LIMIT 5;

To analyze peak days:

SELECT DAY_NUM, COUNT(*) AS total_arrests FROM CRIME_FACT CF JOIN DIM_DATE DD ON CF.DATE_SK = DD.DATE_SK GROUP BY DAY_NUM ORDER BY total_arrests DESC LIMIT 5;

3. What are the top 5 most frequently occurring crimes?

   Relevant - The DIM_CRIMES table contains PD_DESC (precinct description) and OFFENSE_CAT (offense category).

   QUERY:

   SELECT PD_DESC, COUNT(*) AS total_occurrences

   FROM CRIME_FACT CF

   JOIN DIM_CRIMES DC ON CF.CRIMES_SK = DC.CRIMES_SK

   GROUP BY PD_DESC

   ORDER BY total_occurrences DESC

   LIMIT 5;

4. Which crimes have increased or decreased the most over time?

   Relevant - The DIM_DATE table allows us to analyze trends over time.

   QUERY:

   SELECT YEAR_NUM, PD_DESC, COUNT(*) AS total_arrests

   FROM CRIME_FACT CF

   JOIN DIM_CRIMES DC ON CF.CRIMES_SK = DC.CRIMES_SK

   JOIN DIM_DATE DD ON CF.DATE_SK = DD.DATE_SK

   GROUP BY YEAR_NUM, PD_DESC

   ORDER BY YEAR_NUM DESC, total_arrests DESC;

5. Are there specific precincts with higher felony arrests compared to misdemeanors? (Hint: A precinct is a police district within a city.)

   Relevant - The DIM_ARRESTS table contains ARREST_PRECINCT, and DIM_CRIMES has LAW_CATEGORY (which can indicate felony vs. misdemeanor).

   QUERY:

```
SELECT DA.ARREST_PRECINCT,
    DC.LAW_CATEGORY,
    COUNT(*) AS total_arrests
FROM CRIME_FACT CF
JOIN DIM_ARRESTS DA ON CF.ARRESTS_SK = DA.ARRESTS_SK
JOIN DIM_CRIMES DC ON CF.CRIMES_SK = DC.CRIMES_SK
GROUP BY DA.ARREST_PRECINCT, DC.LAW_CATEGORY
ORDER BY DA.ARREST_PRECINCT, total_arrests DESC;
```

6. Which borough has the highest number of arrests? (Hint: A borough is a large administrative division in NYC, such as Manhattan (M), Brooklyn (K), Queens (Q), The Bronx (B), and Staten Island (S).)

   Relevant - DIM_ARRESTS contains ARREST_BOROUGH.

   QUERY:

```
SELECT ARREST_BOROUGH, COUNT(*) AS total_arrests
FROM CRIME_FACT CF
JOIN DIM_ARRESTS DA ON CF.ARRESTS_SK = DA.ARRESTS_SK
GROUP BY ARREST_BOROUGH
ORDER BY total_arrests DESC;
```

7. What is the distribution of arrestees by age, race, and gender?

   Relevant - The DIM_CRIMINAL table contains AGE, SEX, and RACE.

   QUERY:

```
SELECT AGE, SEX, RACE, COUNT(*) AS total_arrests
FROM CRIME_FACT CF
JOIN DIM_CRIMINAL DC ON CF.CRIMINAL_SK = DC.CRIMINAL_SK
GROUP BY AGE, SEX, RACE
ORDER BY total_arrests DESC;
```

8. Can we predict high-crime areas based on past arrest data?

   Partially Relevant - DIM_LOCATION contains LATITUDE, LONGITUDE, and GEOREFERENCE, which help identify crime hotspots. However, this dataset does not include crime severity, time of day, or socioeconomic factors necessary for a full predictive analysis.

```
SELECT DL.LATITUDE, DL.LONGITUDE, COUNT(*) AS total_arrests
FROM CRIME_FACT CF
JOIN DIM_LOCATION DL ON CF.LOCATION_SK = DL.LOCATION_SK
GROUP BY DL.LATITUDE, DL.LONGITUDE
ORDER BY total_arrests DESC
LIMIT 10;
```

-Identifying the Grain:

Each record in the dataset represents a single arrest event, which is the grain of our data.

Dimension Tables:

- DIM_DATE – Stores date-related information (day, month, quarter, year).
- DIM_LOCATION – Stores geographic data (latitude, longitude, borough, precinct).

- DIM_CRIMES – Stores crime-related details (crime type, law code, offense category).
- DIM_ARRESTS – Stores arrest-related attributes (arrest ID, precinct, borough, jurisdiction).
- DIM_CRIMINAL – Stores information about arrestees (age, race, gender).

-Data Profiling Using Alteryx:

The NYPD Arrest dataset was profiled using Alteryx Designer, revealing several data quality issues:

- Missing Values:
    - PD_CD: 8 records are null
    - KY_CD: 32 records are null
    - LAW_CAT_CD: 1390 records are null
    - LATITUDE: 4 records are null
    - LONGITUDE: 4 records are null
    - NEW GEOREFERENCED COLUMN: 4 records are null
- Data Formatting Issues:
    - Date format is MMDDYYYY, but should be converted to YYYYMMDD.
    - Data types need standardization for PD_CD, KY_CD, ARREST_PRECINCT, JURISDICTION_CODE, LATITUDE, and LONGITUDE.

Data Cleansing Approach:

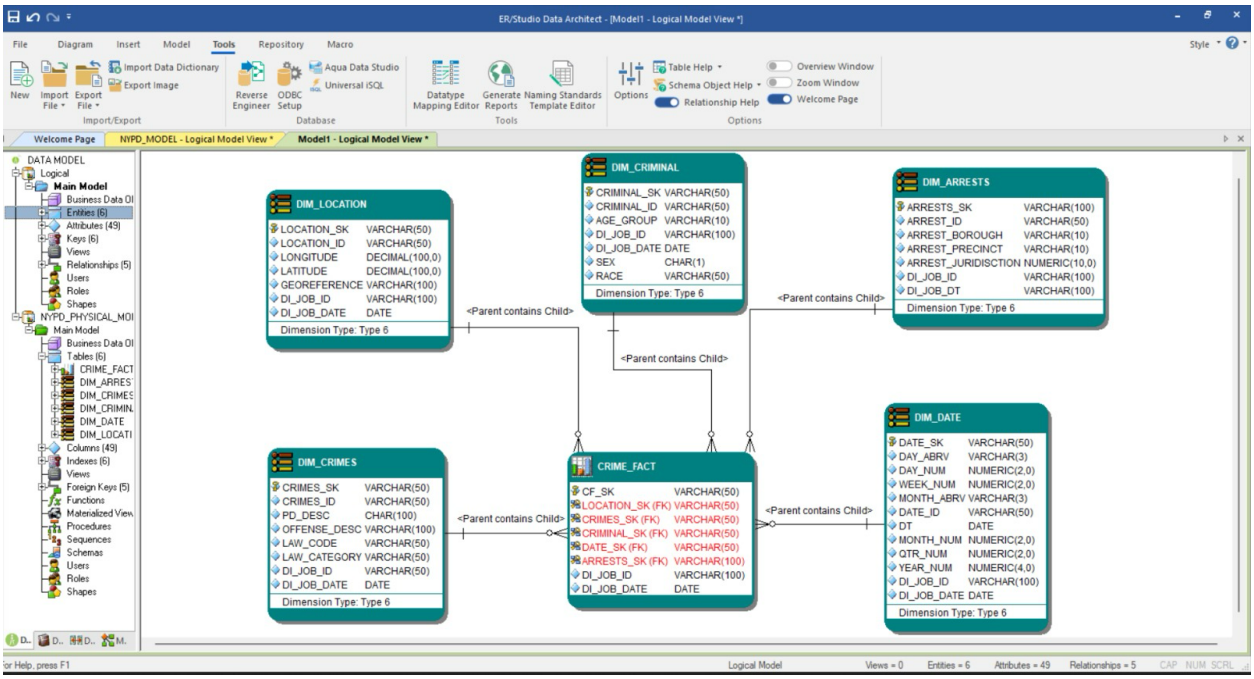
- Replace null values with appropriate placeholders or remove invalid records.
- Convert date format for consistency.
- Standardize data types to maintain integrity in Snowflake.

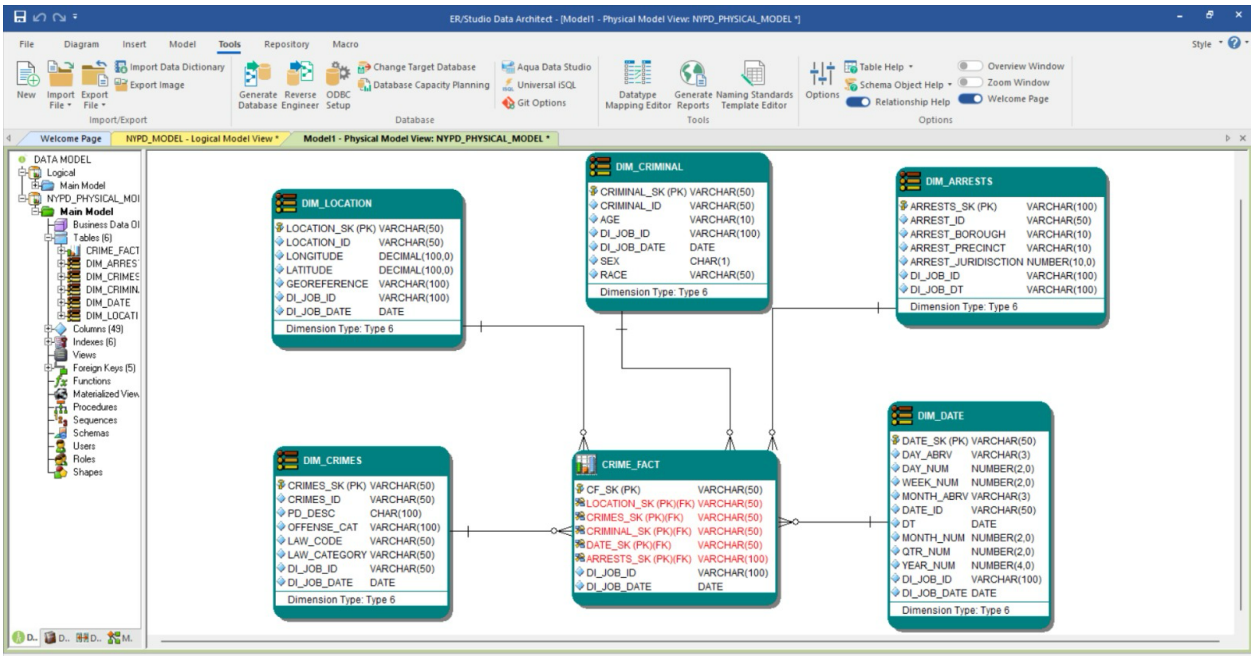
Transformations:

- NULLCLEAR: Handle missing values.
- DateFormat: Convert date format to YYYYMMDD.
- Standardization: Ensure data types match schema requirements.

VARANA NAVADIYA-002308207

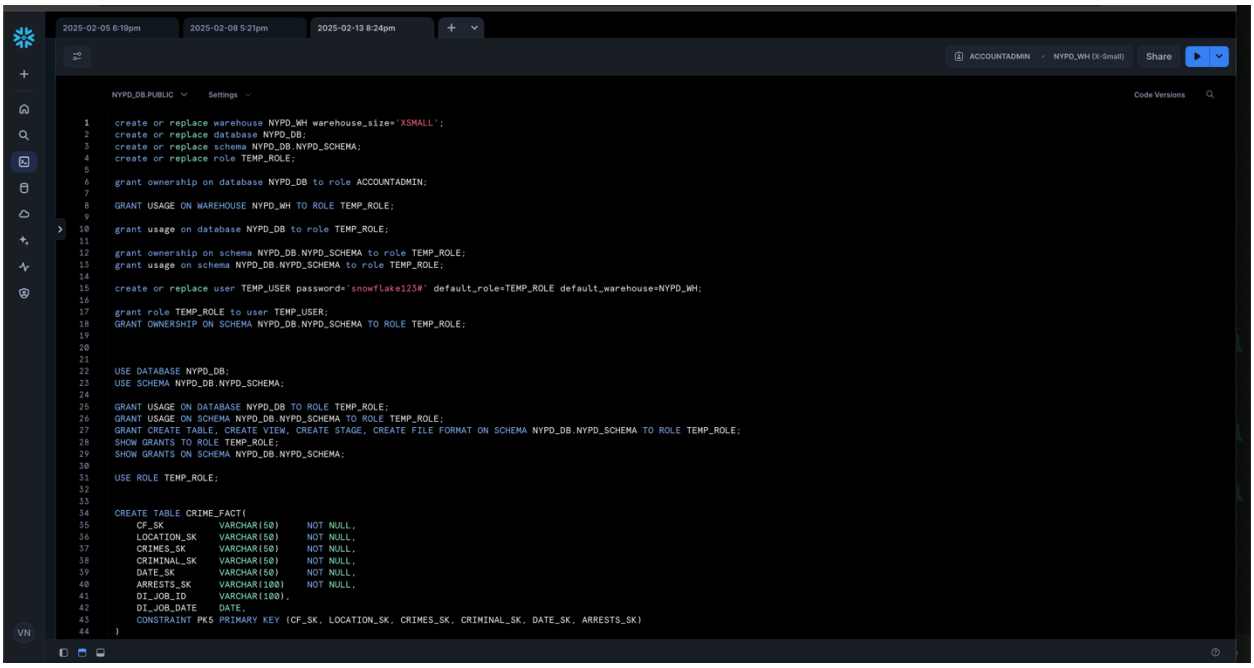
-ER/STUDIO:



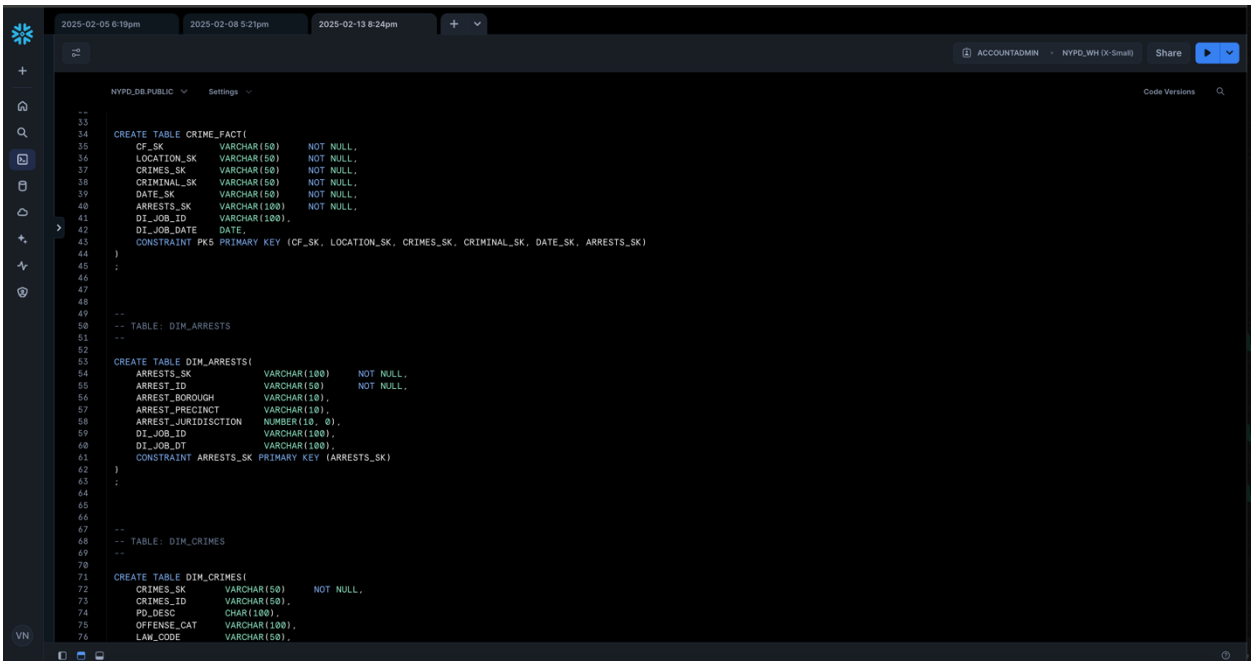


VARANA NAVADIYA-002308207

-SQL SCRIPT:

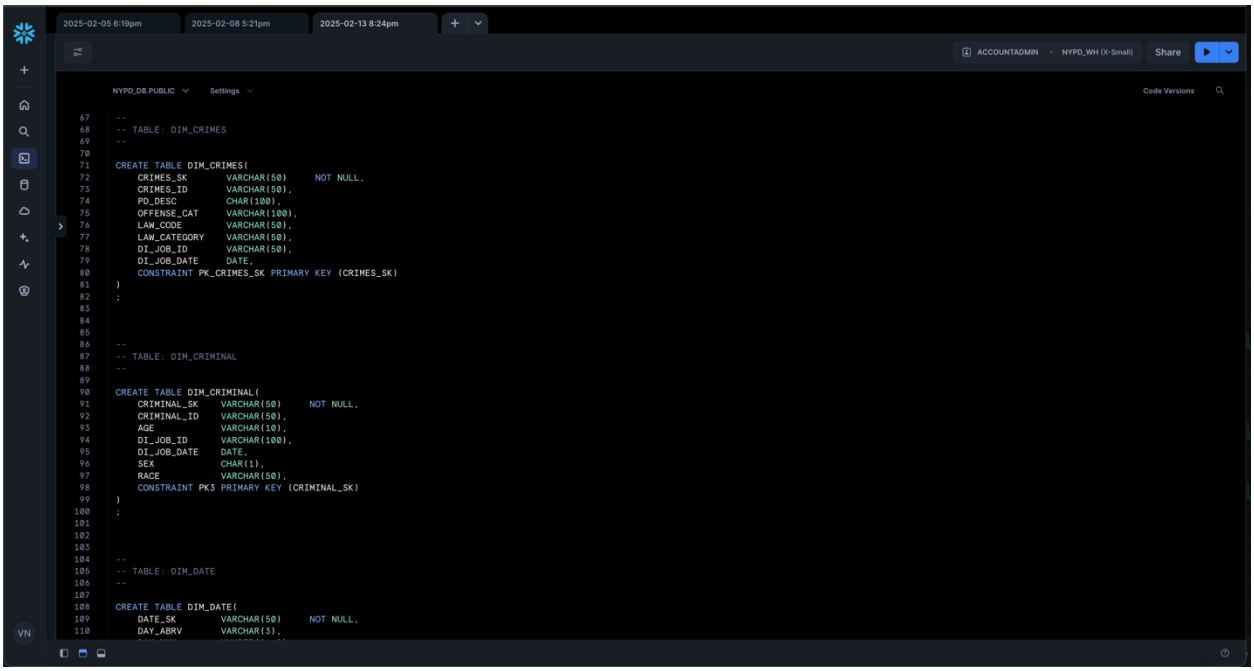


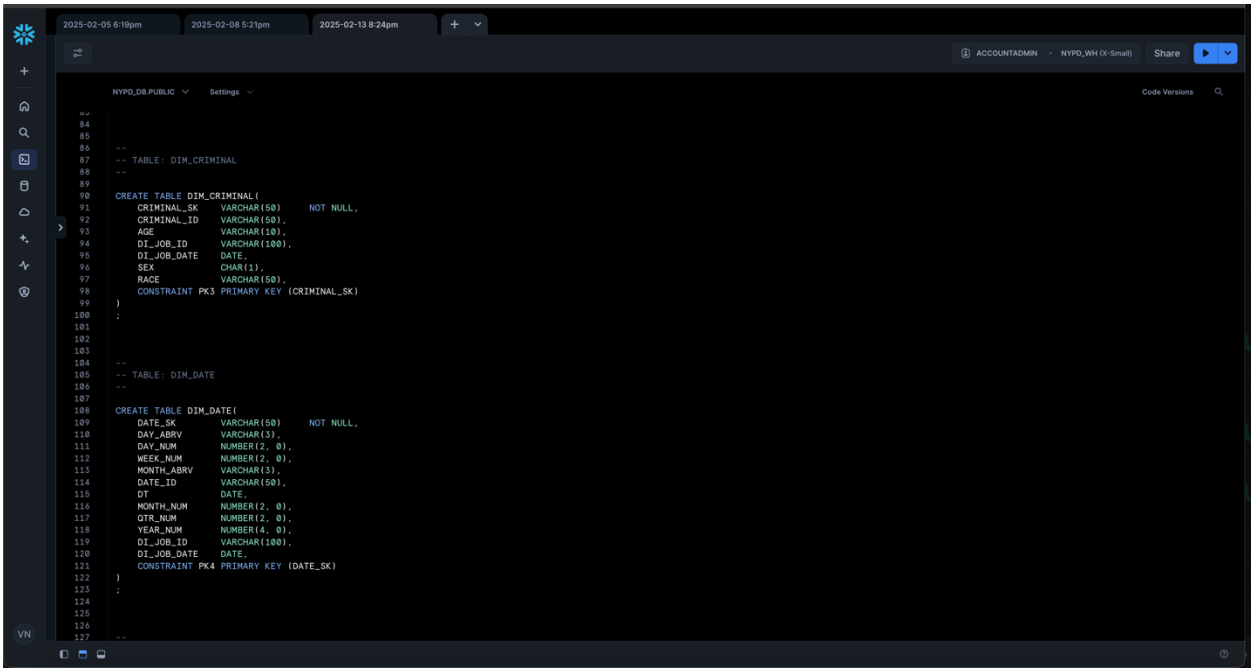
```
create or replace warehouse NYPD_WH warehouse_size='XSMALL';
create or replace database NYPD_DB;
create or replace schema NYPD_DB.NYPD_SCHEMA;
create or replace role TEMP_ROLE;

grant ownership on database NYPD_DB to role ACCOUNTADMIN;

GRANT USAGE ON WAREHOUSE NYPD_WH TO ROLE TEMP_ROLE;

grant usage on database NYPD_DB to role TEMP_ROLE;

grant ownership on schema NYPD_DB.NYPD_SCHEMA to role TEMP_ROLE;
grant usage on schema NYPD_DB.NYPD_SCHEMA to role TEMP_ROLE;

create or replace user TEMP_USER password='snowflake123#' default_role=TEMP_ROLE default_warehouse=NYPD_WH;

grant role TEMP_ROLE to user TEMP_USER;
GRANT OWNERSHIP ON SCHEMA NYPD_DB.NYPD_SCHEMA TO ROLE TEMP_ROLE;



USE DATABASE NYPD_DB;
USE SCHEMA NYPD_DB.NYPD_SCHEMA;

GRANT USAGE ON DATABASE NYPD_DB TO ROLE TEMP_ROLE;
GRANT USAGE ON SCHEMA NYPD_DB.NYPD_SCHEMA TO ROLE TEMP_ROLE;
GRANT CREATE TABLE, CREATE VIEW, CREATE STAGE, CREATE FILE FORMAT ON SCHEMA NYPD_DB.NYPD_SCHEMA TO ROLE TEMP_ROLE;
SHOW GRANTS TO ROLE TEMP_ROLE;
SHOW GRANTS ON SCHEMA NYPD_DB.NYPD_SCHEMA;

USE ROLE TEMP_ROLE;


CREATE TABLE CRIME_FACT(
    CF_SK           VARCHAR(50)     NOT NULL,
    LOCATION_SK     VARCHAR(50)     NOT NULL,
    CRIMES_SK       VARCHAR(50)     NOT NULL,
    CRIMINAL_SK     VARCHAR(50)     NOT NULL,
    DATE_SK         VARCHAR(50)     NOT NULL,
    ARRESTS_SK      VARCHAR(100)    NOT NULL,
    DI_JOB_ID       VARCHAR(100),
    DI_JOB_DATE     DATE,
    CONSTRAINT PK5 PRIMARY KEY (CF_SK, LOCATION_SK, CRIMES_SK, CRIMINAL_SK, DATE_SK, ARRESTS_SK)
)
```



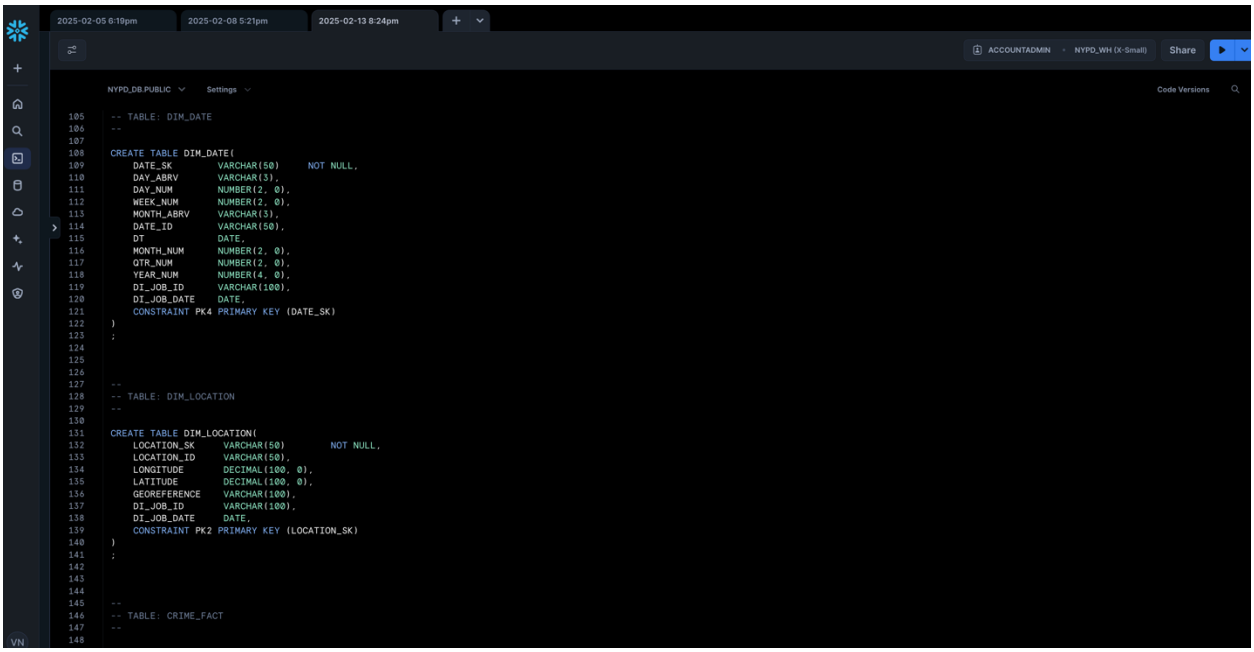
```
--
CREATE TABLE CRIME_FACT(
    CF_SK           VARCHAR(50)     NOT NULL,
    LOCATION_SK     VARCHAR(50)     NOT NULL,
    CRIMES_SK       VARCHAR(50)     NOT NULL,
    CRIMINAL_SK     VARCHAR(50)     NOT NULL,
    DATE_SK         VARCHAR(50)     NOT NULL,
    ARRESTS_SK      VARCHAR(100)    NOT NULL,
    DI_JOB_ID       VARCHAR(100),
    DI_JOB_DATE     DATE,
    CONSTRAINT PK5 PRIMARY KEY (CF_SK, LOCATION_SK, CRIMES_SK, CRIMINAL_SK, DATE_SK, ARRESTS_SK)
)
;



--
-- TABLE: DIM_ARRESTS
--

CREATE TABLE DIM_ARRESTS(
    ARRESTS_SK              VARCHAR(100)     NOT NULL,
    ARREST_ID               VARCHAR(50)      NOT NULL,
    ARREST_BOROUGH          VARCHAR(10),
    ARREST_PRECINCT         VARCHAR(10),
    ARREST_JURIDISCTION     NUMBER(10, 0),
    DI_JOB_ID               VARCHAR(100),
    DI_JOB_DT               VARCHAR(100),
    CONSTRAINT ARRESTS_SK PRIMARY KEY (ARRESTS_SK)
)
;



--
-- TABLE: DIM_CRIMES
--

CREATE TABLE DIM_CRIMES(
    CRIMES_SK       VARCHAR(50)     NOT NULL,
    CRIMES_ID       VARCHAR(50),
    PD_DESC         CHAR(100),
    OFFENSE_CAT     VARCHAR(100),
    LAW_CODE        VARCHAR(50),
```

VARANA NAVADIYA-002308207

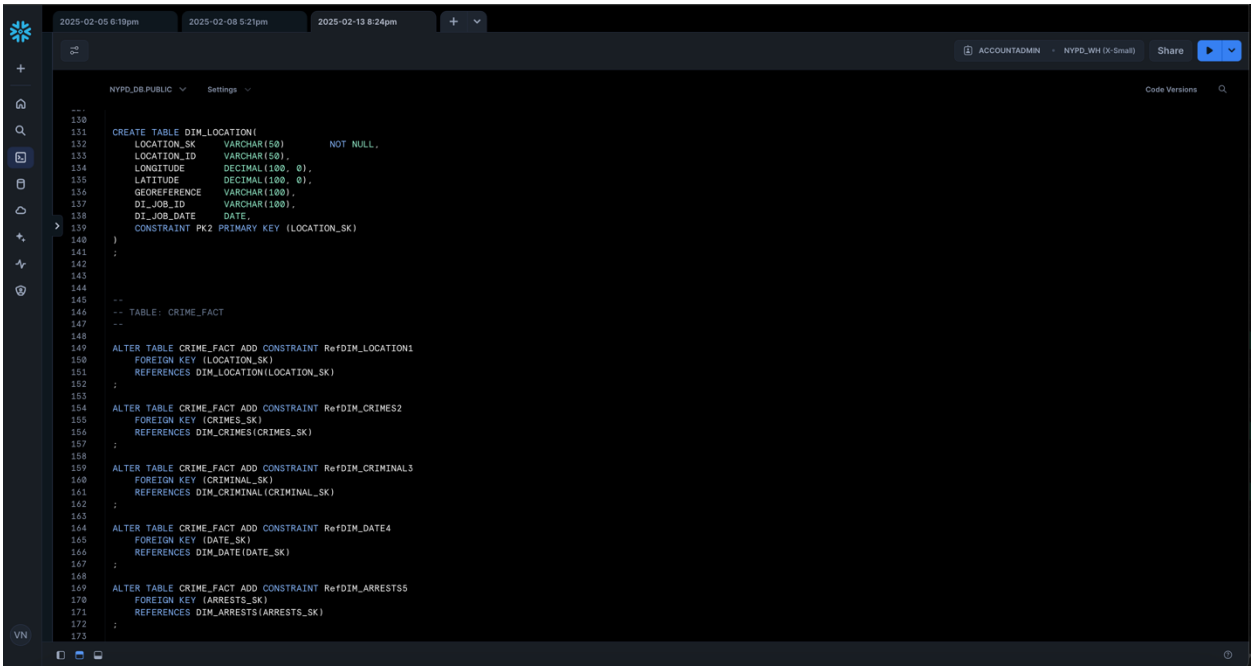




VARANA NAVADIYA-002308207



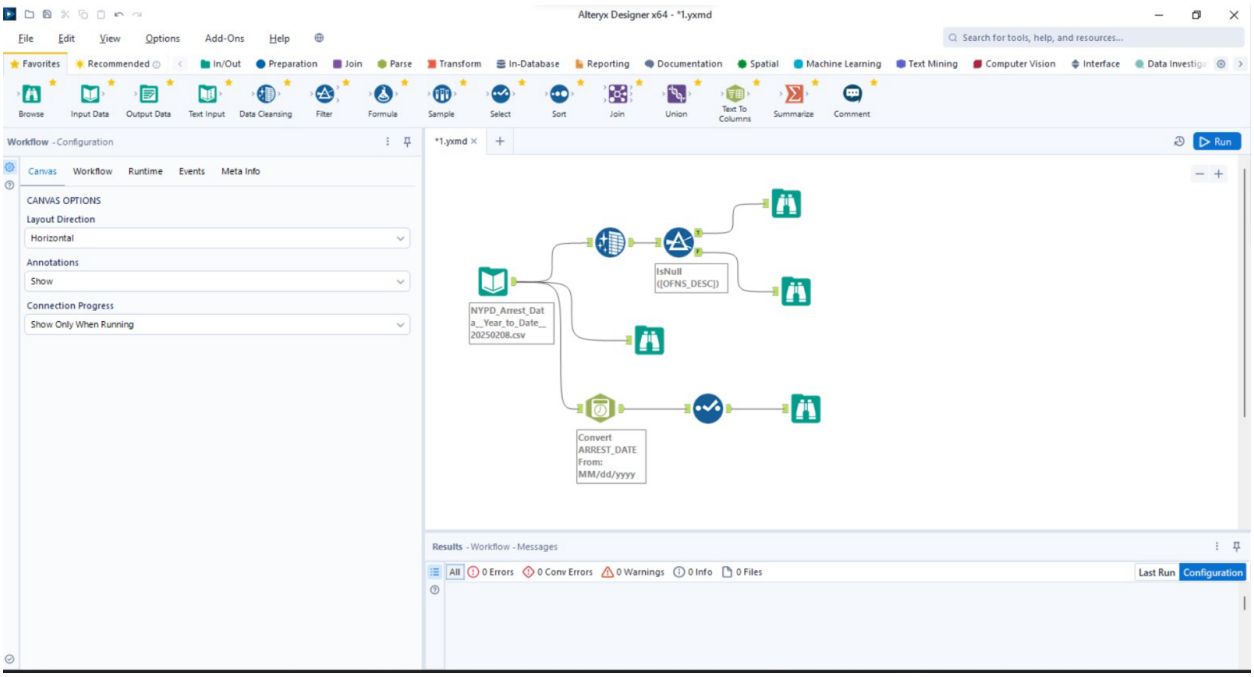
```
-- TABLE: DIM_DATE
--

CREATE TABLE DIM_DATE(
    DATE_SK         VARCHAR(50)     NOT NULL,
    DAY_ABRV        VARCHAR(3),
    DAY_NUM         NUMBER(2, 0),
    WEEK_NUM        NUMBER(2, 0),
    MONTH_ABRV      VARCHAR(3),
    DATE_ID         VARCHAR(50),
    DT              DATE,
    MONTH_NUM       NUMBER(2, 0),
    QTR_NUM         NUMBER(2, 0),
    YEAR_NUM        NUMBER(4, 0),
    DI_JOB_ID       VARCHAR(100),
    DI_JOB_DATE     DATE,
    CONSTRAINT PK4 PRIMARY KEY (DATE_SK)
)
;



--
-- TABLE: DIM_LOCATION
--

CREATE TABLE DIM_LOCATION(
    LOCATION_SK     VARCHAR(50)        NOT NULL,
    LOCATION_ID     VARCHAR(50),
    LONGITUDE       DECIMAL(100, 0),
    LATITUDE        DECIMAL(100, 0),
    GEOREFERENCE    VARCHAR(100),
    DI_JOB_ID       VARCHAR(100),
    DI_JOB_DATE     DATE,
    CONSTRAINT PK2 PRIMARY KEY (LOCATION_SK)
)
;



--
-- TABLE: CRIME_FACT
--
```



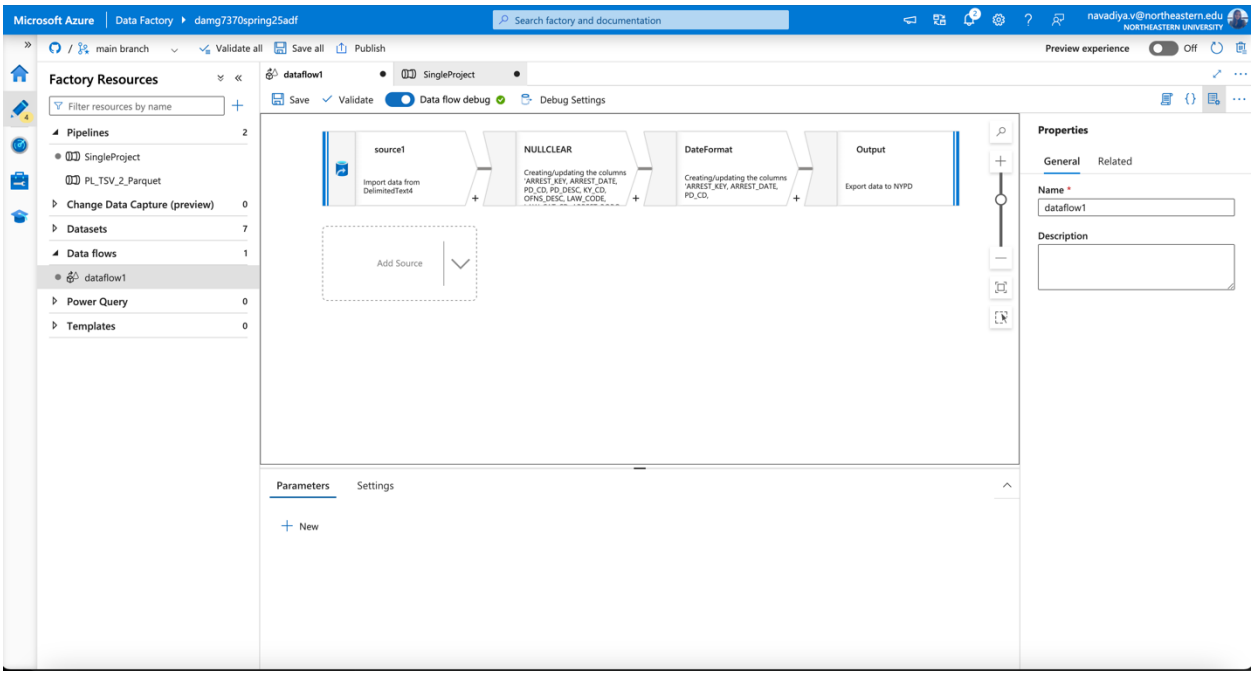
```
CREATE TABLE DIM_LOCATION(
    LOCATION_SK     VARCHAR(50)        NOT NULL,
    LOCATION_ID     VARCHAR(50),
    LONGITUDE       DECIMAL(100, 0),
    LATITUDE        DECIMAL(100, 0),
    GEOREFERENCE    VARCHAR(100),
    DI_JOB_ID       VARCHAR(100),
    DI_JOB_DATE     DATE,
    CONSTRAINT PK2 PRIMARY KEY (LOCATION_SK)
)
;



--
-- TABLE: CRIME_FACT
--

ALTER TABLE CRIME_FACT ADD CONSTRAINT RefDIM_LOCATION1
    FOREIGN KEY (LOCATION_SK)
    REFERENCES DIM_LOCATION(LOCATION_SK)
;

ALTER TABLE CRIME_FACT ADD CONSTRAINT RefDIM_CRIMES2
    FOREIGN KEY (CRIMES_SK)
    REFERENCES DIM_CRIMES(CRIMES_SK)
;

ALTER TABLE CRIME_FACT ADD CONSTRAINT RefDIM_CRIMINAL3
    FOREIGN KEY (CRIMINAL_SK)
    REFERENCES DIM_CRIMINAL(CRIMINAL_SK)
;

ALTER TABLE CRIME_FACT ADD CONSTRAINT RefDIM_DATE4
    FOREIGN KEY (DATE_SK)
    REFERENCES DIM_DATE(DATE_SK)
;

ALTER TABLE CRIME_FACT ADD CONSTRAINT RefDIM_ARRESTS5
    FOREIGN KEY (ARRESTS_SK)
    REFERENCES DIM_ARRESTS(ARRESTS_SK)
;
```

VARANA NAVADIYA-002308207

-ALTERYX:

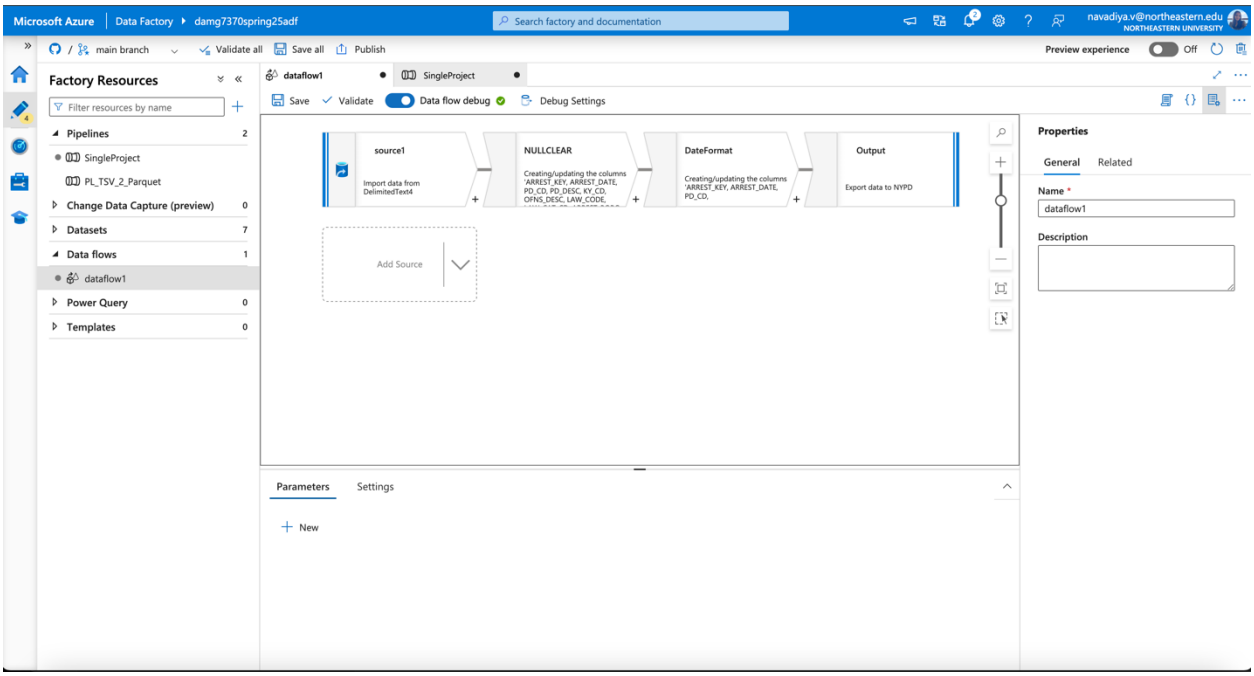


-DATAFLOW:

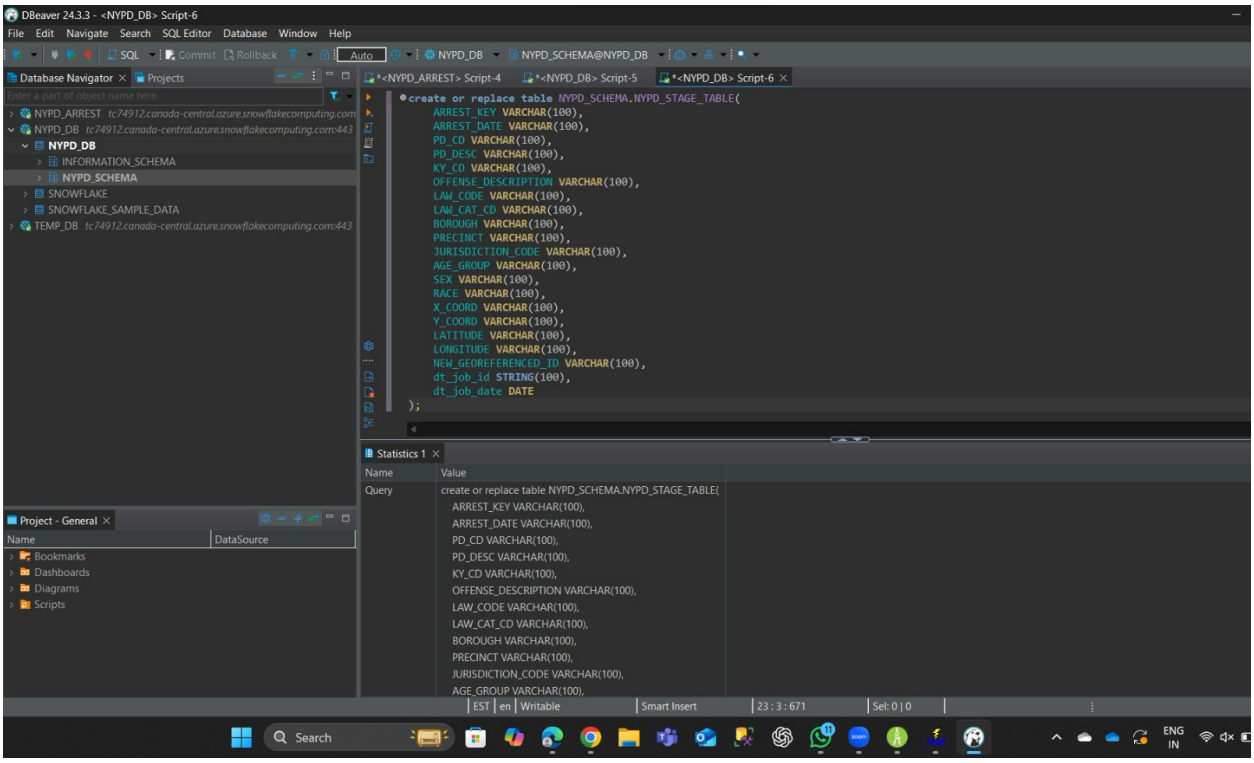


VARANA NAVADIYA-002308207

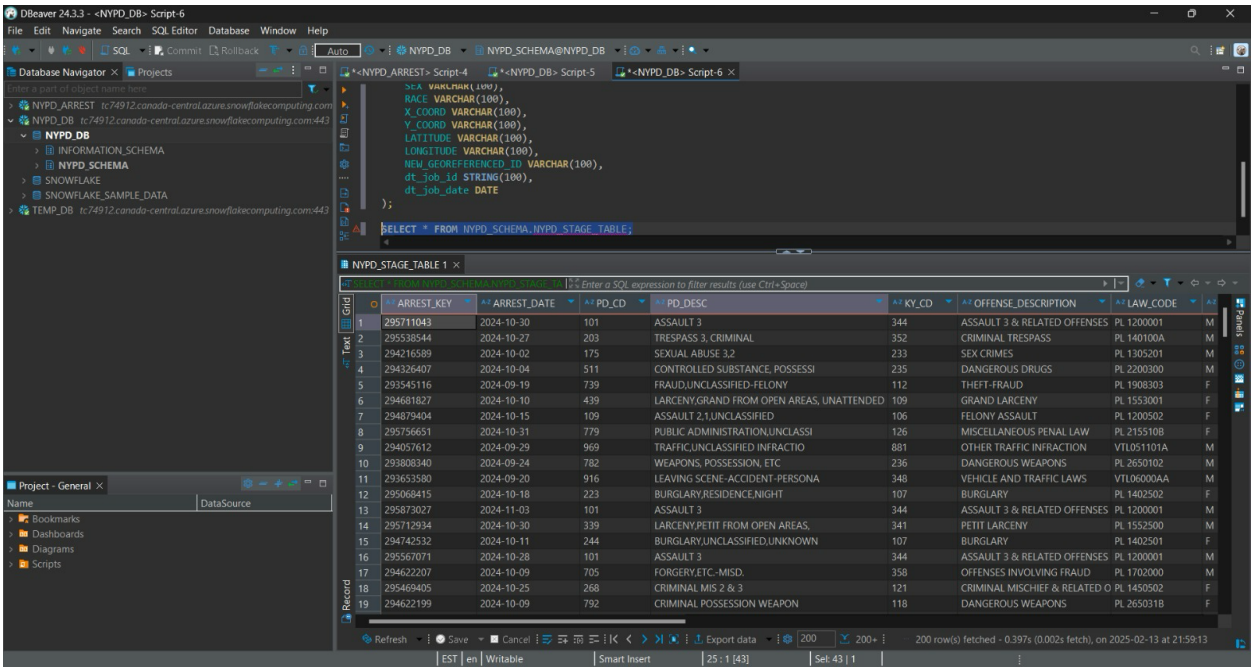
-ADF PIPELINE:



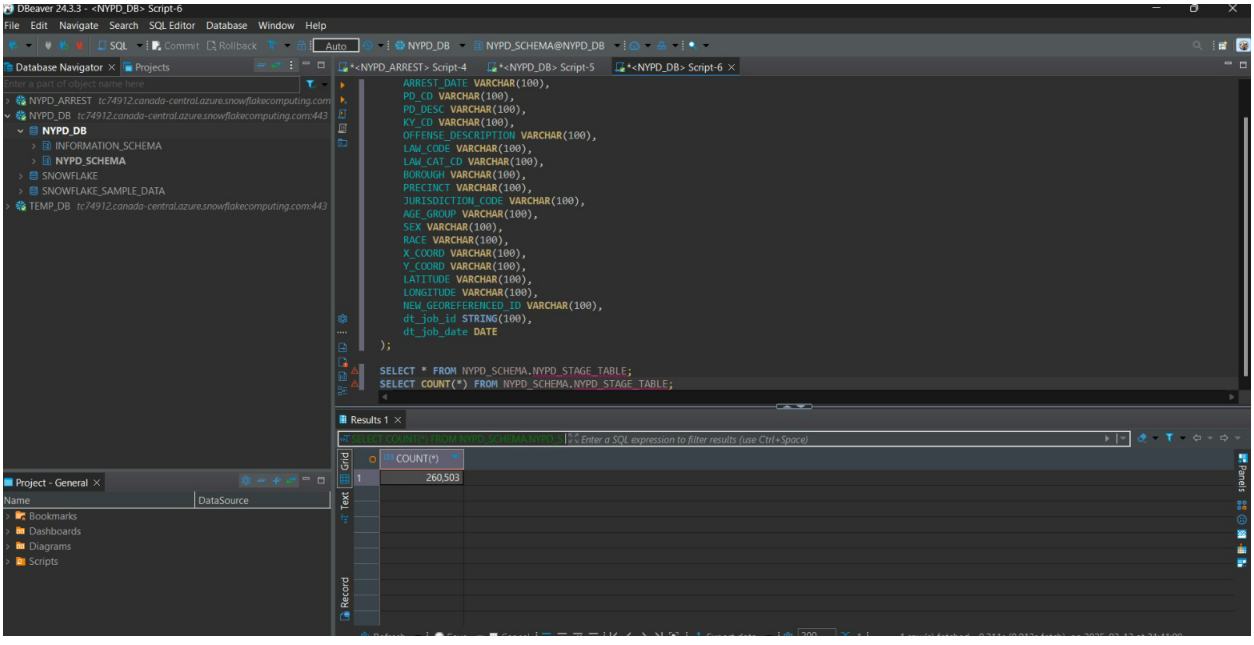
-DATA POPULATED IN SNOWFLAKE DATABASE:



VARANA NAVADIYA-002308207

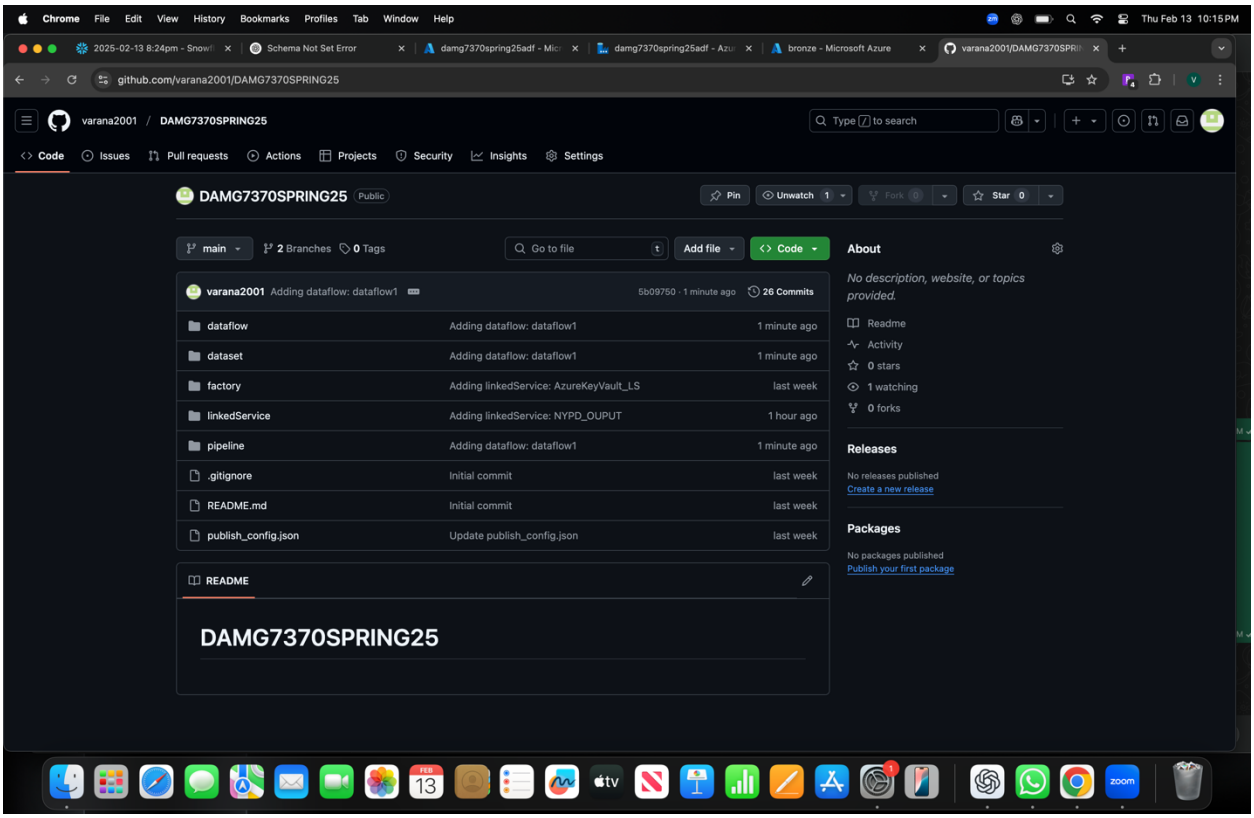


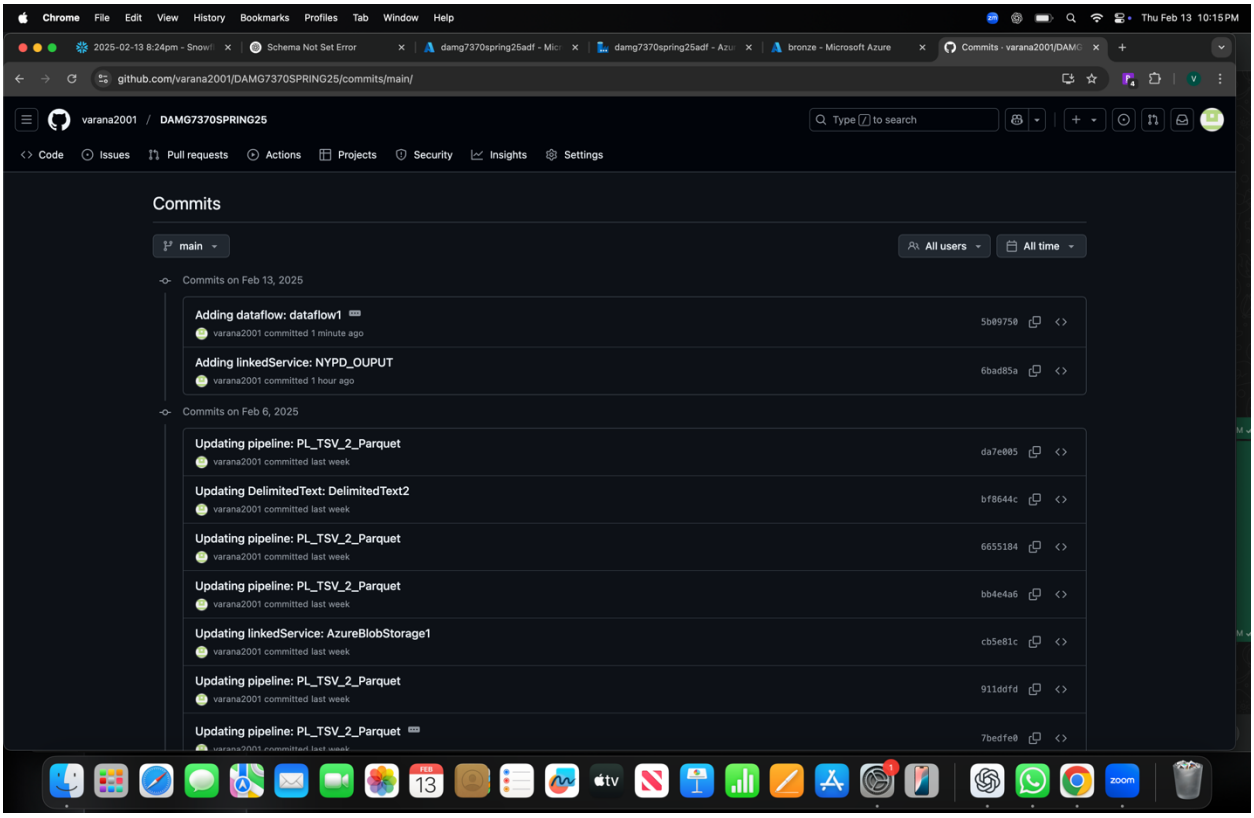
-ROW COUNTS USING SQL QUERY:



VARANA NAVADIYA-002308207

-GITHUB:





VARANA NAVADIYA-002308207

GitHub Link: https://github.com/varana2001/DAMG7370SPRING25

DataFlow Link: https://github.com/varana2001/DAMG7370SPRING25/tree/main/dataflow

Pipeline Link: https://github.com/varana2001/DAMG7370SPRING25/tree/main/pipeline