Individual Assignment

Je Sai Kailash Pulipati

002339774

Overview

Basic Information

- Total Number of records/rows/lines: 260,503 records
- Total Number of columns/fields: 19 Columns
- Source and Target Data types
- Grain: Each individual arrest record

Column Name	Source Data Type	Target Data Type (SQL)
ARREST_KEY	String	VARCHAR
ARREST_DATE	V_String	DATE
PD_CD	V_String	INT
PD_DESC	V_String	VARCHAR
KY_CD	String	INT
OFNS_DESC	String	VARCHAR
LAW_CODE	String	VARCHAR
LAW_CAT_CD	String	VARCHAR
ARREST_BORO	V_String	VARCHAR
ARREST_PRECINCT	String	INT
JURISDICTION_CODE	V_String	INT
AGE_GROUP	V_String	VARCHAR
PERP_SEX	String	VARCHAR
PERP_RACE	String	VARCHAR
X_COORD_CD	V_String	INT
Y_COORD_CD	V_String	INT
Latitude	V_String	FLOAT
Longitude	V_String	FLOAT

New Georeferenced Column	String	VARCHAR

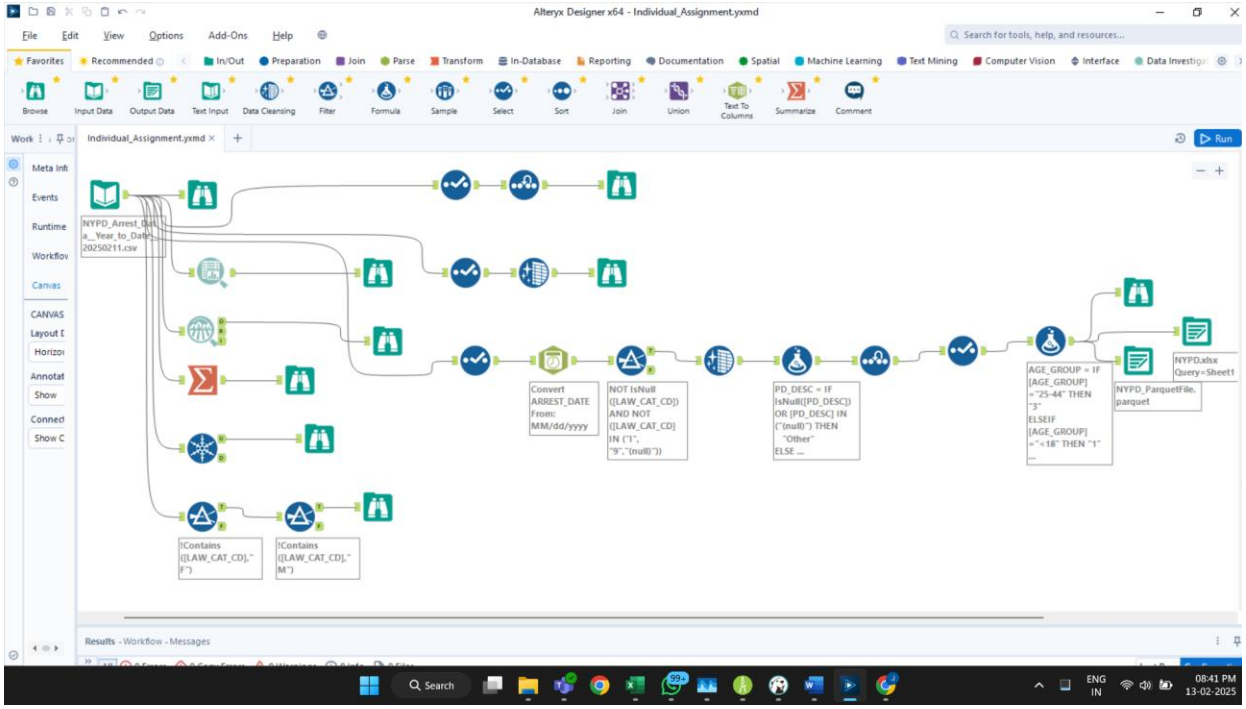
Data Quality Assessment and Data Cleaning

Sno.	Column Name	Issue	Details	Action Plan to fix it
1.	ARREST_DATE	Date in String format	Need to convert it to date format	Use parse DateTime to change the format to DD-MM-YYYY
2.	LAW_CAT_CD	Only 3 levels of Felony present	Other types or Misc felonies reported like “9”, “I”, “(null)”	Used filter Tool to remove unwanted rows
3.	ARREST_DATE, PD_CD, KY_CD,	Null Values present	Contains Null values in between	Used Data cleaning tool to remove the null values and replace with “0”
4.	PD_DESC	Null Values present	Contains Null values like “(null)”	Used formula tool to replace “(null)” to “Other”
5.	Longitude and Latitude	Null Values	Contains Null values	Used Data Imputation to replace null values with Mean
6.	Select Tool	Wrong Data format	Need to change Data formats	Used Select tool to change the data types
7.	AGE_GROUP	Unusable age group divisions	Need to remove divisions and use some unique identifying values	Used Formula tool to divide age groups into values
8.	LAW_CAT_CD	Jurisdiction out of range	Only 0,1,2 Jurisdictions should be present	Use filter tool to separate rows which are above 2

Insights and Observations

1. The LAW_CAT_CD column has violations other than M, F, V.
2. There are some Jurisdiction codes which are outside the NYPD which can be removed
3. Null values are identified using the Summarise tool
4. There are no duplicate Arrest_key records
5. Some longitudes, latitudes have “0” so used data imputation to fill with average values so that they won’t be affected by becoming outliers
6. Converted Age_groups into categories like 1,2,3,4,5.
7. Some column names are not in proper standart format, so converted all column names into Capital with no spaces.

Alteryx Workflow



NYPD_Arrest_Data__Year_to_Date_20250211.csv

Convert ARREST_DATE From: MM/dd/yyyy

NOT IsNull ([LAW_CAT_CD]) AND NOT ([LAW_CAT_CD] IN ("I", "9","(null)"))

PD_DESC = IF IsNull([PD_DESC]) OR [PD_DESC] IN ("(null)") THEN "Other" ELSE ...

AGE_GROUP = IF [AGE_GROUP] ="25-44" THEN "3" ELSEIF [AGE_GROUP] ="<18" THEN "1" ...

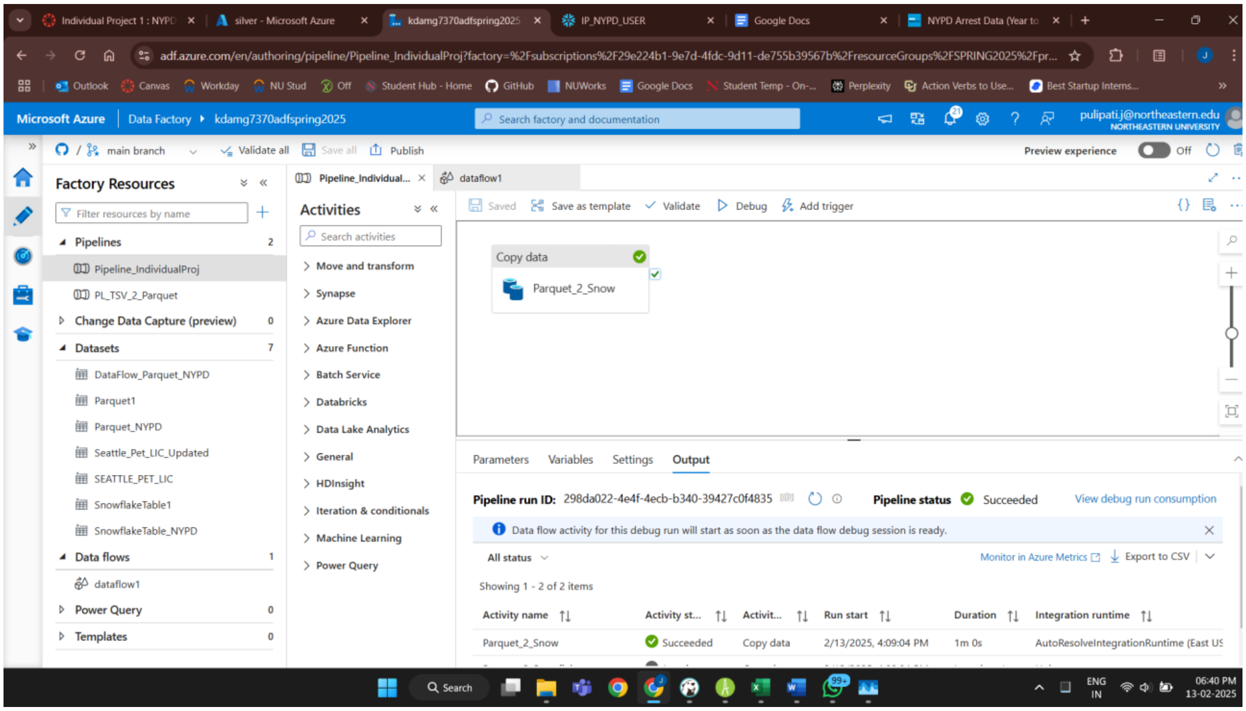
NYPD_ParquetFile.parquet

NYPD.xlsx Query=Sheet1

!Contains ([LAW_CAT_CD]," F")

!Contains ([LAW_CAT_CD]," M")

ADF-Snowflake Pipeline



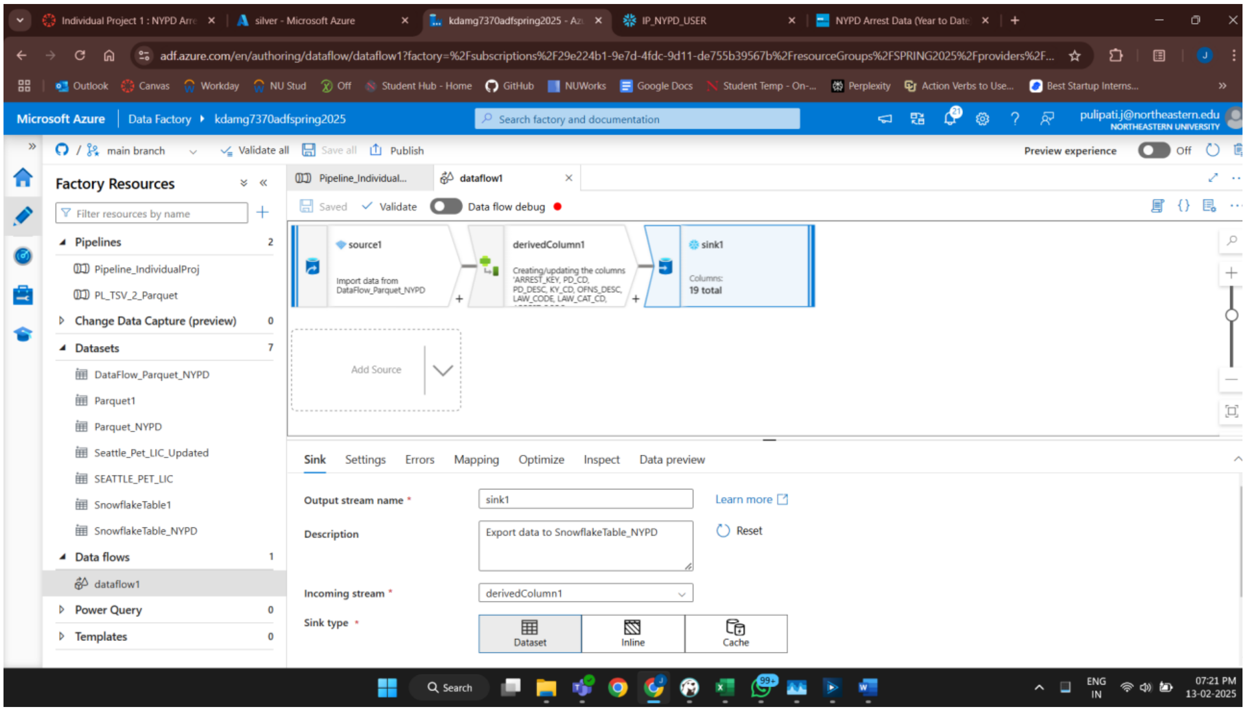
Pipeline run ID: 298da022-4e4f-4ecb-b340-39427c0f4835 **Pipeline status** Succeeded View debug run consumption

Data flow activity for this debug run will start as soon as the data flow debug session is ready.

All status | Monitor in Azure Metrics | Export to CSV

Showing 1 - 2 of 2 items

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime
Parquet_2_Snow	Succeeded	Copy data	2/13/2025, 4:09:04 PM	1m 0s	AutoResolveIntegrationRuntime (East US



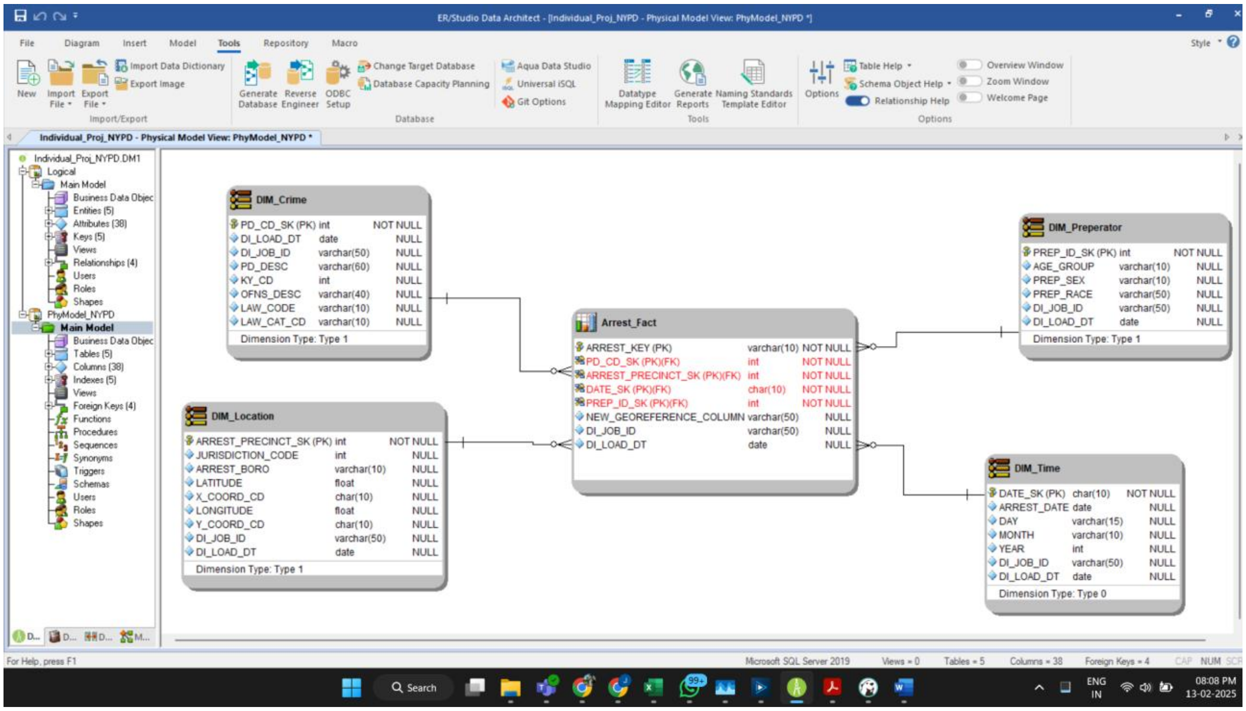
Output stream name * sink1

Description: Export data to SnowflakeTable_NYPD

Incoming stream * derivedColumn1

Sink type * Dataset | Inline | Cache

ER Diagram:


ER/Studio Data Architect - [Individual_Proj_NYPD - Physical Model View: PhyModel_NYPD *]
Individual_Proj_NYPD - Physical Model View: PhyModel_NYPD *
Microsoft SQL Server 2019
Views = 0
Tables = 5
Columns = 38
Foreign Keys = 4

DIM_Crime

PD_CD_SK (PK)	int	NOT NULL
DI_LOAD_DT	date	NULL
DI_JOB_ID	varchar(50)	NULL
PD_DESC	varchar(60)	NULL
KY_CD	int	NULL
OFNS_DESC	varchar(40)	NULL
LAW_CODE	varchar(10)	NULL
LAW_CAT_CD	varchar(10)	NULL

Dimension Type: Type 1

DIM_Location

ARREST_PRECINCT_SK (PK)	int	NOT NULL
JURISDICTION_CODE	int	NULL
ARREST_BORO	varchar(10)	NULL
LATITUDE	float	NULL
X_COORD_CD	char(10)	NULL
LONGITUDE	float	NULL
Y_COORD_CD	char(10)	NULL
DI_JOB_ID	varchar(50)	NULL
DI_LOAD_DT	date	NULL

Dimension Type: Type 1

Arrest_Fact

ARREST_KEY (PK)	varchar(10)	NOT NULL
PD_CD_SK (PK)(FK)	int	NOT NULL
ARREST_PRECINCT_SK (PK)(FK)	int	NOT NULL
DATE_SK (PK)(FK)	char(10)	NOT NULL
PREP_ID_SK (PK)(FK)	int	NOT NULL
NEW_GEOREFERENCE_COLUMN	varchar(50)	NULL
DI_JOB_ID	varchar(50)	NULL
DI_LOAD_DT	date	NULL

DIM_Preperator

PREP_ID_SK (PK)	int	NOT NULL
AGE_GROUP	varchar(10)	NULL
PREP_SEX	varchar(10)	NULL
PREP_RACE	varchar(50)	NULL
DI_JOB_ID	varchar(50)	NULL
DI_LOAD_DT	date	NULL

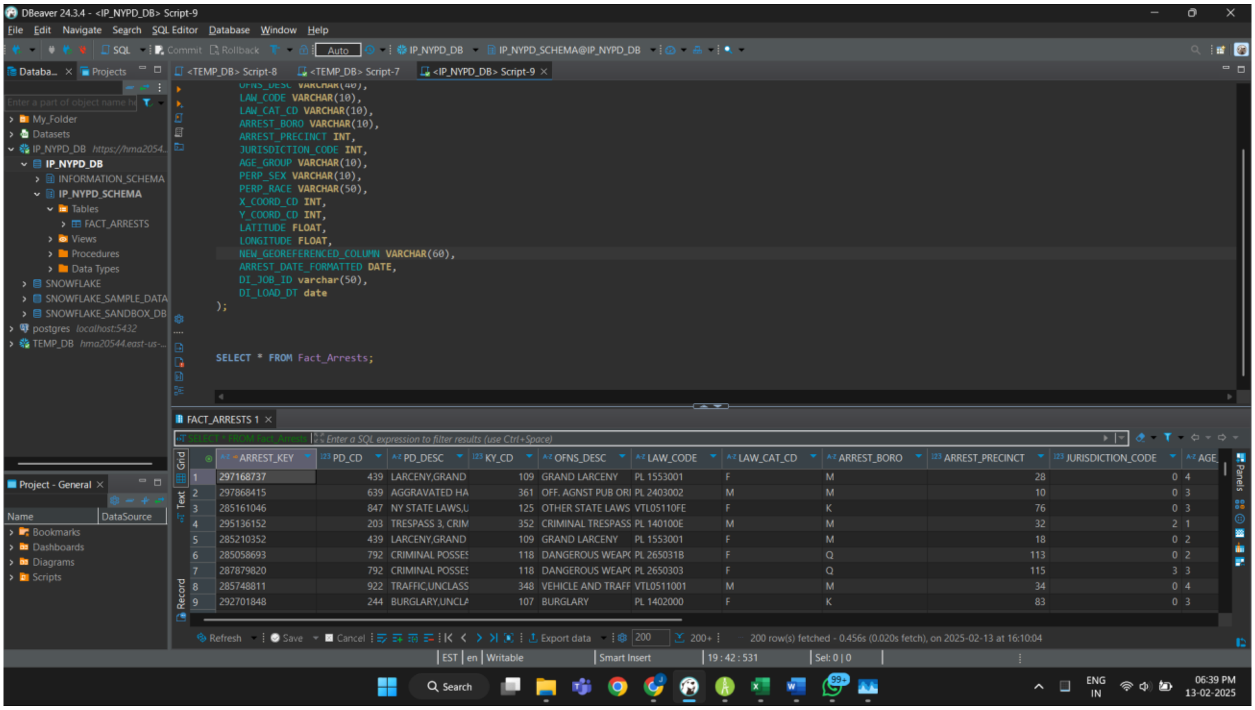
Dimension Type: Type 1

DIM_Time

DATE_SK (PK)	char(10)	NOT NULL
ARREST_DATE	date	NULL
DAY	varchar(15)	NULL
MONTH	varchar(10)	NULL
YEAR	int	NULL
DI_JOB_ID	varchar(50)	NULL
DI_LOAD_DT	date	NULL

Dimension Type: Type 0

Final Created Table pushed to Snowflake:

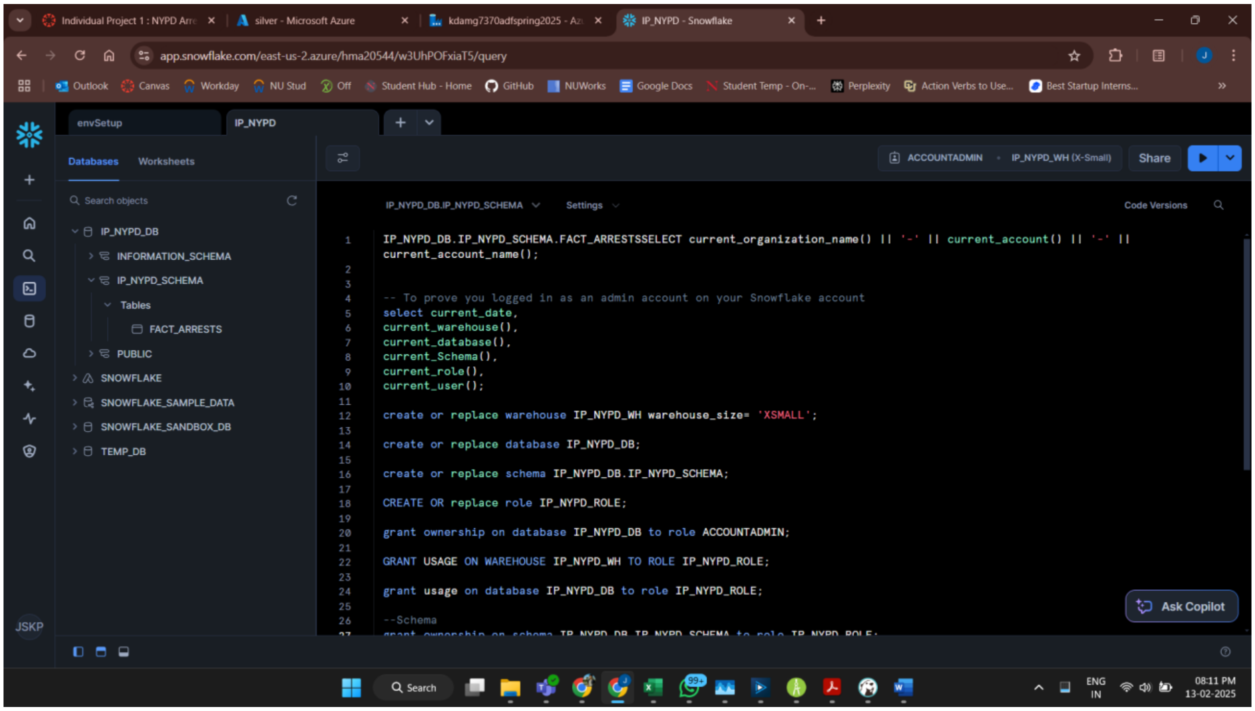


```
    OFNS_DESC VARCHAR(40),
    LAW_CODE VARCHAR(10),
    LAW_CAT_CD VARCHAR(10),
    ARREST_BORO VARCHAR(10),
    ARREST_PRECINCT INT,
    JURISDICTION_CODE INT,
    AGE_GROUP VARCHAR(10),
    PERP_SEX VARCHAR(10),
    PERP_RACE VARCHAR(50),
    X_COORD_CD INT,
    Y_COORD_CD INT,
    LATITUDE FLOAT,
    LONGITUDE FLOAT,
    NEW_GEOREFERENCED_COLUMN VARCHAR(60),
    ARREST_DATE_FORMATTED DATE,
    DI_JOB_ID varchar(50),
    DI_LOAD_DT date
);


SELECT * FROM Fact_Arrests;
```

	ARREST_KEY	PD_CD	PD_DESC	KY_CD	OFNS_DESC	LAW_CODE	LAW_CAT_CD	ARREST_BORO	ARREST_PRECINCT	JURISDICTION_CODE	AGE_
1	297168737	439	LARCENY,GRAND	109	GRAND LARCENY	PL 1553001	F	M	28	0	4
2	297868415	639	AGGRAVATED HA	361	OFF. AGNST PUB ORI	PL 2403002	M	M	10	0	3
3	285161046	847	NY STATE LAWS,U	125	OTHER STATE LAWS	VTL05110FE	F	K	76	0	3
4	295136152	203	TRESPASS 3, CRIM	352	CRIMINAL TRESPASS	PL 140100E	M	M	32	2	1
5	285210352	439	LARCENY,GRAND	109	GRAND LARCENY	PL 1553001	F	M	18	0	2
6	285058693	792	CRIMINAL POSSES	118	DANGEROUS WEAP(	PL 265031B	F	Q	113	0	2
7	287879820	792	CRIMINAL POSSES	118	DANGEROUS WEAP(	PL 2650303	F	Q	115	3	3
8	285748811	922	TRAFFIC,UNCLASS	348	VEHICLE AND TRAFF	VTL0511001	M	M	34	0	4
9	292701848	244	BURGLARY,UNCLA	107	BURGLARY	PL 1402000	F	K	83	0	3

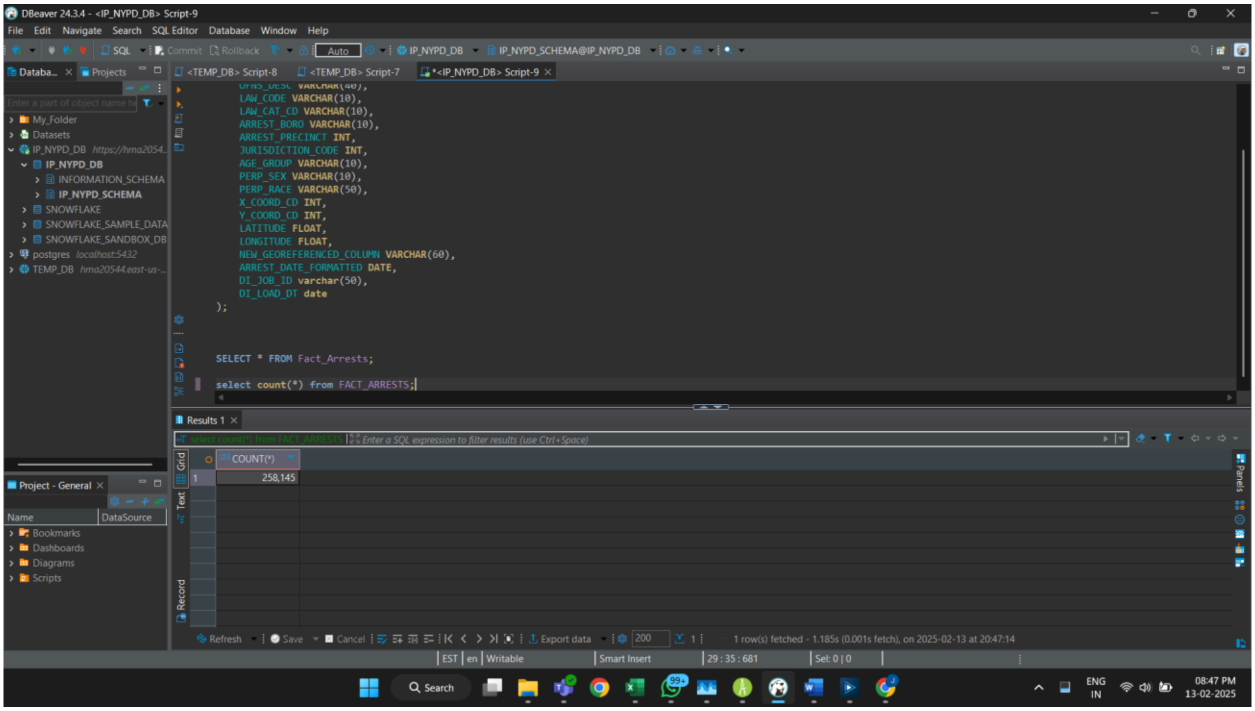
200 row(s) fetched - 0.456s (0.020s fetch), on 2025-02-13 at 16:10:04

Snowflake DB Creation



```
IP_NYPD_DB.IP_NYPD_SCHEMA.FACT_ARRESTSSELECT current_organization_name() || '-' || current_account() || '-' ||
current_account_name();


-- To prove you logged in as an admin account on your Snowflake account
select current_date,
current_warehouse(),
current_database(),
current_Schema(),
current_role(),
current_user();

create or replace warehouse IP_NYPD_WH warehouse_size= 'XSMALL';

create or replace database IP_NYPD_DB;

create or replace schema IP_NYPD_DB.IP_NYPD_SCHEMA;

CREATE OR replace role IP_NYPD_ROLE;

grant ownership on database IP_NYPD_DB to role ACCOUNTADMIN;

GRANT USAGE ON WAREHOUSE IP_NYPD_WH TO ROLE IP_NYPD_ROLE;

grant usage on database IP_NYPD_DB to role IP_NYPD_ROLE;

--Schema
grant ownership on schema IP_NYPD_DB.IP_NYPD_SCHEMA to role IP_NYPD_ROLE;
```

Count (*) rows from Table

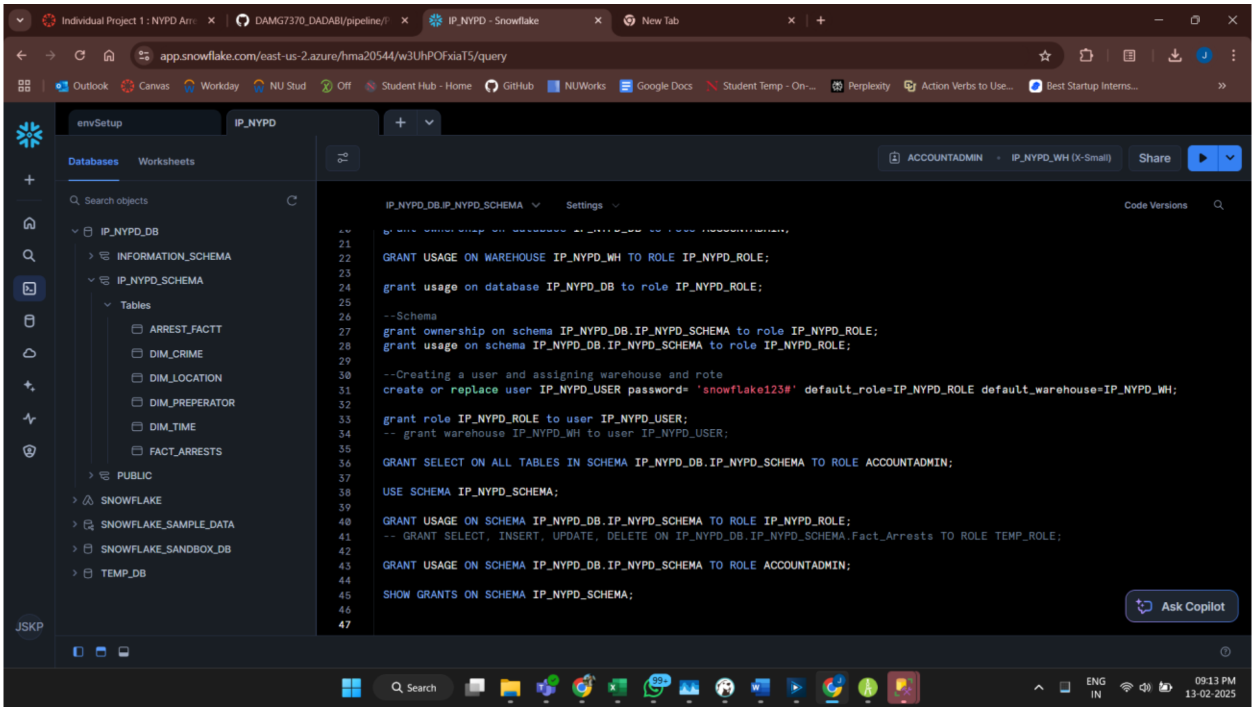


```
    OFNS_DESC VARCHAR(40),
    LAW_CODE VARCHAR(10),
    LAW_CAT_CD VARCHAR(10),
    ARREST_BORO VARCHAR(10),
    ARREST_PRECINCT INT,
    JURISDICTION_CODE INT,
    AGE_GROUP VARCHAR(10),
    PERP_SEX VARCHAR(10),
    PERP_RACE VARCHAR(50),
    X_COORD_CD INT,
    Y_COORD_CD INT,
    LATITUDE FLOAT,
    LONGITUDE FLOAT,
    NEW_GEOREFERENCED_COLUMN VARCHAR(60),
    ARREST_DATE_FORMATTED DATE,
    DI_JOB_ID varchar(50),
    DI_LOAD_DT date
);


SELECT * FROM Fact_Arrests;

select count(*) from FACT_ARRESTS;
```

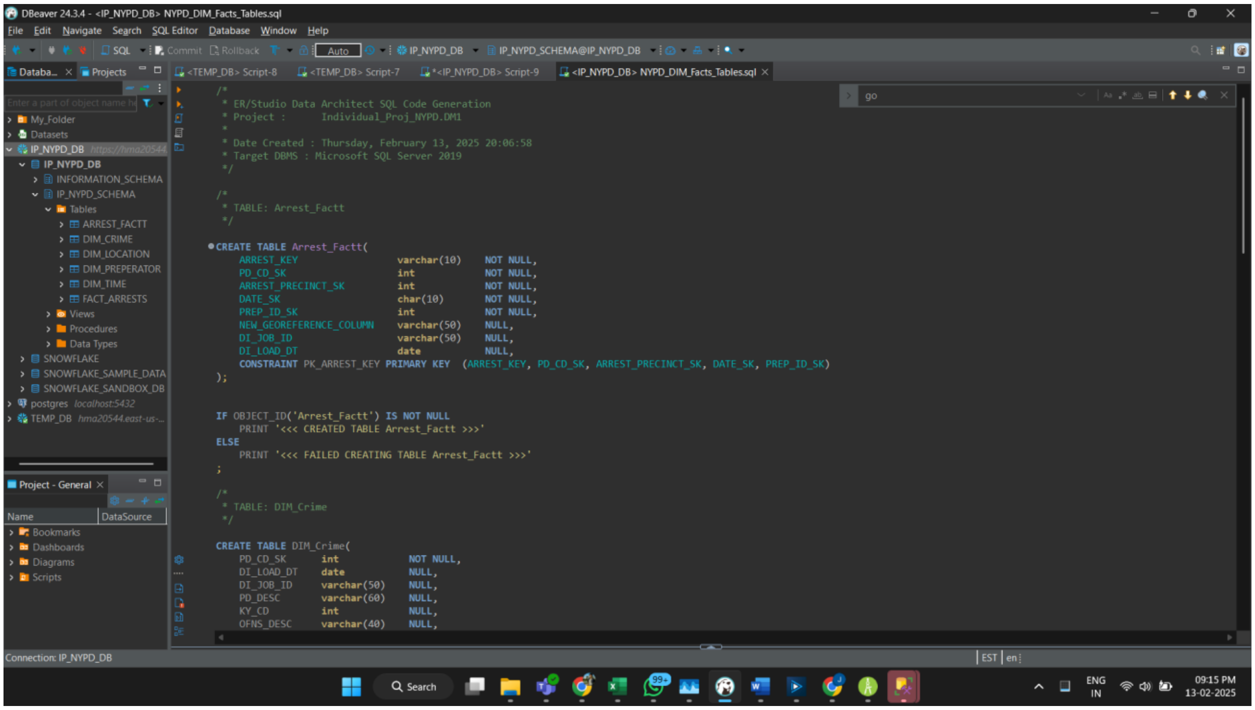
	COUNT(*)
1	258,145

```
1 row(s) fetched - 1.185s (0.001s fetch), on 2025-02-13 at 20:47:14
```

DIMENSIONS AND FACTS TABLES:



```
grant ownership on database IP_NYPD_DB to role ACCOUNTADMIN;

GRANT USAGE ON WAREHOUSE IP_NYPD_WH TO ROLE IP_NYPD_ROLE;

grant usage on database IP_NYPD_DB to role IP_NYPD_ROLE;

--Schema
grant ownership on schema IP_NYPD_DB.IP_NYPD_SCHEMA to role IP_NYPD_ROLE;
grant usage on schema IP_NYPD_DB.IP_NYPD_SCHEMA to role IP_NYPD_ROLE;

--Creating a user and assigning warehouse and rote
create or replace user IP_NYPD_USER password= 'snowflake123#' default_role=IP_NYPD_ROLE default_warehouse=IP_NYPD_WH;

grant role IP_NYPD_ROLE to user IP_NYPD_USER;
-- grant warehouse IP_NYPD_WH to user IP_NYPD_USER;

GRANT SELECT ON ALL TABLES IN SCHEMA IP_NYPD_DB.IP_NYPD_SCHEMA TO ROLE ACCOUNTADMIN;

USE SCHEMA IP_NYPD_SCHEMA;

GRANT USAGE ON SCHEMA IP_NYPD_DB.IP_NYPD_SCHEMA TO ROLE IP_NYPD_ROLE;
-- GRANT SELECT, INSERT, UPDATE, DELETE ON IP_NYPD_DB.IP_NYPD_SCHEMA.Fact_Arrests TO ROLE TEMP_ROLE;

GRANT USAGE ON SCHEMA IP_NYPD_DB.IP_NYPD_SCHEMA TO ROLE ACCOUNTADMIN;

SHOW GRANTS ON SCHEMA IP_NYPD_SCHEMA;
```



```sql
/*
 * ER/Studio Data Architect SQL Code Generation
 * Project :      Individual_Proj_NYPD.DM1
 *
 * Date Created : Thursday, February 13, 2025 20:06:58
 * Target DBMS : Microsoft SQL Server 2019
 */

/*
 * TABLE: Arrest_Factt
 */

CREATE TABLE Arrest_Factt(
    ARREST_KEY                 varchar(10)    NOT NULL,
    PD_CD_SK                   int            NOT NULL,
    ARREST_PRECINCT_SK         int            NOT NULL,
    DATE_SK                    char(10)       NOT NULL,
    PREP_ID_SK                 int            NOT NULL,
    NEW_GEOREFERENCE_COLUMN    varchar(50)    NULL,
    DI_JOB_ID                  varchar(50)    NULL,
    DI_LOAD_DT                 date           NULL,
    CONSTRAINT PK_ARREST_KEY PRIMARY KEY  (ARREST_KEY, PD_CD_SK, ARREST_PRECINCT_SK, DATE_SK, PREP_ID_SK)
);


IF OBJECT_ID('Arrest_Factt') IS NOT NULL
    PRINT '<<< CREATED TABLE Arrest_Factt >>>'
ELSE
    PRINT '<<< FAILED CREATING TABLE Arrest_Factt >>>'
;

/*
 * TABLE: DIM_Crime
 */

CREATE TABLE DIM_Crime(
    PD_CD_SK      int            NOT NULL,
    DI_LOAD_DT    date           NULL,
    DI_JOB_ID     varchar(50)    NULL,
    PD_DESC       varchar(60)    NULL,
    KY_CD         int            NULL,
    OFNS_DESC     varchar(40)    NULL,
```

SQL SCRIPT

```
/*
 * ER/Studio Data Architect SQL Code Generation
 * Project :      Individual_Proj_NYPD.DM1
 *
 * Date Created : Thursday, February 13, 2025 20:06:58
 * Target DBMS : Microsoft SQL Server 2019
 */


/*
 * TABLE: Arrest_Fact
 */


CREATE TABLE Arrest_Fact(
    ARREST_KEY              varchar(10)    NOT NULL,
    PD_CD_SK                int            NOT NULL,
    ARREST_PRECINCT_SK        int            NOT NULL,
  DATE_SK              char(10)     NOT NULL,
  PREP_ID_SK            int         NOT NULL,
  NEW_GEOREFERENCE_COLUMN   varchar(50)   NULL,
  DI_JOB_ID             varchar(50)   NULL,
  DI_LOAD_DT            date         NULL,
  CONSTRAINT PK_ARREST_KEY PRIMARY KEY NONCLUSTERED (ARREST_KEY, PD_CD_SK,
ARREST_PRECINCT_SK, DATE_SK, PREP_ID_SK)
)

go



IF OBJECT_ID('Arrest_Fact') IS NOT NULL
  PRINT '<<< CREATED TABLE Arrest_Fact >>>'
ELSE
  PRINT '<<< FAILED CREATING TABLE Arrest_Fact >>>'
go


/*
 * TABLE: DIM_Crime
 */


CREATE TABLE DIM_Crime(
  PD_CD_SK    int         NOT NULL,
  DI_LOAD_DT   date        NULL,
  DI_JOB_ID    varchar(50)   NULL,
  PD_DESC     varchar(60)   NULL,
  KY_CD       int         NULL,
  OFNS_DESC    varchar(40)   NULL,
    LAW_CODE      varchar(10)    NULL,
    LAW_CAT_CD    varchar(10)    NULL,
    CONSTRAINT PK2 PRIMARY KEY NONCLUSTERED (PD_CD_SK)
)

go



IF OBJECT_ID('DIM_Crime') IS NOT NULL
    PRINT '<<< CREATED TABLE DIM_Crime >>>'
ELSE
    PRINT '<<< FAILED CREATING TABLE DIM_Crime >>>'
go

/*
 * TABLE: DIM_Location
 */

CREATE TABLE DIM_Location(
    ARREST_PRECINCT_SK    int           NOT NULL,
    JURISDICTION_CODE     int           NULL,
    ARREST_BORO           varchar(10)   NULL,
    LATITUDE              float         NULL,
    X_COORD_CD            char(10)      NULL,
    LONGITUDE             float         NULL,
    Y_COORD_CD            char(10)      NULL,
    DI_JOB_ID             varchar(50)   NULL,
    DI_LOAD_DT            date          NULL,
    CONSTRAINT PK_ARREST_PRECINCT PRIMARY KEY NONCLUSTERED (ARREST_PRECINCT_SK)
```

```
)

go


IF OBJECT_ID('DIM_Location') IS NOT NULL
    PRINT '<<< CREATED TABLE DIM_Location >>>'
ELSE
    PRINT '<<< FAILED CREATING TABLE DIM_Location >>>'
go

/*
 * TABLE: DIM_Preperator
 */

CREATE TABLE DIM_Preperator(
    PREP_ID_SK    int           NOT NULL,
    AGE_GROUP     varchar(10)    NULL,
    PREP_SEX      varchar(10)    NULL,
    PREP_RACE     varchar(50)    NULL,
    DI_JOB_ID     varchar(50)    NULL,
    DI_LOAD_DT    date           NULL,
    CONSTRAINT PK4 PRIMARY KEY NONCLUSTERED (PREP_ID_SK)
)

go


IF OBJECT_ID('DIM_Preperator') IS NOT NULL
```

```
    PRINT '<<< CREATED TABLE DIM_Preperator >>>'
ELSE
    PRINT '<<< FAILED CREATING TABLE DIM_Preperator >>>'
go

/*
 * TABLE: DIM_Time
 */

CREATE TABLE DIM_Time(
    DATE_SK       char(10)      NOT NULL,
    ARREST_DATE    date         NULL,
    DAY          varchar(15)    NULL,
    MONTH        varchar(10)    NULL,
    YEAR         int           NULL,
    DI_JOB_ID     varchar(50)    NULL,
    DI_LOAD_DT     date         NULL,
    CONSTRAINT PK_DATE_ID PRIMARY KEY NONCLUSTERED (DATE_SK)
)

go



IF OBJECT_ID('DIM_Time') IS NOT NULL
    PRINT '<<< CREATED TABLE DIM_Time >>>'
ELSE
    PRINT '<<< FAILED CREATING TABLE DIM_Time >>>'
go
```

```
/*
 * TABLE: Arrest_Fact
 */

ALTER TABLE Arrest_Fact ADD CONSTRAINT RefDIM_Crime1
    FOREIGN KEY (PD_CD_SK)
    REFERENCES DIM_Crime(PD_CD_SK)
go

ALTER TABLE Arrest_Fact ADD CONSTRAINT RefDIM_Location2
    FOREIGN KEY (ARREST_PRECINCT_SK)
    REFERENCES DIM_Location(ARREST_PRECINCT_SK)
go

ALTER TABLE Arrest_Fact ADD CONSTRAINT RefDIM_Time3
    FOREIGN KEY (DATE_SK)
    REFERENCES DIM_Time(DATE_SK)
go

ALTER TABLE Arrest_Fact ADD CONSTRAINT RefDIM_Preperator4
    FOREIGN KEY (PREP_ID_SK)
    REFERENCES DIM_Preperator(PREP_ID_SK)
go
```

DDL Scripts

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

```
SELECT dt.YEAR, dt.MONTH, COUNT(*) AS Total_Arrests FROM Arrest_Fact af

JOIN DIM_Time dt ON af.DATE_SK = dt.DATE_SK GROUP BY dt.YEAR, dt.MONTH

ORDER BY dt.YEAR, dt.MONTH;
```

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

```
SELECT dt.MONTH, COUNT(*) AS Arrest_Count FROM Arrest_Fact af

JOIN DIM_Time dt ON af.DATE_SK = dt.DATE_SK GROUP BY dt.MONTH ORDER BY Arrest_Count DESC;
```

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

```
SELECT dc.OFNS_DESC, COUNT(*) AS Crime_Count

FROM Arrest_Fact af JOIN DIM_Crime dc ON af.PD_CD_SK = dc.PD_CD_SK

GROUP BY dc.OFNS_DESC ORDER BY Crime_Count DESC;
```

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

```
SELECT dt.YEAR, dc.OFNS_DESC, COUNT(*) AS Arrest_Count FROM Arrest_Fact af JOIN DIM_Time dt ON af.DATE_SK =
dt.DATE_SK JOIN DIM_Crime dc ON af.PD_CD_SK = dc.PD_CD_SK GROUP BY dt.YEAR, dc.OFNS_DESC ORDER BY
dt.YEAR, Arrest_Count DESC;
```

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

```
SELECT dl.ARREST_PRECINCT, dc.LAW_CAT_CD, COUNT(*) AS Arrest_Count

FROM Arrest_Fact af JOIN DIM_Location dl ON af.ARREST_PRECINCT_SK = dl.ARREST_PRECINCT_SK

JOIN DIM_Crime dc ON af.PD_CD_SK = dc.PD_CD_SK

WHERE dc.LAW_CAT_CD IN ('F', 'M') GROUP BY dl.ARREST_PRECINCT, dc.LAW_CAT_CD ORDER BY Arrest_Count;
```

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).)

```
SELECT dl.ARREST_BORO, COUNT(*) AS Arrest_Count FROM Arrest_Fact af JOIN DIM_Location dl ON
af.ARREST_PRECINCT_SK = dl.ARREST_PRECINCT_SK

GROUP BY dl.ARREST_BORO ORDER BY Arrest_Count DESC;
```

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

```
SELECT dp.AGE_GROUP, dp.PREP_RACE, dp.PREP_SEX, COUNT(*) AS Arrest_Count

FROM Arrest_Fact af JOIN DIM_Preperator dp ON af.PREP_ID_SK = dp.PREP_ID_SK

GROUP BY dp.AGE_GROUP, dp.PREP_RACE, dp.PREP_SEX ORDER BY Arrest_Count DESC;
```

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

```
SELECT dl.LATITUDE, dl.LONGITUDE, COUNT(*) AS Arrest_Count FROM Arrest_Fact af JOIN DIM_Location dl ON
af.ARREST_PRECINCT_SK = dl.ARREST_PRECINCT_SK GROUP BY dl.LATITUDE, dl.LONGITUDE ORDER BY Arrest_Count
DESC;
```