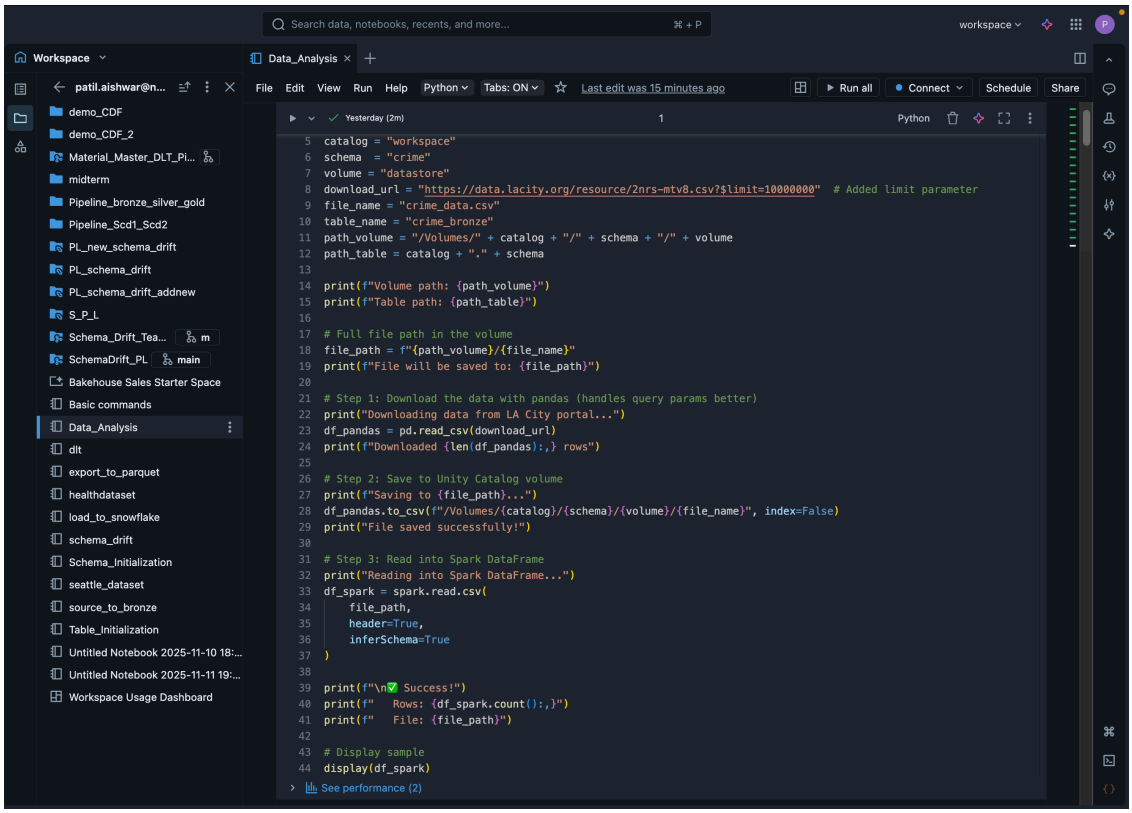
LA Crime Analysis Pipeline

Section 1: Data Ingestion from LA City Portal

Downloaded crime incident data from LA City's open data API and store it in Databricks Unity Catalog Volume
catalog = "workspace" schema = "crime" volume = "datastore"



Downloaded via pandas to handle API query parameters and download full dataset
Saved to Unity Catalog Volume - Persists CSV file to /Volumes/workspace/crime/datastore/crime_data.csv
Load into Spark DataFrame - Reads CSV into Spark with schema inference for distributed processing;
And then created Bronze Table - Registers data as workspace.crime.crime_bronze table

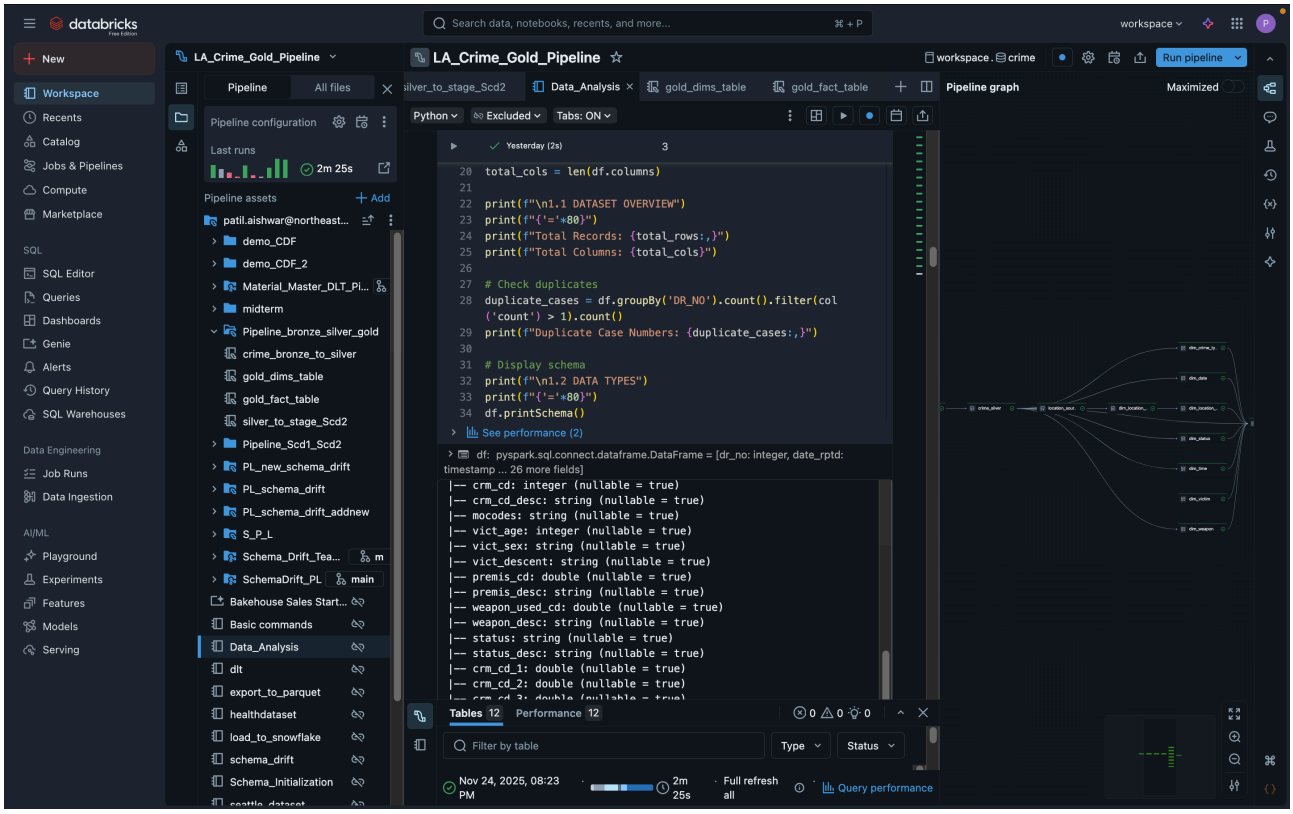
Expected Output:

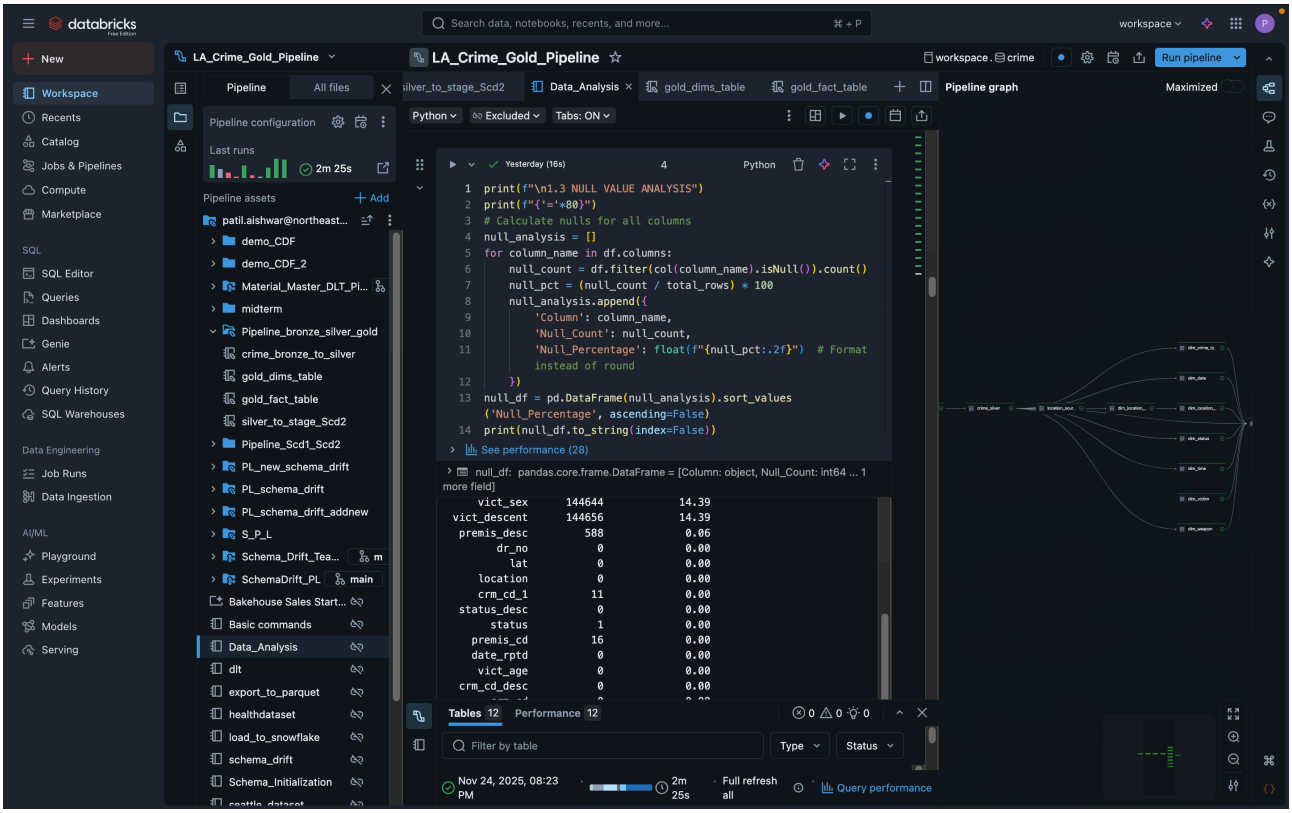
- Dataset: ~1 million crime incident records
- Columns: 28 attributes covering incident details, location, victim demographics, and case status

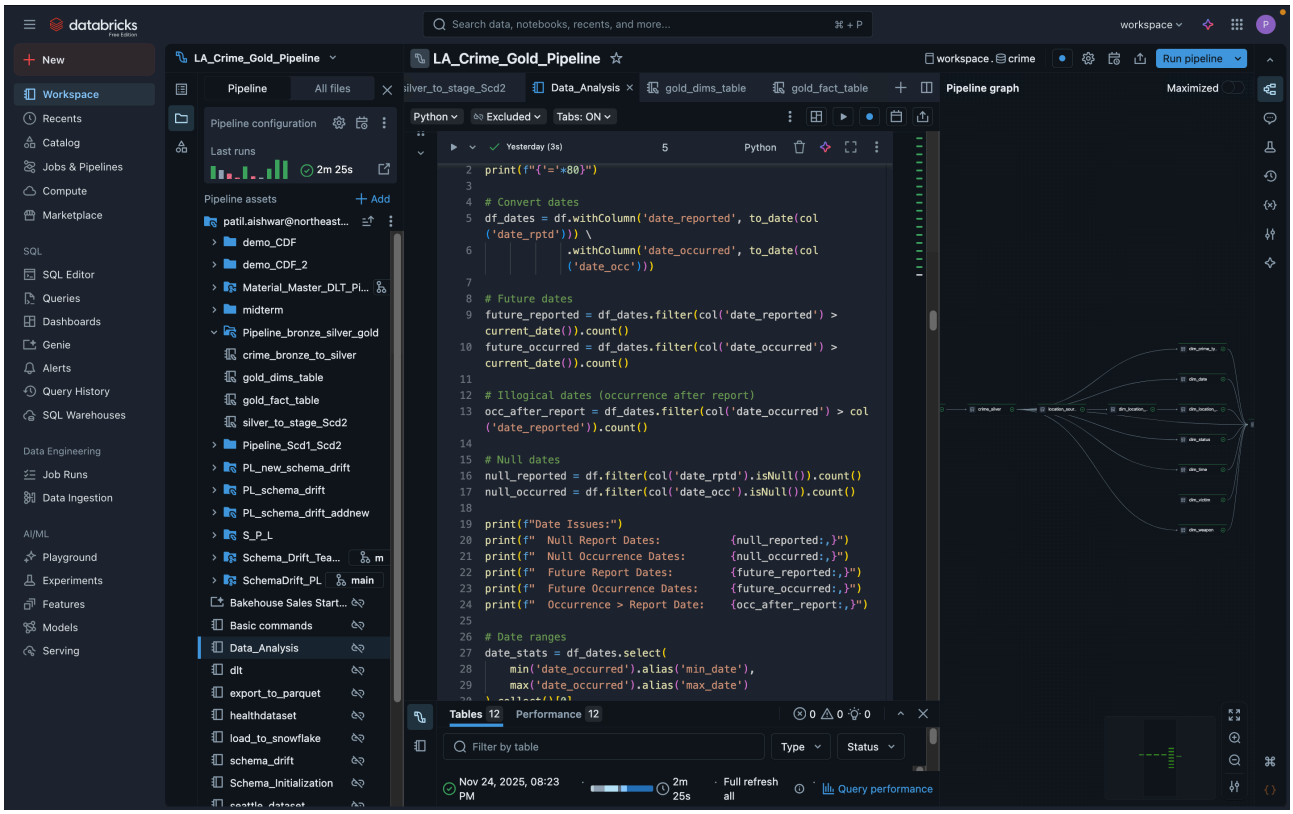
Dataset Overview

Checks Performed:

- Total row count and column count
- Primary key integrity (duplicate DR_NO detection)
- Schema validation (column names and data types)


databricks
Search data, notebooks, recents, and more...
workspace
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
AI/ML
Playground
Experiments
Features
Models
Serving
LA_Crime_Gold_Pipeline
Pipeline
All files
Pipeline configuration
Last runs
2m 25s
Pipeline assets
Add
patil.aishwar@northeast...
demo_CDF
demo_CDF_2
Material_Master_DLT_Pi...
midterm
Pipeline_bronze_silver_gold
crime_bronze_to_silver
gold_dims_table
gold_fact_table
silver_to_stage_Scd2
Pipeline_Scd1_Scd2
PL_new_schema_drift
PL_schema_drift
PL_schema_drift_addnew
S_P_L
Schema_Drift_Tea...
SchemaDrift_PL
main
Bakehouse Sales Start...
Basic commands
Data_Analysis
dlt
export_to_parquet
healthdataset
load_to_snowflake
schema_drift
Schema_Initialization
LA_Crime_Gold_Pipeline
workspace . crime
Run pipeline
silver_to_stage_Scd2
Data_Analysis
gold_dims_table
gold_fact_table
Pipeline graph
Maximized
Python
Excluded
Tabs: ON
Yesterday (2s)
3
20 total_cols = len(df.columns)
21
22 print(f"\n1.1 DATASET OVERVIEW")
23 print(f"{'='*80}")
24 print(f"Total Records: {total_rows:,}")
25 print(f"Total Columns: {total_cols}")
26
27 # Check duplicates
28 duplicate_cases = df.groupBy('DR_NO').count().filter(col
('count') > 1).count()
29 print(f"Duplicate Case Numbers: {duplicate_cases:,}")
30
31 # Display schema
32 print(f"\n1.2 DATA TYPES")
33 print(f"{'='*80}")
34 df.printSchema()
See performance (2)
df: pyspark.sql.connect.dataframe.DataFrame = [dr_no: integer, date_rptd:
timestamp ... 26 more fields]
|-- crm_cd: integer (nullable = true)
|-- crm_cd_desc: string (nullable = true)
|-- mocodes: string (nullable = true)
|-- vict_age: integer (nullable = true)
|-- vict_sex: string (nullable = true)
|-- vict_descent: string (nullable = true)
|-- premis_cd: double (nullable = true)
|-- premis_desc: string (nullable = true)
|-- weapon_used_cd: double (nullable = true)
|-- weapon_desc: string (nullable = true)
|-- status: string (nullable = true)
|-- status_desc: string (nullable = true)
|-- crm_cd_1: double (nullable = true)
|-- crm_cd_2: double (nullable = true)
Tables 12
Performance 12
Filter by table
Type
Status
Nov 24, 2025, 08:23 PM
2m 25s
Full refresh all
Query performance


databricks
Search data, notebooks, recents, and more...
workspace
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
AI/ML
Playground
Experiments
Features
Models
Serving
LA_Crime_Gold_Pipeline
Pipeline
All files
Pipeline configuration
Last runs
2m 25s
Pipeline assets
Add
patil.aishwar@northeast...
demo_CDF
demo_CDF_2
Material_Master_DLT_Pi...
midterm
Pipeline_bronze_silver_gold
crime_bronze_to_silver
gold_dims_table
gold_fact_table
silver_to_stage_Scd2
Pipeline_Scd1_Scd2
PL_new_schema_drift
PL_schema_drift
PL_schema_drift_addnew
S_P_L
Schema_Drift_Tea...
SchemaDrift_PL
main
Bakehouse Sales Start...
Basic commands
Data_Analysis
dlt
export_to_parquet
healthdataset
load_to_snowflake
schema_drift
Schema_Initialization
LA_Crime_Gold_Pipeline
workspace . crime
Run pipeline
silver_to_stage_Scd2
Data_Analysis
gold_dims_table
gold_fact_table
Pipeline graph
Maximized
Python
Excluded
Tabs: ON
Yesterday (16s)
4
Python
1 print(f"\n1.3 NULL VALUE ANALYSIS")
2 print(f"{'='*80}")
3 # Calculate nulls for all columns
4 null_analysis = []
5 for column_name in df.columns:
6 null_count = df.filter(col(column_name).isNull()).count()
7 null_pct = (null_count / total_rows) * 100
8 null_analysis.append({
9 'Column': column_name,
10 'Null_Count': null_count,
11 'Null_Percentage': float(f"{null_pct:.2f}") # Format
instead of round
12 })
13 null_df = pd.DataFrame(null_analysis).sort_values
('Null_Percentage', ascending=False)
14 print(null_df.to_string(index=False))
See performance (28)
null_df: pandas.core.frame.DataFrame = [Column: object, Null_Count: int64 ... 1
more field]
vict_sex 144644 14.39
vict_descent 144656 14.39
premis_desc 588 0.06
dr_no 0 0.00
lat 0 0.00
location 0 0.00
crm_cd_1 11 0.00
status_desc 0 0.00
status 1 0.00
premis_cd 16 0.00
date_rptd 0 0.00
vict_age 0 0.00
crm_cd_desc 0 0.00
Tables 12
Performance 12
Filter by table
Type
Status
Nov 24, 2025, 08:23 PM
2m 25s
Full refresh all
Query performance



Data Types Analysis

Used Spark's printSchema() to display column names, data types, and nullable status.

Key Observations:

- Date fields (date_rptd, date_occ) - String format requiring conversion
- Time field (time_occ) - Integer in military time format (0-2359)
- Geographic fields (lat, lon) - Double precision required for mapping
- Demographic fields (vict_age, vict_sex, vict_descent) - Mixed types requiring standardization

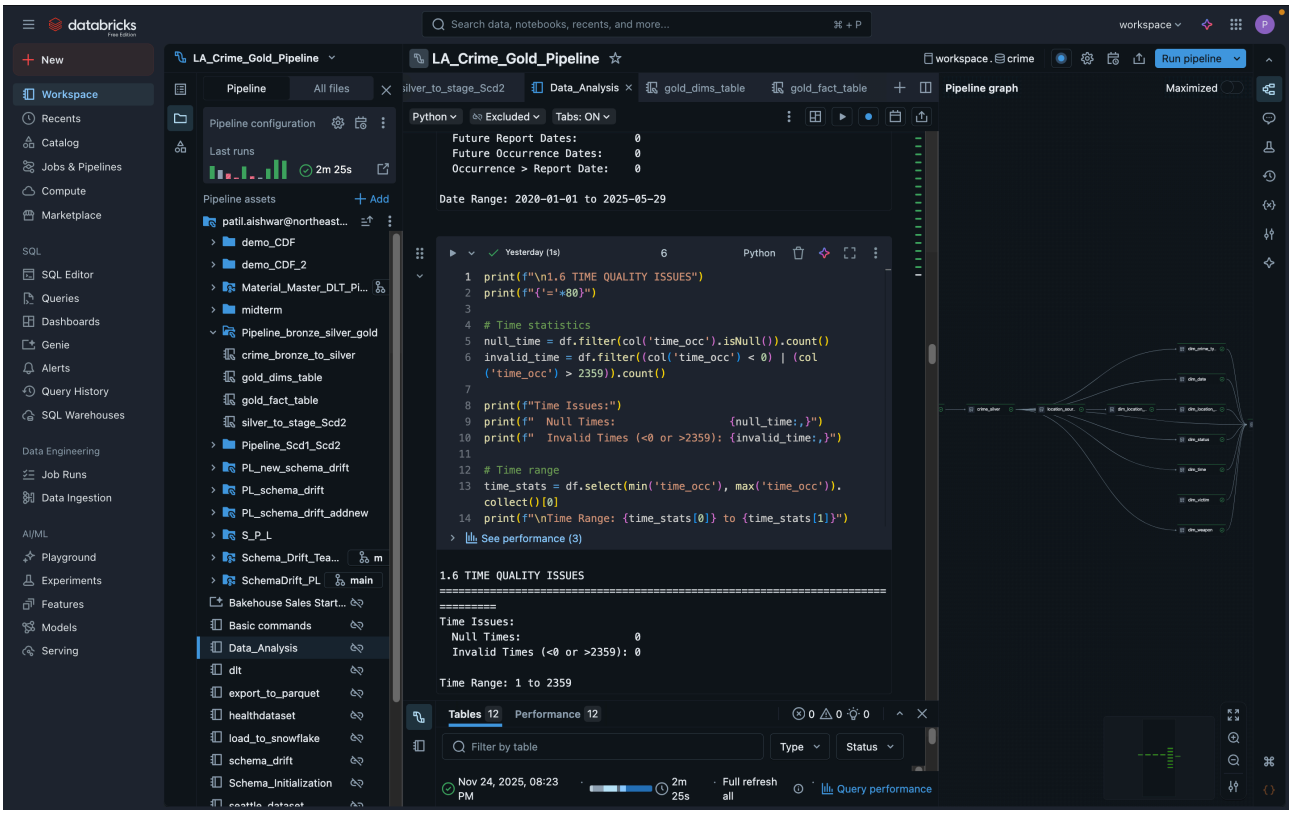
NULL Value Analysis

- Iterates through all 28 columns
- Counts NULL values per column
- Calculates NULL percentage relative to total rows
- Sorts by NULL percentage (descending)

Date Quality Issues

Validation Checks:

1. NULL Date Detection: Missing date_rptd or date_occ values
2. Future Date Detection: Crimes reported/occurred in the future (impossible - data error)
3. Logical Timeline Validation: date_occurred ≤ date_reported (crime can't be reported before it happened)



Time Quality Issues
Validation Rules:

- Format: Military time (24-hour format without colon)
- Valid Range: 0000 to 2359
- NULL Times: Missing time_occ values
- Invalid Times: Values < 0 or > 2359 (impossible times)

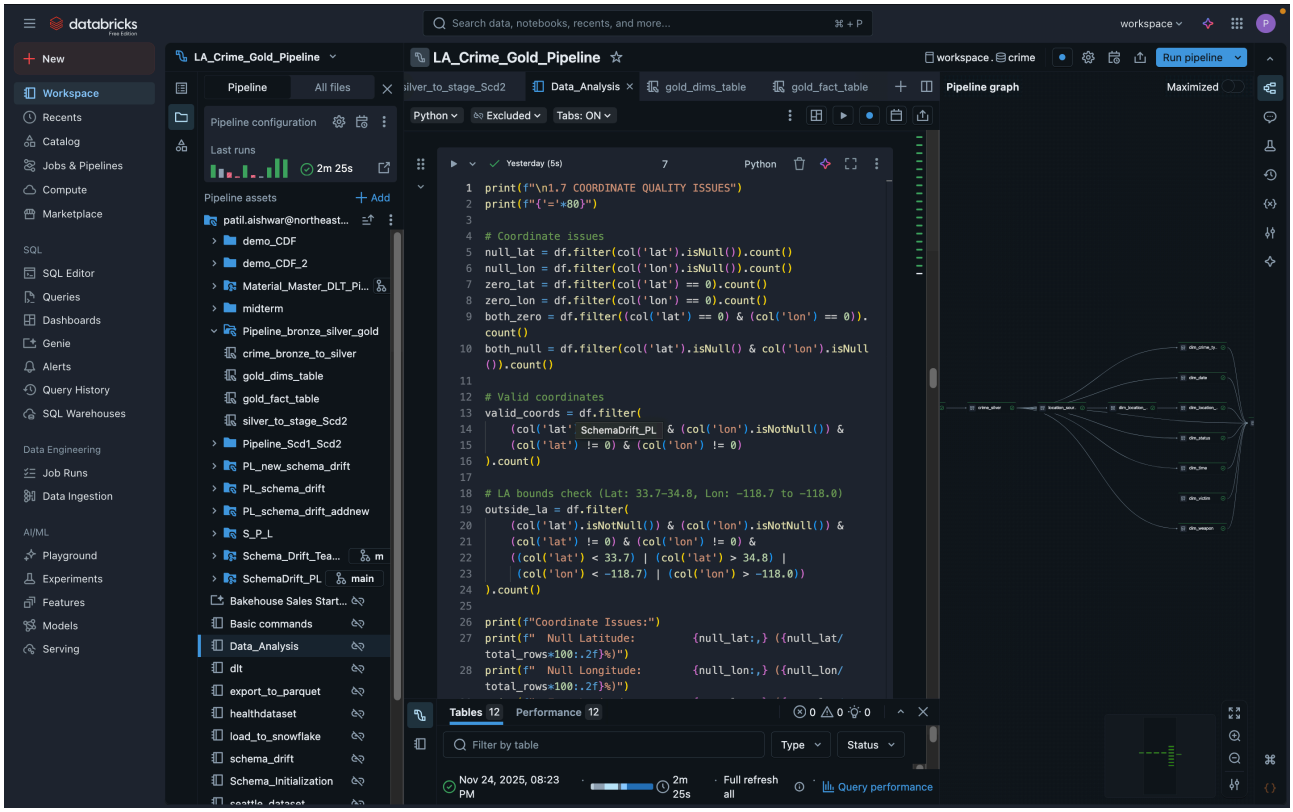
Geographic Data Quality Issues

LA Geographic Boundaries:

- Latitude Range: 33.7°N to 34.8°N
- Longitude Range: -118.7°W to -118.0°W

Quality Issues Detected:

1. NULL Coordinates: Missing latitude/longitude values
2. Zero Coordinates (0, 0): Placeholder values indicating missing location data (points to Gulf of Guinea, Africa)

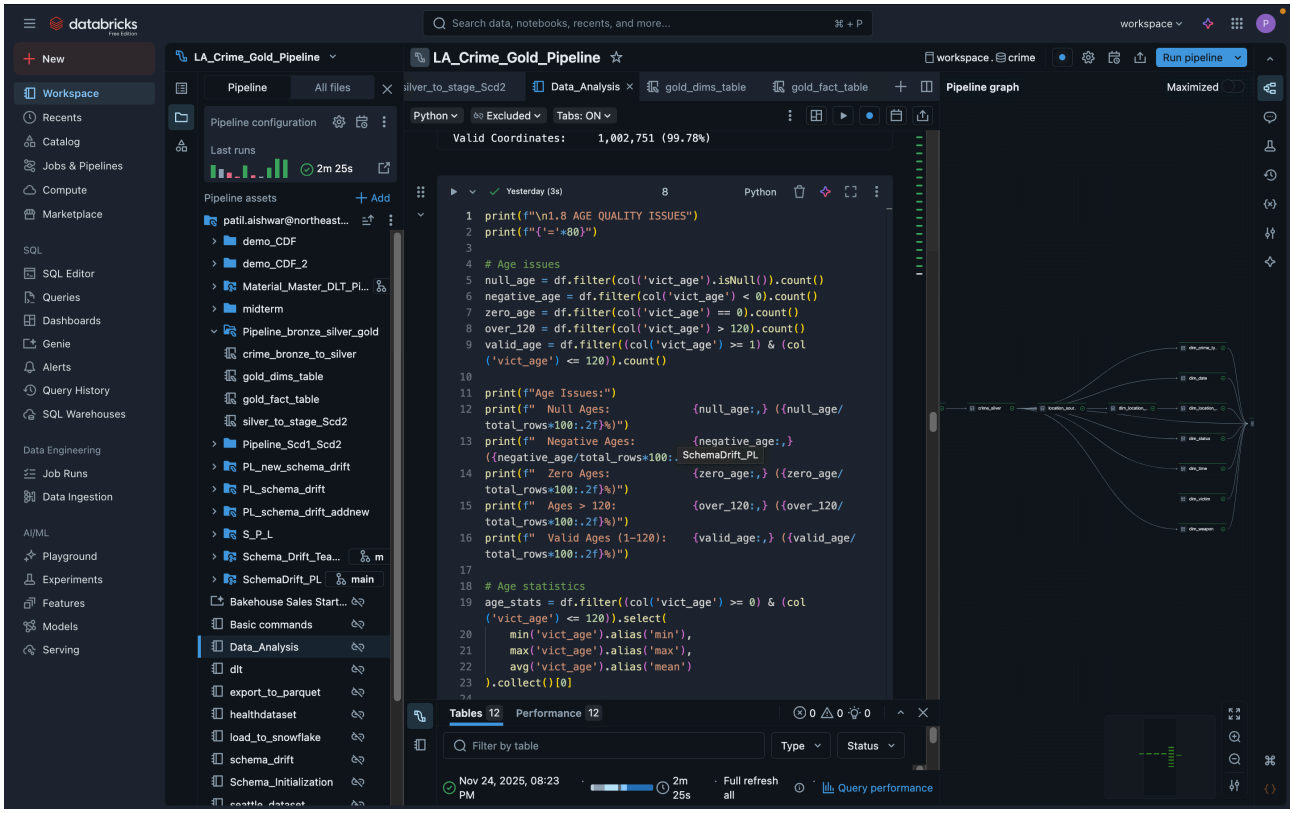


Victim Age Quality Issues

Valid Age Range: 1 to 120 years

Invalid Patterns Detected:

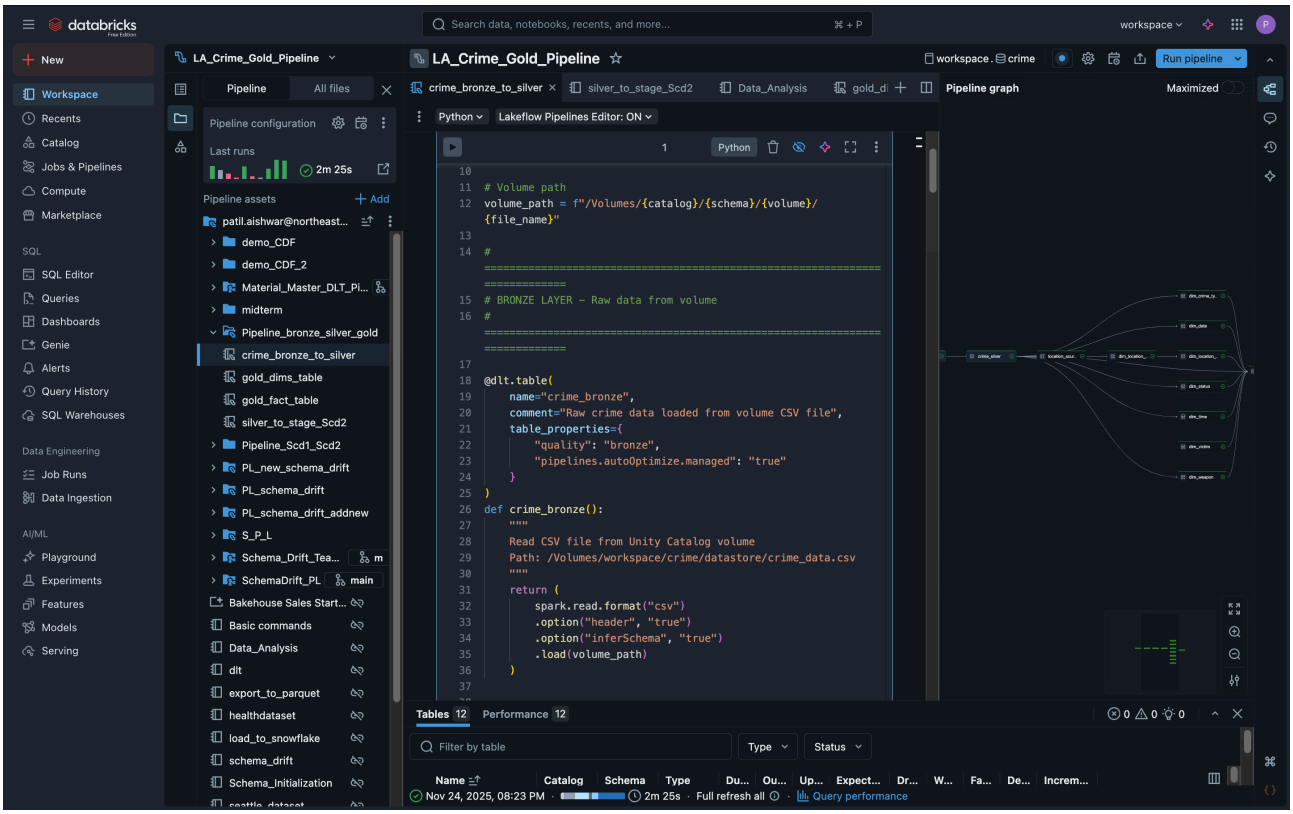
1. Negative Ages: Data entry errors (impossible values)
2. Zero Ages: Placeholder for unknown victim age
3. NULL Ages: Missing demographic data
4. Ages > 120: Unrealistic values (data errors or placeholder codes)



Bronze Layer: Ingesting raw CSV data from Unity Catalog Volume into DLT-managed table
Silver Layer: Data cleaning and transforming

Create a DLT-managed table that reads raw CSV data from Unity Catalog Volume with no transformations.

Source: /Volumes/workspace/crime/datastore/crime_data.csv
Target: workspace.crime.crime_bronze (DLT-managed table)



Expected Output

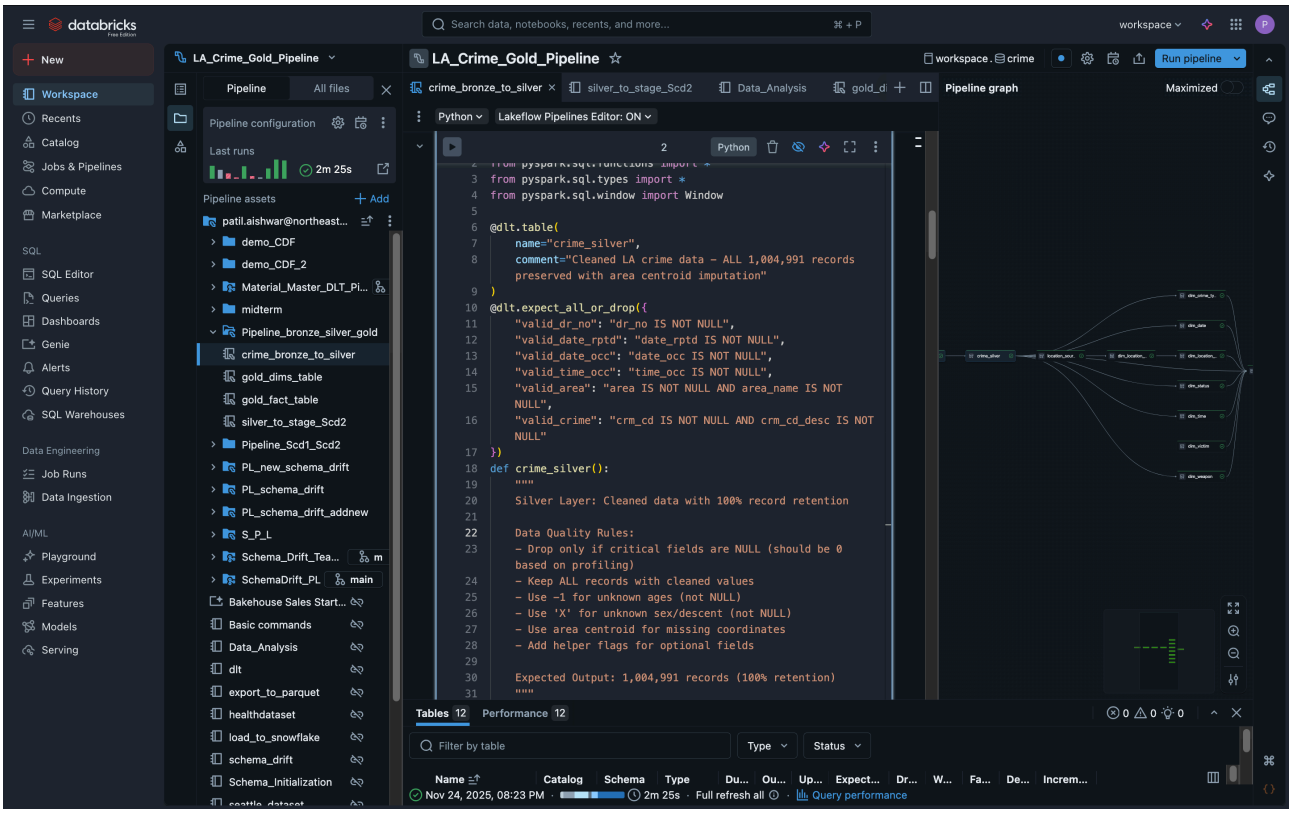
- Records: 1,004,991 crime incidents (all records from CSV)
- Columns: 28 original attributes
- Quality: Raw data, no transformations applied
- Table Type: DLT-managed Delta table

Silver Layer - Data Cleaning & Transformation

Transform raw Bronze data into clean, analysis-ready Silver layer with:

- 100% record retention - No data loss
- Standardized values - Consistent encoding for unknown/missing data
- Geographic imputation - Area centroid assignment for missing coordinates
- Quality flags - Enable flexible filtering in Gold layer

Result: Clean, unified dataset ready for dimensional modeling with NO data loss.

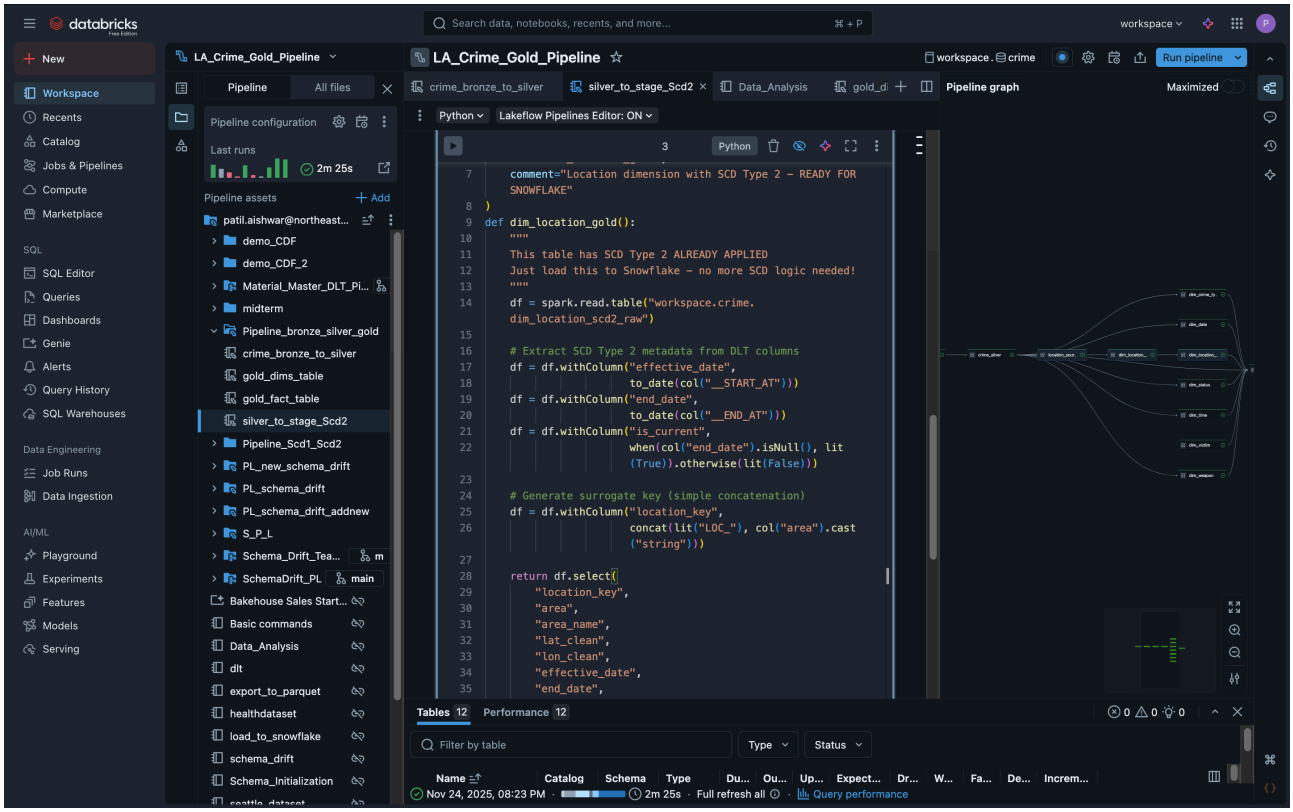


SCD Type 2 for Location

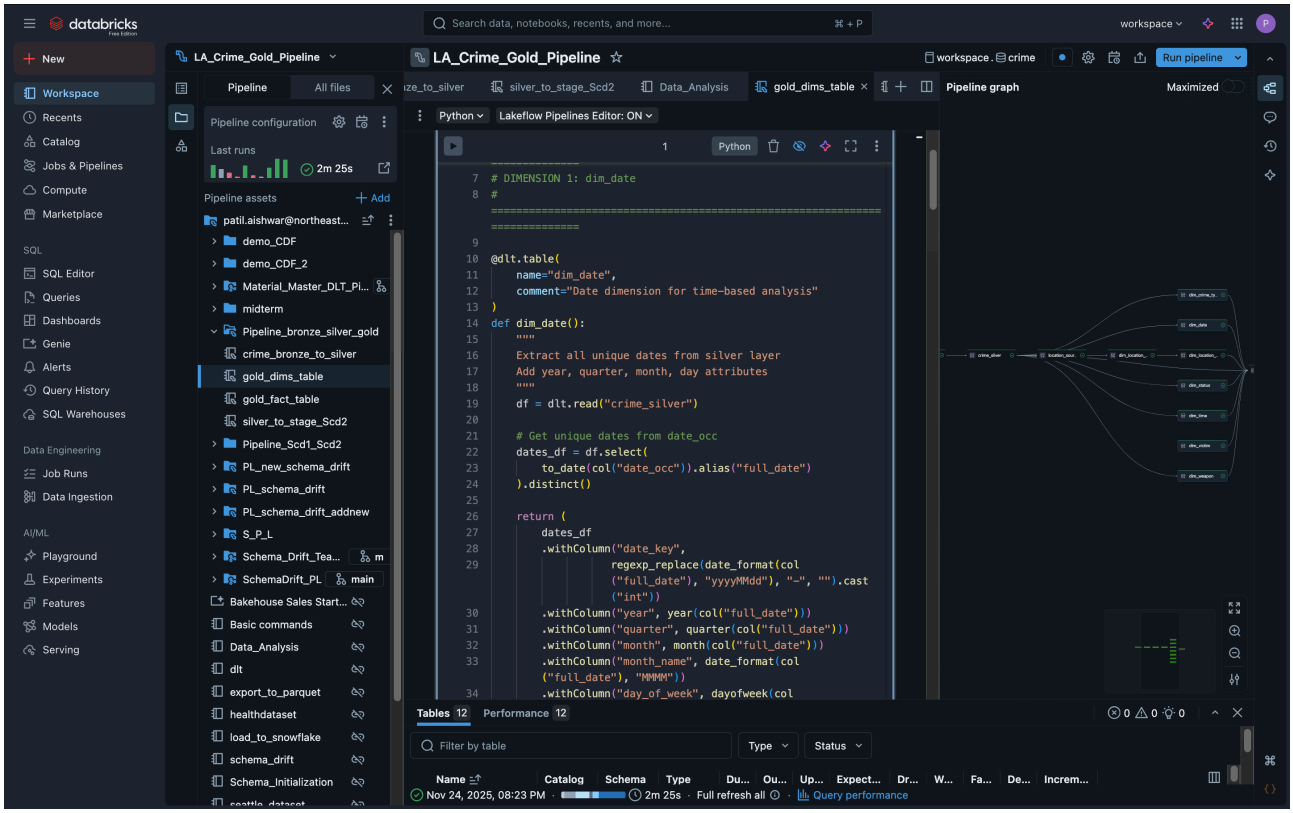
We felt that Location data can change over time in the LA crime dataset, so that's why we thought sad 2 was needed here

SCD Type 2 preserves this history by:

- Creating a new record when location attributes change (new lat/lon or area_name)
- Marking old records as is_current = False with an end_date



Gold Layer Dimensions: Loaded all the remaining six dimension apart from location



Dimension Tables

1. dim_date - Date Dimension
Purpose: Enable time-series analysis for crime trends (yearly, quarterly, monthly, daily patterns)
Source: Unique date_occ values from Silver layer
Grain: One row per unique date
Key Attributes:

- date_key (PK) - Integer format: YYYYMMDD (e.g., 20230615)
- full_date - Full date value
- year, quarter, month, month_name - Hierarchical attributes
- day_of_week (1=Sunday, 7=Saturday), day_name - Day analysis
- created_by, created_dt - Metadata

2. dim_time - Time of Day Dimension
Purpose: Analyze crime patterns by time of day and time periods
Source: Unique time_occ values from Silver layer
Grain: One row per unique military time value (0000-2359)
Key Attributes:

- time_key (PK) - Integer: military time (e.g., 1430 = 2:30 PM)
- time_occ - Original time value
- time_period - Classification: Morning (6:00-11:59), Afternoon (12:00-16:59), Evening (17:00-20:59), Night (21:00-5:59)
- created_by, created_dt - Metadata

3. dim_crime_type - Crime Classification Dimension
Purpose: Categorize crimes by type code and FBI classification
Source: Unique crm_cd and crm_cd_desc from Silver layer
Grain: One row per unique crime code
Key Attributes:

- crime_type_key (PK) - Same as crm_cd (crime code)
- crm_cd - Crime code (e.g., 624 = Battery - Simple Assault)
- crm_cd_desc - Crime description
- part_1_2 - FBI Uniform Crime Reporting classification:
  - 1 = Part I crimes (serious: murder, rape, robbery, burglary, etc.)
  - 2 = Part II crimes (less serious: simple assault, vandalism, etc.)
- created_by, created_dt - Metadata

4. dim_weapon - Weapon Classification Dimension
Purpose: Track weapons used in crimes
Source: Unique weapon_used_cd_clean and weapon_desc_clean from Silver layer
Grain: One row per unique weapon code
Key Attributes:

- weapon_key (PK) - Same as weapon_used_cd
- weapon_used_cd - Weapon code (0 = no weapon, 400 = firearm, etc.)
- weapon_desc - Weapon description (cleaned with "NO WEAPON/UNKNOWN" for nulls)
- created_by, created_dt - Metadata

5. dim_victim - Victim Demographics Dimension
Purpose: Analyze crime patterns by victim age groups and sex
Source: Pre-generated combinations of age groups and sex categories
Grain: One row per unique age_group + sex combination
Key Attributes:

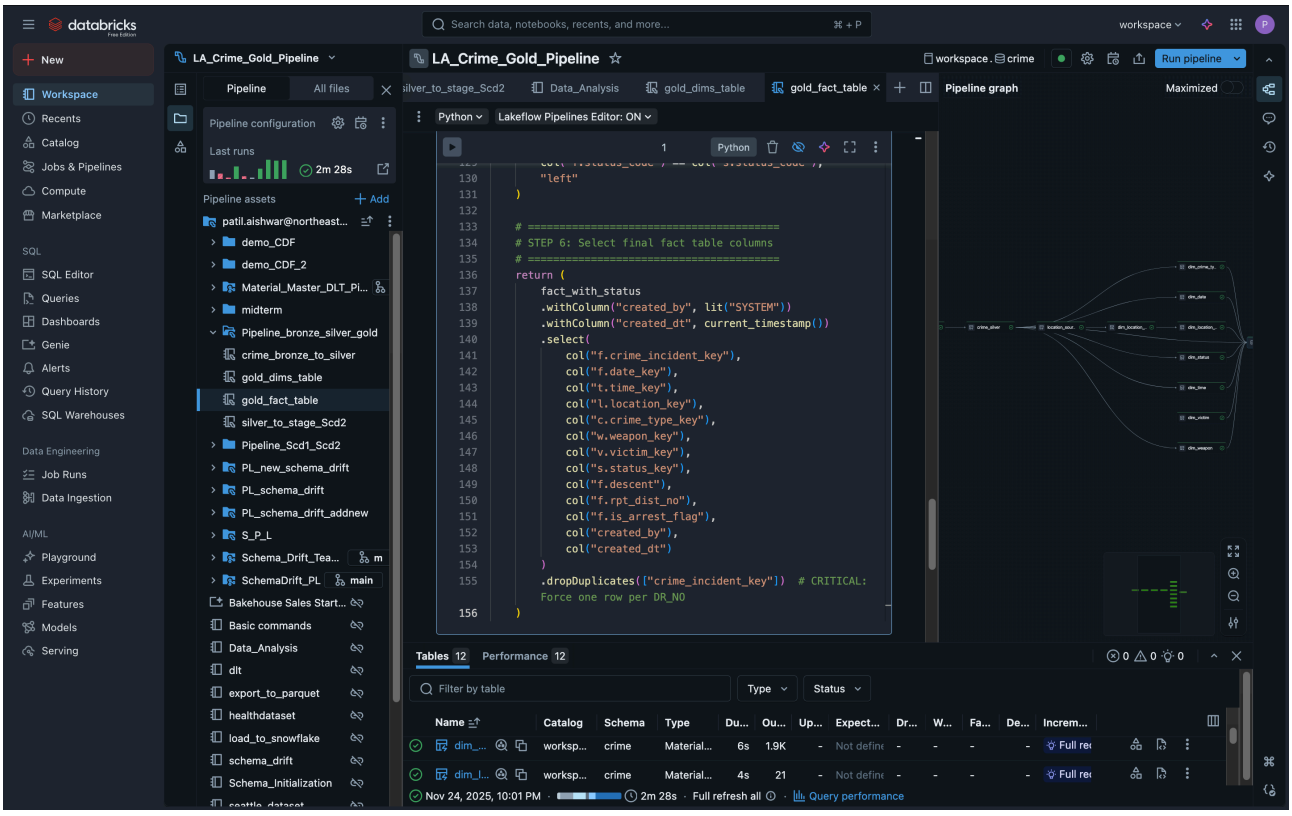
- victim_key (PK) - Sequential integer (1-18)

- age_group - Categorical: Unknown, 0-17 (Juvenile), 18-25 (Young Adult), 26-35 (Adult), 36-50 (Middle Age), 51+ (Senior)
- sex - M (Male), F (Female), X (Unknown)
- created_by, created_dt - Metadata

6. dim_status - Case Status Dimension
Purpose: Track investigation and arrest status
Source: Unique status_clean and status_desc from Silver layer
Grain: One row per unique status code
Key Attributes:

- status_key (PK) - Sequential integer (auto-generated)
- status_code - Status code (IC, AA, JA, JO, etc.)
- status_desc - Status description
    - IC = Investigation Continues (active case)
    - AA = Adult Arrest
    - JA = Juvenile Arrest
    - JO = Juvenile Other
- created_by, created_dt - Metadata

Gold Layer Fact Table



Fact Table Design

Grain: One row per crime incident (uniquely identified by DR_NO)

Fact Table Type
Transaction Fact Table - Records individual crime events at the most granular level

Fact Table Schema

Dimension Foreign Keys (7)

Column	Type	References
date_key	Integer	dim_date.date_key
time_key	Integer	dim_time.time_key
location_key	String	dim_location.location_key
crime_type_key	Integer	dim_crime_type.crime_type_key
weapon_key	Integer	dim_weapon.weapon_key
victim_key	Integer	dim_victim.victim_key
status_key	Integer	dim_status.status_key

Degenerate Dimensions (2)

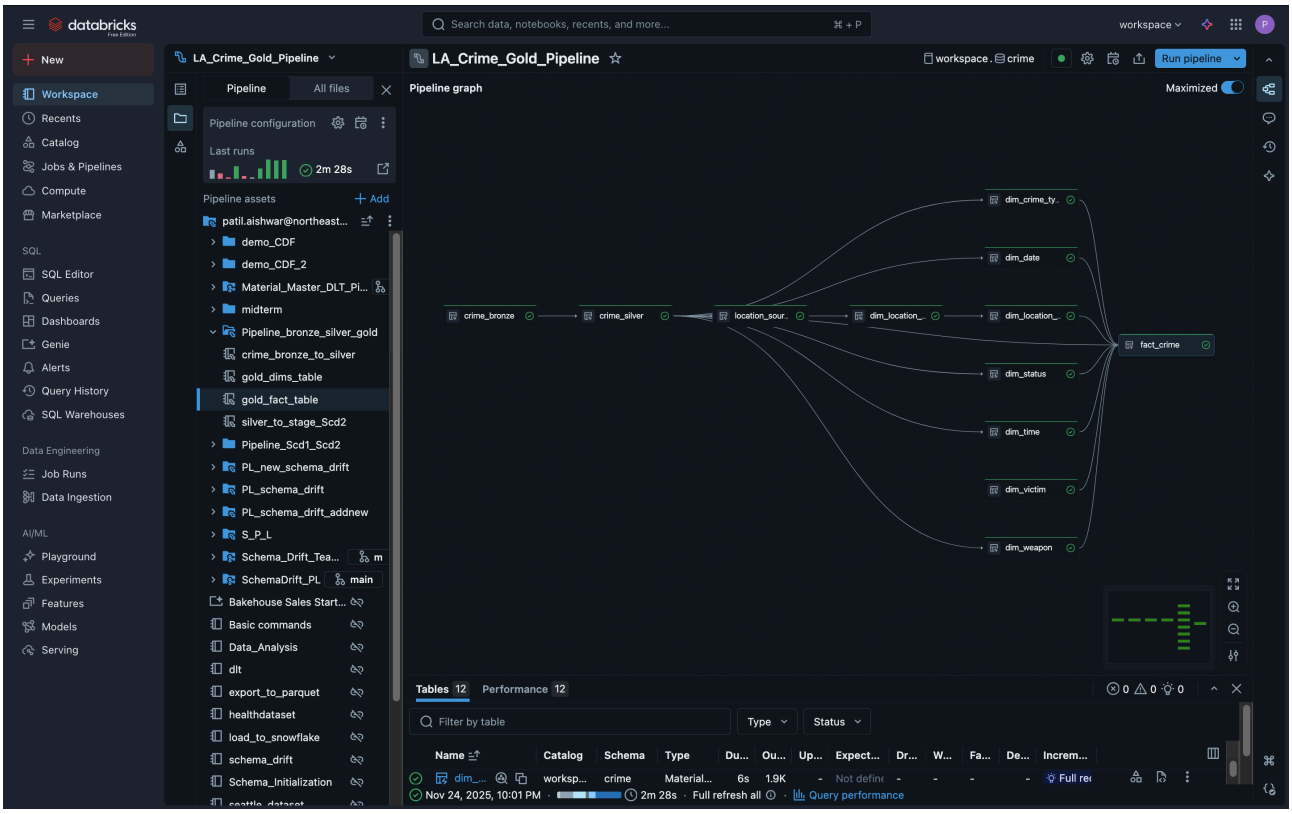
Column	Type
crime_incident_key	String
descent	String
rpt_dist_no	Integer

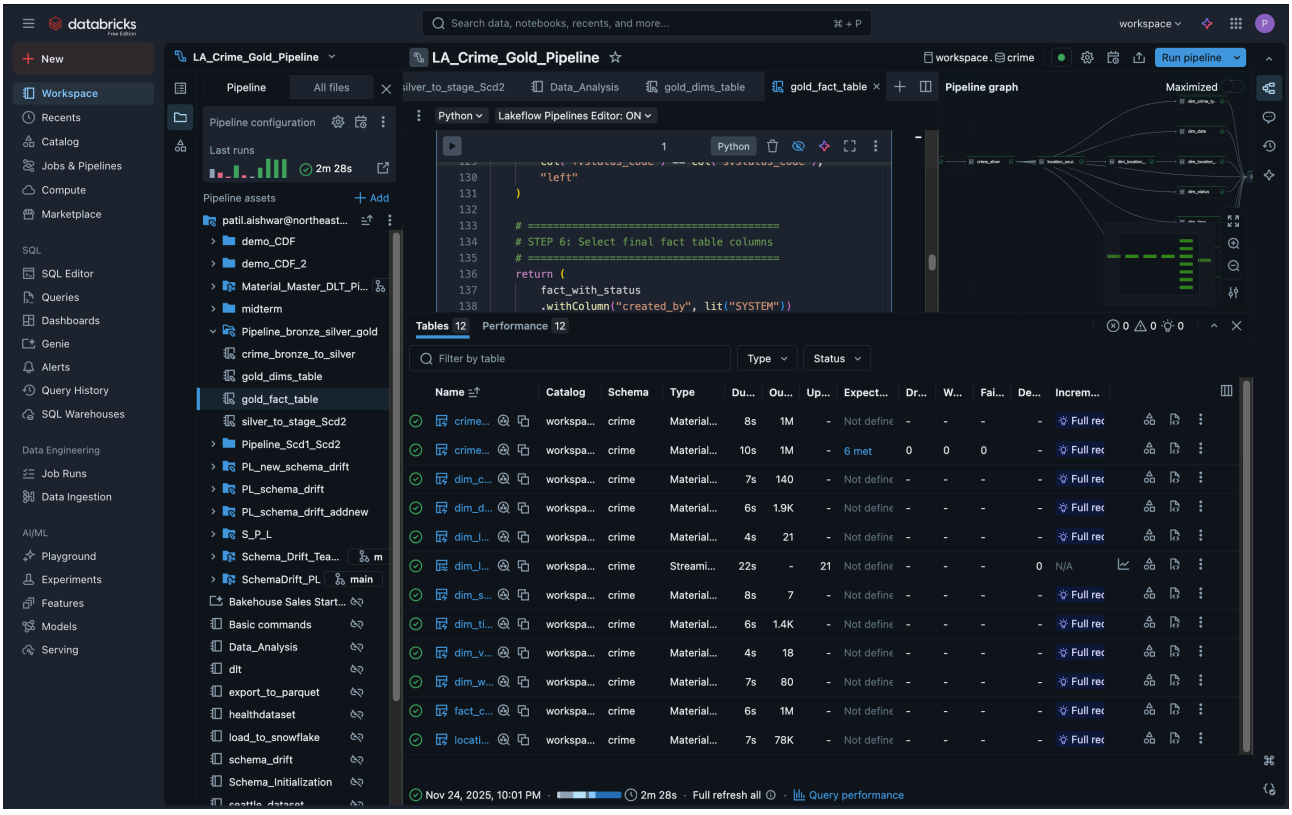
Fact Measures (1)

Column	Type	Purpose
is_arrest_flag	Boolean	Whether crime resulted in arrest (AA or JA status)

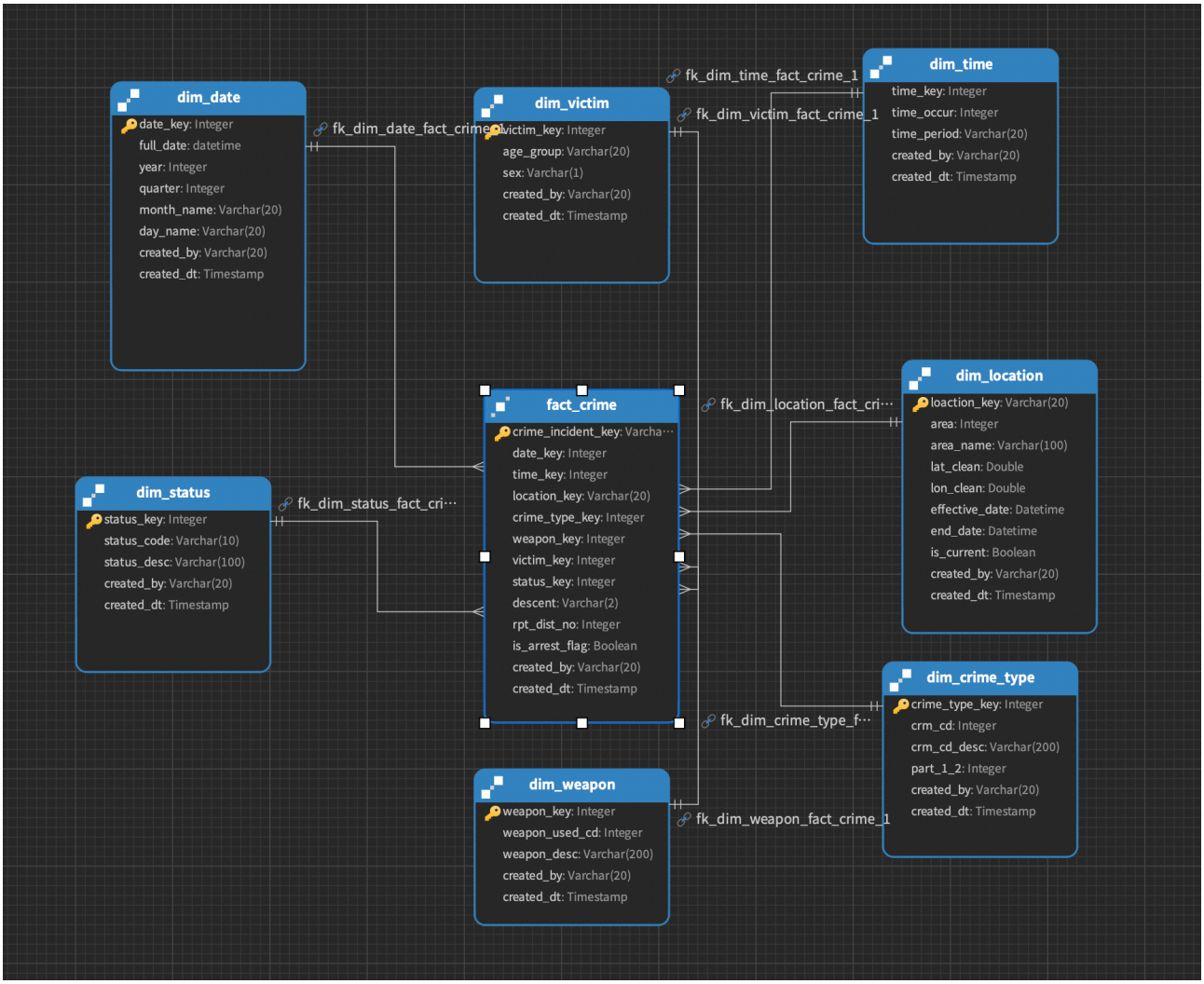
Metadata (2)

Column	Type	Purpose
created_by	String	"SYSTEM" - audit trail
created_dt	Timestamp	When record was loaded


databricks
Search data, notebooks, recents, and more...
workspace
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
AI/ML
Playground
Experiments
Features
Models
Serving
LA_Crime_Gold_Pipeline
Pipeline
All files
Pipeline configuration
Last runs
2m 28s
Pipeline assets
Add
patil.aishwar@northeast...
demo_CDF
demo_CDF_2
Material_Master_DLT_Pi...
midterm
Pipeline_bronze_silver_gold
crime_bronze_to_silver
gold_dims_table
gold_fact_table
silver_to_stage_Scd2
Pipeline_Scd1_Scd2
PL_new_schema_drift
PL_schema_drift
PL_schema_drift_addnew
S_P_L
Schema_Drift_Tea...
SchemaDrift_PL
Bakehouse Sales Start...
Basic commands
Data_Analysis
dlt
export_to_parquet
healthdataset
load_to_snowflake
schema_drift
Schema_Initialization
LA_Crime_Gold_Pipeline
workspace . crime
Run pipeline
Pipeline graph
Maximized
crime_bronze
crime_silver
location_sour...
dim_location_...
dim_crime_ty...
dim_date
dim_location_...
fact_crime
dim_status
dim_time
dim_victim
dim_weapon
Tables 12
Performance 12
Filter by table
Type
Status
Name
Catalog
Schema
Type
Du...
Ou...
Up...
Expect...
Dr...
W...
Fa...
De...
Increm...
dim_...
worksp...
crime
Material...
6s
1.9K
Not define
Full re
Nov 24, 2025, 10:01 PM
2m 28s
Full refresh all
Query performance


databricks
Search data, notebooks, recents, and more...
workspace
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
AI/ML
Playground
Experiments
Features
Models
Serving
LA_Crime_Gold_Pipeline
Pipeline
All files
Pipeline configuration
Last runs
2m 28s
Pipeline assets
Add
patil.aishwar@northeast...
demo_CDF
demo_CDF_2
Material_Master_DLT_Pi...
midterm
Pipeline_bronze_silver_gold
crime_bronze_to_silver
gold_dims_table
gold_fact_table
silver_to_stage_Scd2
Pipeline_Scd1_Scd2
PL_new_schema_drift
PL_schema_drift
PL_schema_drift_addnew
S_P_L
Schema_Drift_Tea...
SchemaDrift_PL
Bakehouse Sales Start...
Basic commands
Data_Analysis
dlt
export_to_parquet
healthdataset
load_to_snowflake
schema_drift
Schema_Initialization
LA_Crime_Gold_Pipeline
workspace . crime
Run pipeline
Data_Analysis
gold_dims_table
gold_fact_table
Pipeline graph
Maximized
Python
Lakeflow Pipelines Editor: ON
"left"
STEP 6: Select final fact table columns
return (
fact_with_status
.withColumn("created_by", lit("SYSTEM"))
Tables 12
Performance 12
Filter by table
Type
Status
Name
Catalog
Schema
Type
Du...
Ou...
Up...
Expect...
Dr...
W...
Fai...
De...
Increm...
crime...
workspa...
crime
Material...
8s
1M
Not define
Full rec
crime...
workspa...
crime
Material...
10s
1M
6 met
0
0
0
Full rec
dim_c...
workspa...
crime
Material...
7s
140
Not define
Full rec
dim_d...
workspa...
crime
Material...
6s
1.9K
Not define
Full rec
dim_l...
workspa...
crime
Material...
4s
21
Not define
Full rec
dim_l...
workspa...
crime
Streami...
22s
21
Not define
0
N/A
dim_s...
workspa...
crime
Material...
8s
7
Not define
Full rec
dim_ti...
workspa...
crime
Material...
6s
1.4K
Not define
Full rec
dim_v...
workspa...
crime
Material...
4s
18
Not define
Full rec
dim_w...
workspa...
crime
Material...
7s
80
Not define
Full rec
fact_c...
workspa...
crime
Material...
6s
1M
Not define
Full rec
locati...
workspa...
crime
Material...
7s
78K
Not define
Full rec
Nov 24, 2025, 10:01 PM
2m 28s
Full refresh all
Query performance

This is the revised dimension model that we created

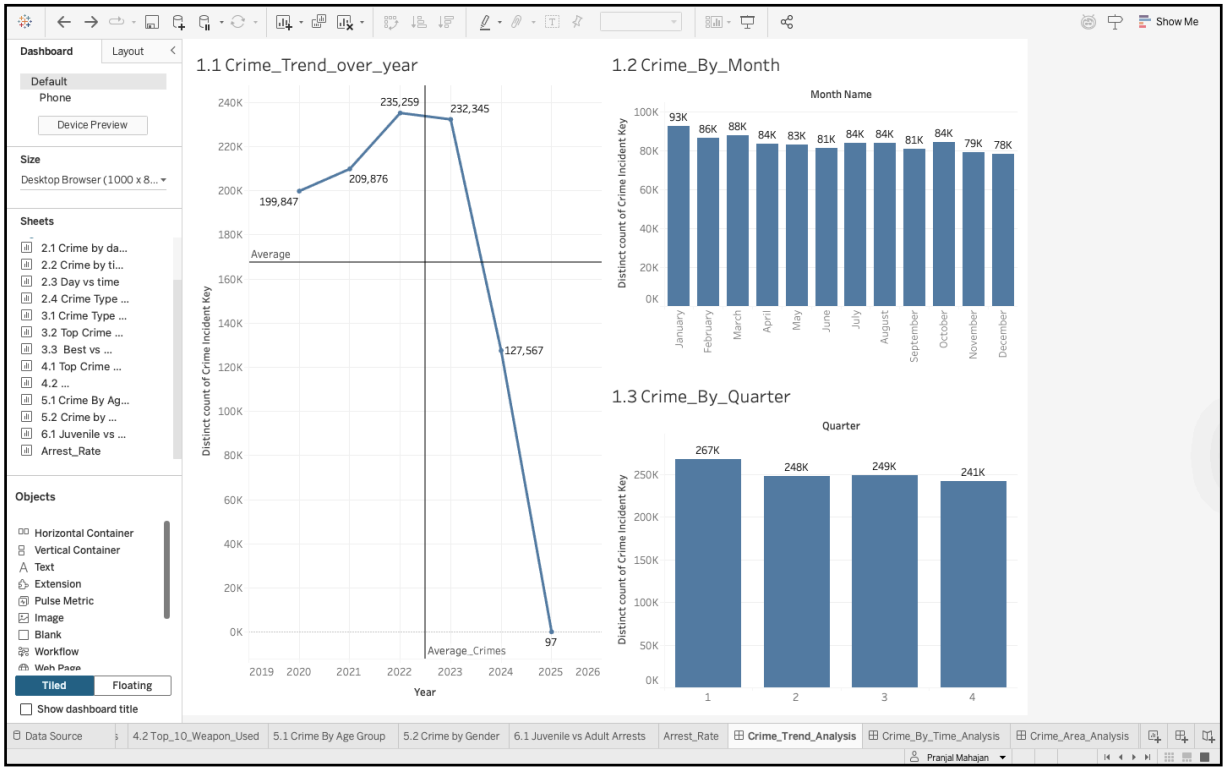


We established a connection between Tableau and Databricks, imported all dimension and fact tables, configured the dimensional relationships, and developed visualizations to address the following business requirements:

Business Questions:

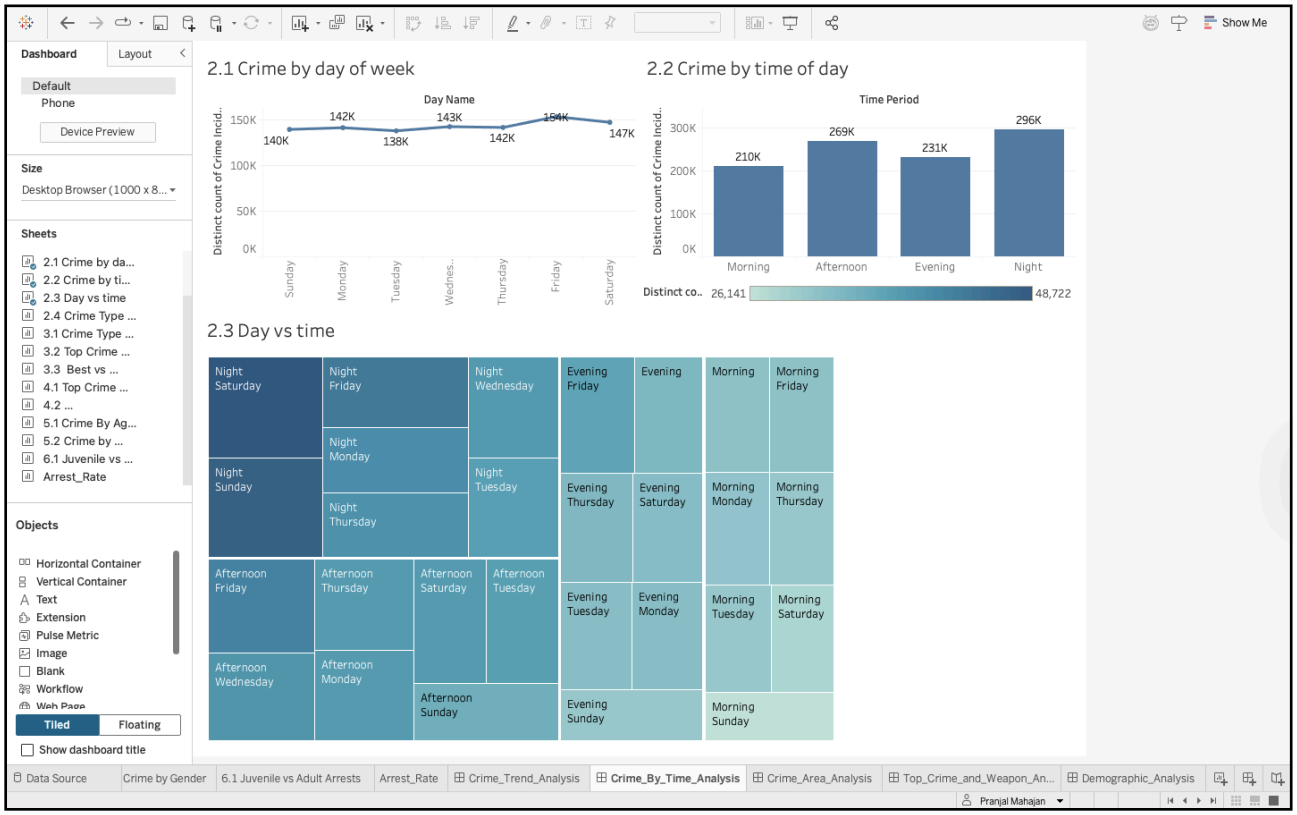
Crime Rates Over Time:

- **What is the overall trend in crime rates over the years?**
- **How have crime rates changed on a monthly basis?**
- **How have crime rates changed on a quarterly basis?**



Day Time and Week Factors:

- **Is there a correlation between the day of the week and the number of reported crimes?**
- **Do certain types of crimes tend to occur at specific times of the day?**

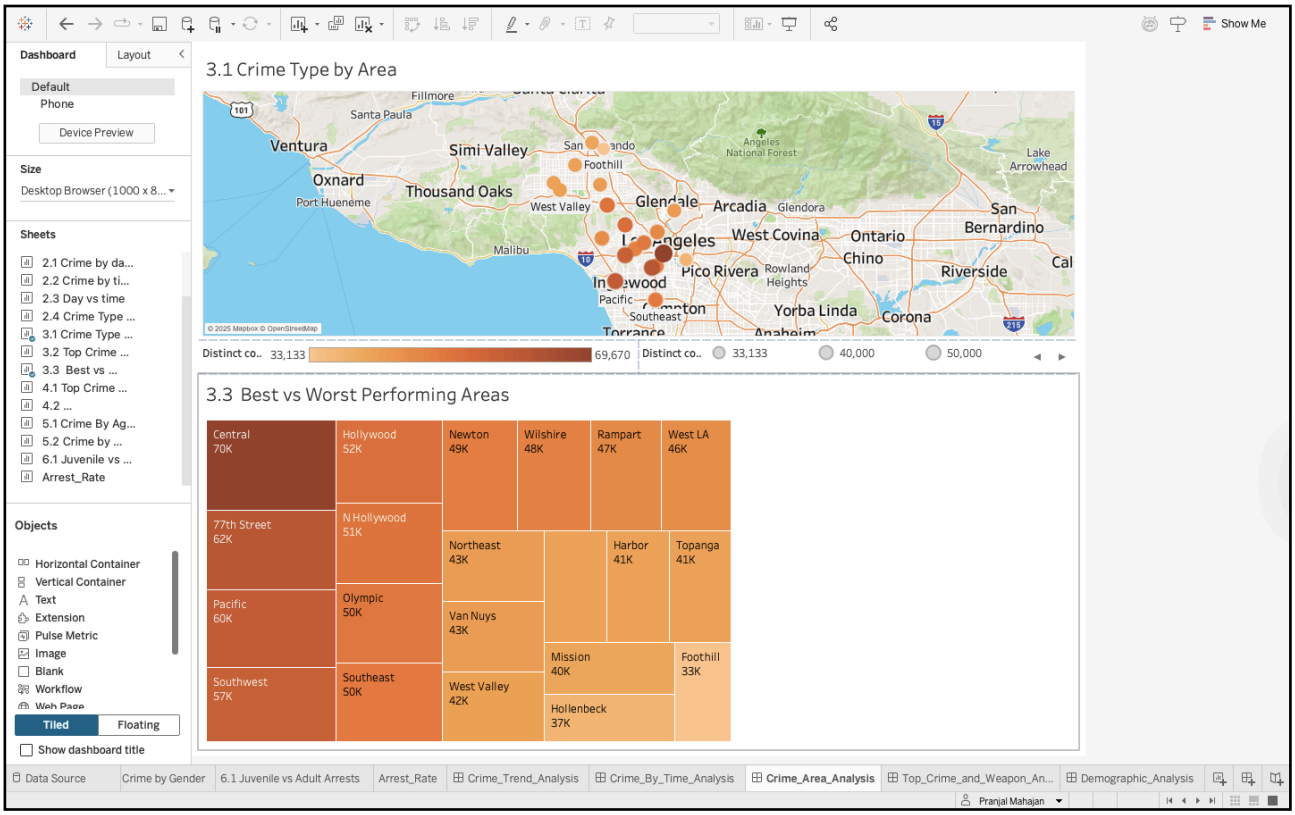


Crime by Location:

- **Where are the high-crime areas in Los Angeles?**

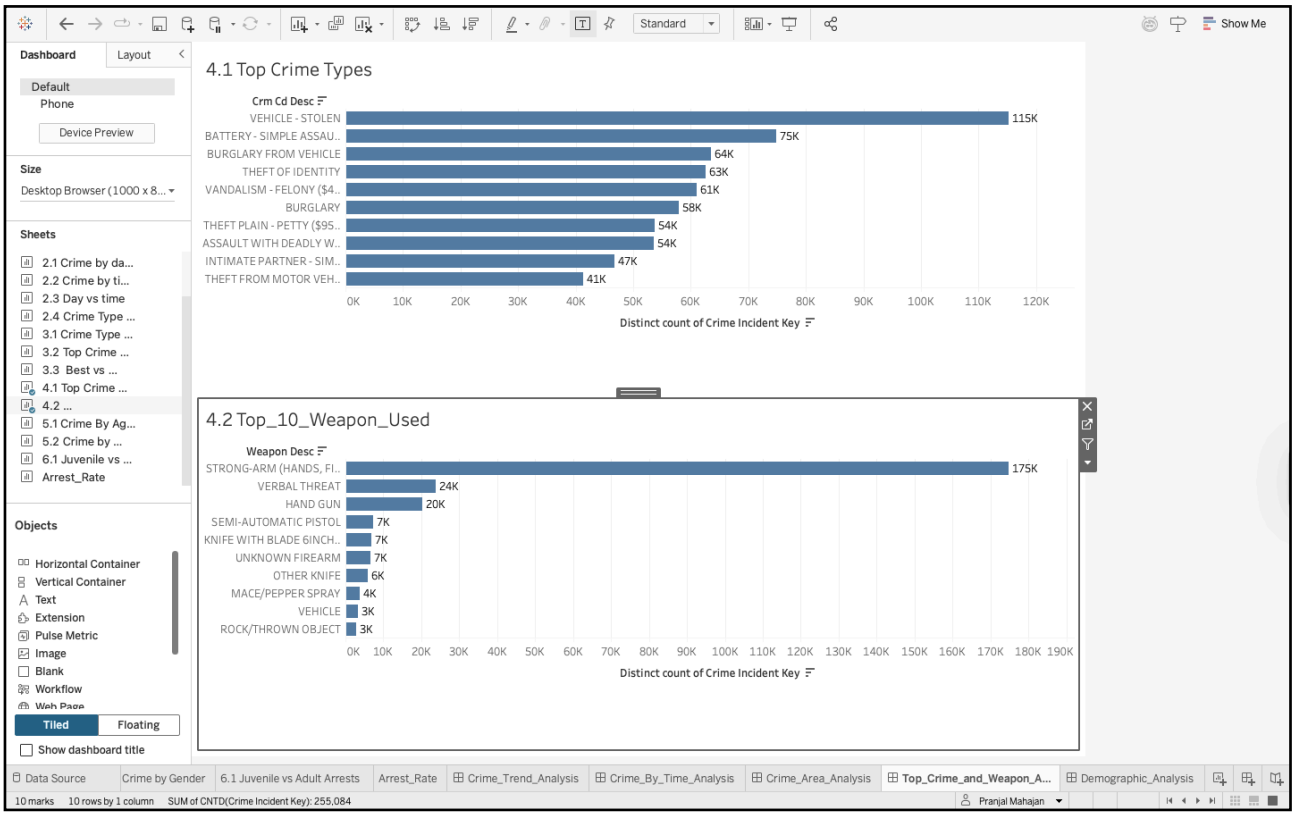
We want to know (your inferences on):

- **By seeing the geo map, what are the crime hotspots?**
- **Can you identify areas where Los Angeles performs better or worse in terms of crime rates?**



Types of Crime:

- **What are the most commonly used weapons in reported crimes?**



Demographic Analysis:

• Show in the visualization the age patterns in crime.

• Show in the visualization the gender-related patterns in crime.

Arrests Ratio:

• What percentage of reported crimes result in Juvenile, Adult arrests?