

Contents:

1. Overview of the project
2. About the Datasets
3. Profiling - Inferences, and Staging
4. Data Cleaning, Transformations, and Intermediate Staging.
5. ER Modeling, Fact and Dimensional Loading
6. Visualizations

Overview of the Project:

This final project on Motor Collisions - focuses on data modeling, data engineering, analysis with the help of visualizations pipeline using the datasets provided (Austin, Chicago, Montgomery, NYC).

Project Objectives

1. Analyze the given data sets thoroughly
2. Design and implement an ETL (Extract, Transform, Load) pipeline
3. Create a dimensional model for efficient data storage and retrieval
4. Develop visualizations to showcase findings
5. Present a narrative that tells the story hidden within the data

This project serves as a capstone, allowing for the application of skills acquired throughout the course in a real-world scenario. It provides an opportunity to demonstrate proficiency in data handling, analysis, and presentation, showcasing of abilities in the field of data engineering and analytics.

About the Datasets

Each of the datasets has been obtained from the government websites and profiling has been done individually for each one of them, and the details of the same are presented in the profiling document. The following section covers only the key details about the datasets.

1. Austin Dataset:

No. of Columns: 43

No. of Rows: 212,834

Source: https://data.austintexas.gov/Transportation-and-Mobility/Austin-Crash-Report-Data-Crash-Level-Records/y2wy-tgr5/about_data

2. Chicago Dataset:

No. of Columns: 48

No. of Rows: 896,756

Source: https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data

3. New York City Dataset:

No. of Columns: 29

No. of Rows: 2139048

Source: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data

4. Montgomery Dataset:

No. of Columns: 37

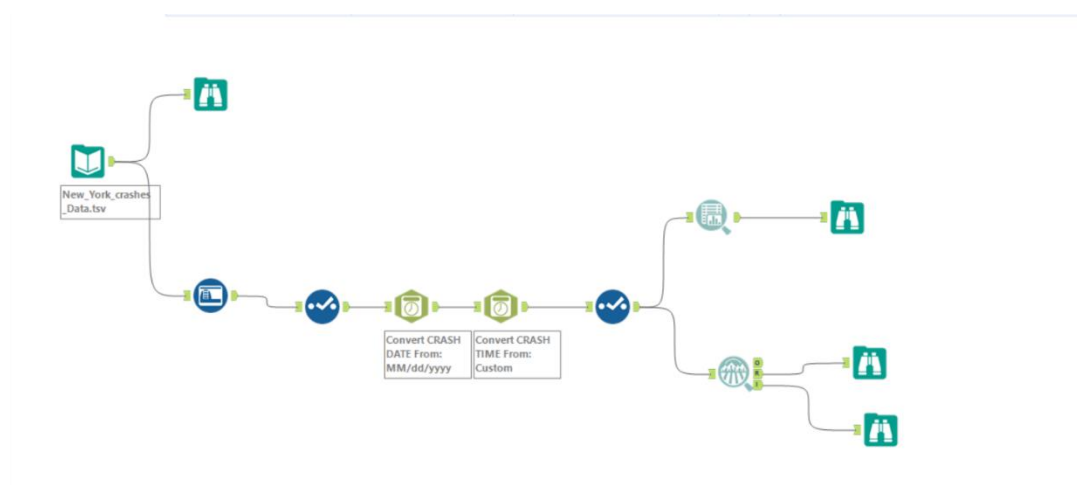
No. of Rows: 107003

Source: https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Incidents-Data/bhju-22kf/about_data

Profiling - Inferences, and Staging

Using Alteryx Designer and Y-Data Profiling, we have performed the profiling on all the given datasets. The following section showcases only the important inferences obtained from the datasets.

NYC Profiling:

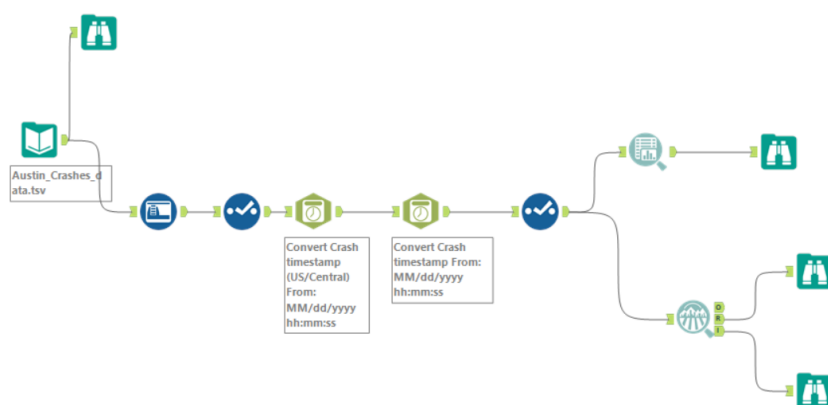


Key Findings

1. **Data Integrity:** A critical data integrity issue was identified and resolved in record 1269803, where a missing close quote was fixed.
2. **Location Data Incompleteness:** Significant null values (>10% on average) exist in location-related columns such as latitude, longitude, zipcode, and street names. Approximately 30,000 records lack all location data.
3. **Time Format Inconsistency:** The Crash_Time column uses a 24-hour format, differing from other datasets.
4. **Primary Key Issues:** The Collision_ID column, intended as a unique identifier, contains null values and inconsistent ID lengths.
5. **Derivable Data:** Potential to derive zipcode information from the Borough column.
6. **Vehicle Classification:** The Vehicle type codes column requires categorization into broader groups.

7. Recommendations
8. Address Completion: Develop a strategy to complete the accident_reported_address using a combination of available location data.
9. Time Standardization: Convert Crash_Time to include AM/PM designations for consistency with other datasets.
10. Unique Identifier Standardization: Implement a uniform Primary Key format to ensure each record is uniquely and consistently identified.
11. Data Enrichment:
12. Derive missing zip codes from Borough information where possible.
13. Categorize vehicle types into broader, more manageable groups.
14. Data Cleaning Protocol: Establish a robust data cleaning protocol to address missing values and ensure data consistency across all fields.

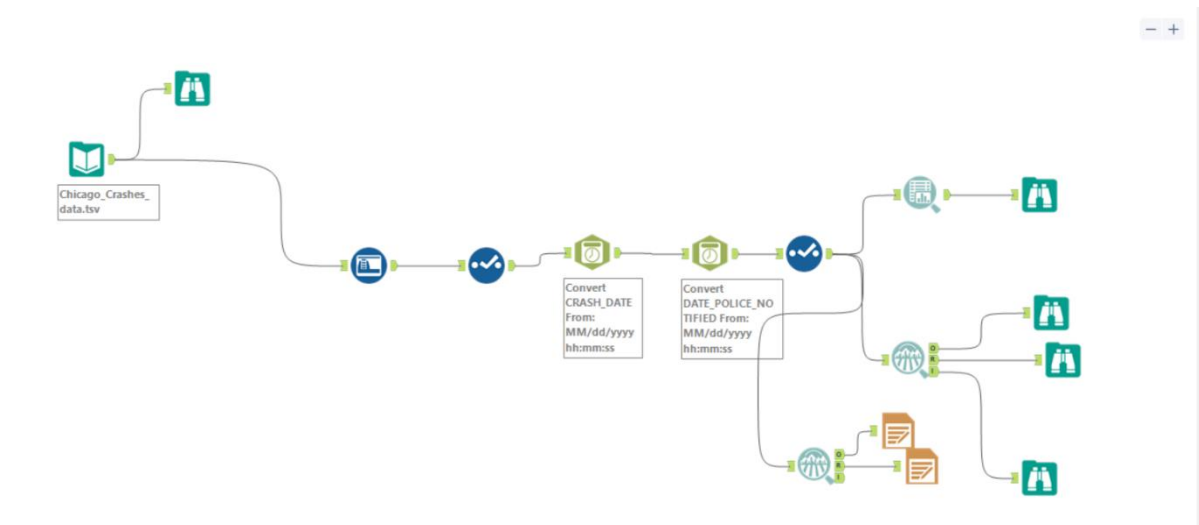
Austin Dataset:



Key Observations:

1. The Austin dataset has three unique columns, namely ID, Crash_ID, and Case_ID. However, after profiling, it is found that the Case_ID column has 1.4% missing values, making it unsuitable for assigning a PK to move forward.
2. Both the Crash Timestamp and Crash Timestamp (US/Central) columns have multi-valued data separated by spaces. These columns can either be divided further for finer granularity or cleaned to standardize the values into a proper format.
3. The Crash Speed Limit column contains the value -1 for accidents that are either under investigation or uncertain. The count for these values is 38,244.
4. Significant null values are found in rpt_street_sfx and rpt_block_num, which can be derived from the rpt_street_name column.
5. Latitude and Longitude have an equal number of null values, which need to be addressed by assigning a non-existent latitude and longitude coordinate instead of using "NA" or 0.
6. There are no records where both the Latitude and Longitude columns are empty, but the Point column is not empty.
7. The reported_street_prefix column is entirely null and can be removed down the line.
8. Fields like micromobility_death_count, bicycle_death_count, and motorcycle_death_count have limited unique values, indicating categorical data.

Chicago Dataset:



Key Observations and Recommendations

1. Columns like DOORING_I (99.7% missing), WORKERS_PRESENT_I (99.9%), WORK_ZONE_I (99.4%), and WORK_ZONE_TYPE (99.6%) have an overwhelmingly high percentage of missing values and could potentially be dropped.
2. Although columns like LANE_CNT and INTERSECTION_RELATED_I have 77.8% and 77.0% null values respectively, they might still be relevant for analysis. These columns can also be dropped if proven to have no significant analytical value.
3. Columns such as STATEMENTS_TAKEN_I, PHOTOS_TAKEN_I, and HIT_AND_RUN_I have Boolean-like values (Y/N) but include missing entries. These should be converted into a binary format (1/0) for easier processing, with missing values explicitly handled, for example, assigning a 0 for missing entries.
4. For geographical analysis, apart from latitude and longitude, as well as street_no and street_name, other columns such as street_direction and location can be dropped as they do not add much value.
5. The CRASH_DATE and DATE_POLICE_NOTIFIED columns can be split into separate DATE and TIME components to enable temporal analysis.
6. There are no records where latitude and longitude are null while the location column is not.
7. The INJURIES_TOTAL column could be recalculated by summing the values of other columns denoting injuries.

It was also observed that there are no records where the INJURIES_TOTAL is greater than the aggregated value of related columns, indicating that no category has been left out.

```
19 select * from stg_chicago_crashes
20 where INJURIES_TOTAL > (INJURIES_FATAL + INJURIES_INCAPACITATING + INJURIES_NON_INCAPACITATING + INJURIES_REPORTED_NOT_EVIDENT + INJURIES_NO_INDICATION +
21 INJURIES_UNKNOWN)
22
```

Results | Chart

CRASH_RECORD_ID	CRASH_DATE	CRASH_DATE_EST_I	POSTED_SPEED_LIMIT	TRAFFIC_CONTROL_DEVICE	DEVICE_CONDITION	WEATHER_CONDITION
Query produced no results						

Query Details

Query duration	498ms
Rows	0

Additionally, around 876k records were found where the INJURIES_TOTAL does not match the sum of other injury columns. Therefore, the INJURIES_TOTAL column could be recalculated for consistency.

```
19 select * from stg_chicago_crashes
20 where INJURIES_TOTAL != (INJURIES_FATAL + INJURIES_INCAPACITATING + INJURIES_NON_INCAPACITATING + INJURIES_REPORTED_NOT_EVIDENT +
21 INJURIES_NO_INDICATION + INJURIES_UNKNOWN)
22
```

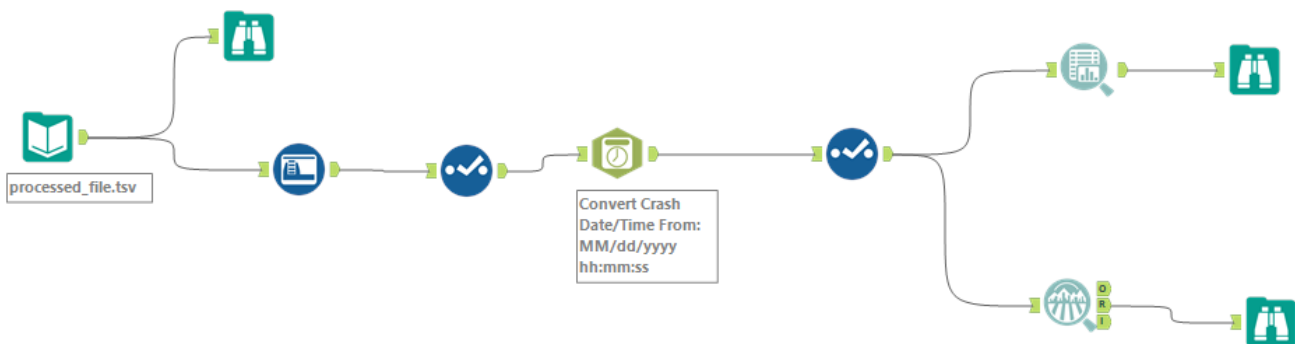
	CRASH_RECORD_ID	CRASH_DATE	CRASH_DATE_EST_I	POSTED
1	5dd7fd7f8283655c01858337f00d486bcccddc19bd06df49db543ff615d9446ba	03/24/2023 03:00:00 PM	null	30
2	5dd80e7f610f0252c4d0ddc4b1dc10d9253cc691d9c426f9cb965d3052f35c7et	01/20/2024 09:47:00 PM	null	40
3	5dd811e7bcafd011707c681ccf005ed30af02b76aaa041dd1b1167552fccc9b6	04/22/2021 04:10:00 PM	null	20

Query Details

Query duration 51ms

Rows 876.7K

4. Montgomery Profile:



Key Observations:

1. There are few records that have \n in between the values, causing a record to breakdown into multiple records. This must be addressed.

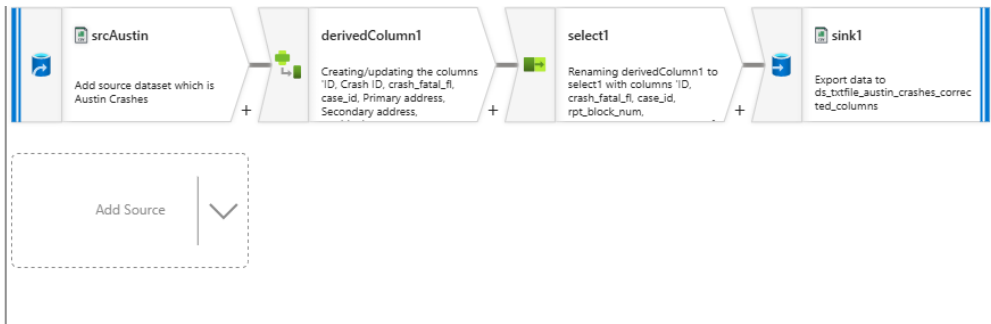
Staging-As-Is Tables Using ADF:

ADF was used to create pipelines for the initial staging-as-is tables. The "Medallion Architecture" was implemented to separate operations. The "bronze" layer was utilized to stage all source files, and the "silver" layer was used to stage all Parquet files in the first step of the pipeline. Source files were converted to Parquet format and then loaded into Snowflake tables. Screenshots of the job and the counts in Snowflake are provided.

For Austin:



Since the Austin columns are not as per ADF, we had to create a dataflow to address this issue and make the column names compliant.



Counts of the as-is Staging Tables and the dataset:

Views: 7,592 Downloads: 3,753

Data Provided by: City of Austin, Texas Dataset Owner: transportation.data@austintexas.gov

Attachments: dataset_cover_photo

Topics: Category: Transportation and Mobility Tags: crash, vision zero

What's in this Dataset?

Rows	Columns	Each Row
213K	43	Crash

Columns (43)

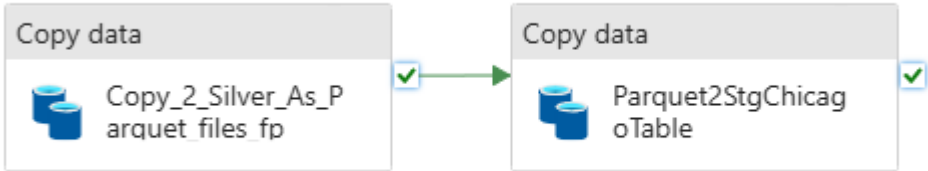
Column Name	Description	Field Name	Data Type
ID	The ID	id	Number
Crash ID	TxDOT C.R.I.S. system-generated unique identifying number for a crash	cris_crash_id	Number
crash_fatal_fl	Fatal Crash Identifier - Indicates that the crash involved one or more fatalities	crash_fatal_fl	Checkbox
case_id	Case ID	case_id	Text
Primary address	The primary address where the crash is reported to have occurred	address_primary	Text
Secondary address	The secondary address where the crash is reported to have occurred	address_secondary	Text

```

1 | SELECT COUNT(*) FROM STG_AUSTIN_CRASHES;
  
```

	COUNT(*)
1	213053

For Chicago:



Chicago's dataset contains valid column names hence we didn't use dataflow to correct the column names.

CHICAGO DATA PORTAL Chicago Data Portal

Browse Tutorial Feedback

About Data Related Content

Traffic Crashes - Crashes

Search

CRAS...	CRASH_DATE_EST_J	CRASH_DATE	POSTED_SPEED_LIMIT	TRAFFIC_CONTROL_DEVICE	DEVICE_CONDITION	WEATHER_CONDITION	LIGHTING_CONDITION	FIRST_CRASH_T
3cafe9ed5	Y	12/01/2024 09:55:00 PM	30	STOP SIGN/FLASHER	FUNCTIONING PROPERLY	CLEAR	DARKNESS, LIGHTED ROAD	
974cb6a8a		12/01/2024 09:45:00 PM	30		FUNCTIONING PROPERLY	CLEAR	DARKNESS, LIGHTED ROAD	
af6f0fe4c6		12/01/2024 09:30:00 PM	30		FUNCTIONING PROPERLY	CLEAR	DUSK	
3615af1d8		12/01/2024 09:28:00 PM	30		NO CONTROLS	CLEAR	DARKNESS, LIGHTED ROAD	SIDESWIPE
c8a40e18c		12/01/2024 09:10:00 PM	30		NO CONTROLS	CLEAR	DARKNESS	PARKE
88faa58c6c		12/01/2024 09:02:00 PM	30		NO CONTROLS	CLEAR	DARKNESS, LIGHTED ROAD	PARKE
06d359f9d		12/01/2024 08:56:00 PM	30		UNKNOWN	CLEAR	DARKNESS, LIGHTED ROAD	PARKE
e3251a0c1	Y	12/01/2024 08:25:00 PM	30		NO CONTROLS	CLEAR	DARKNESS	
ed10205fe1		12/01/2024 08:25:00 PM	30	NO CONTROLS	NO CONTROLS	CLEAR	DARKNESS, LIGHTED ROAD	
6d8c39ca7		12/01/2024 08:12:00 PM	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	CLEAR	DARKNESS, LIGHTED ROAD	
7e074cdact		12/01/2024 07:43:00 PM	30	STOP SIGN/FLASHER	FUNCTIONING PROPERLY	CLEAR	DARKNESS, LIGHTED ROAD	PARKE
8add6e17c		12/01/2024 07:41:00 PM	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	CLEAR	DARKNESS, LIGHTED ROAD	
3a5d365ed		12/01/2024 07:35:00 PM	30	STOP SIGN/FLASHER	FUNCTIONING PROPERLY	CLEAR	UNKNOWN	SIDESWIPE

Export dataset

Traffic Crashes - Crashes

Download file API endpoint

Export format
CSV

All data (897880 rows)

Cancel Download

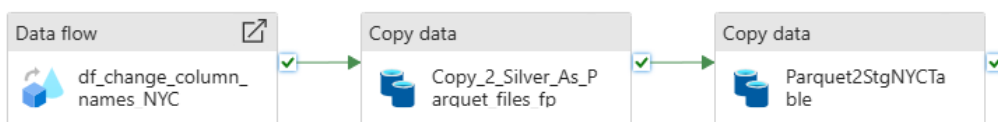
5 | SELECT COUNT(*) FROM STG_CHICAGO_CRASHES

Results Chart

	COUNT(*)
1	897880

3. For NYC:

Like Austin we used Dataflows to modify the column names.



NYC OpenData

Home Data About Learn Contact Us

About Data Related Content

Motor Vehicle Collisions - Crashes Public Safety

The Motor Vehicle Collisions crash data contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions crash data is required to be filled out for collisions involving a motor vehicle.

Read more

About this Dataset

Updated December 3, 2024

Data Last Updated December 3, 2024 Metadata Last Updated April 19, 2021

Date Created April 26, 2014

Views 601K Downloads 204K

Data Provided by Police Department (NYPD) Dataset Owner NYC OpenData

Agency Police Department (NYPD)

Update

Update Frequency Daily

Automation Yes

Date Made Public 5/7/2014

Attachments

MVCollisionsDataDictionary_20190813_ERD.xlsx

Export dataset

Motor Vehicle Collisions - Crashes

Download file API endpoint

Export format
TSV for Excel

All data (2139048 rows)

Cancel Download

SELECT COUNT(*) FROM STG_NYC_CRASHES;	
results	Chart
	COUNT(*)
	2139048

4. For Montgomery:

Similarly for Montgomery.



dataMontgomery

Crash Reporting - Incidents Data

This dataset provides general information about collisions occurring on county and city roads, collected via the Automated Crash Reporting System (ACRS) and reported by the Montgomery County Police, and reported by the Montgomery County Police.

Updated **November 29, 2024**

Export dataset

Crash Reporting - Incidents Data

Download file API endpoint

Export format
CSV

All data (107003 rows)

Cancel Download

5	SELECT COUNT(*) FROM STG_MONTGOMERY_CRASHES	
Results	Chart	
		COUNT(*)
1		107003

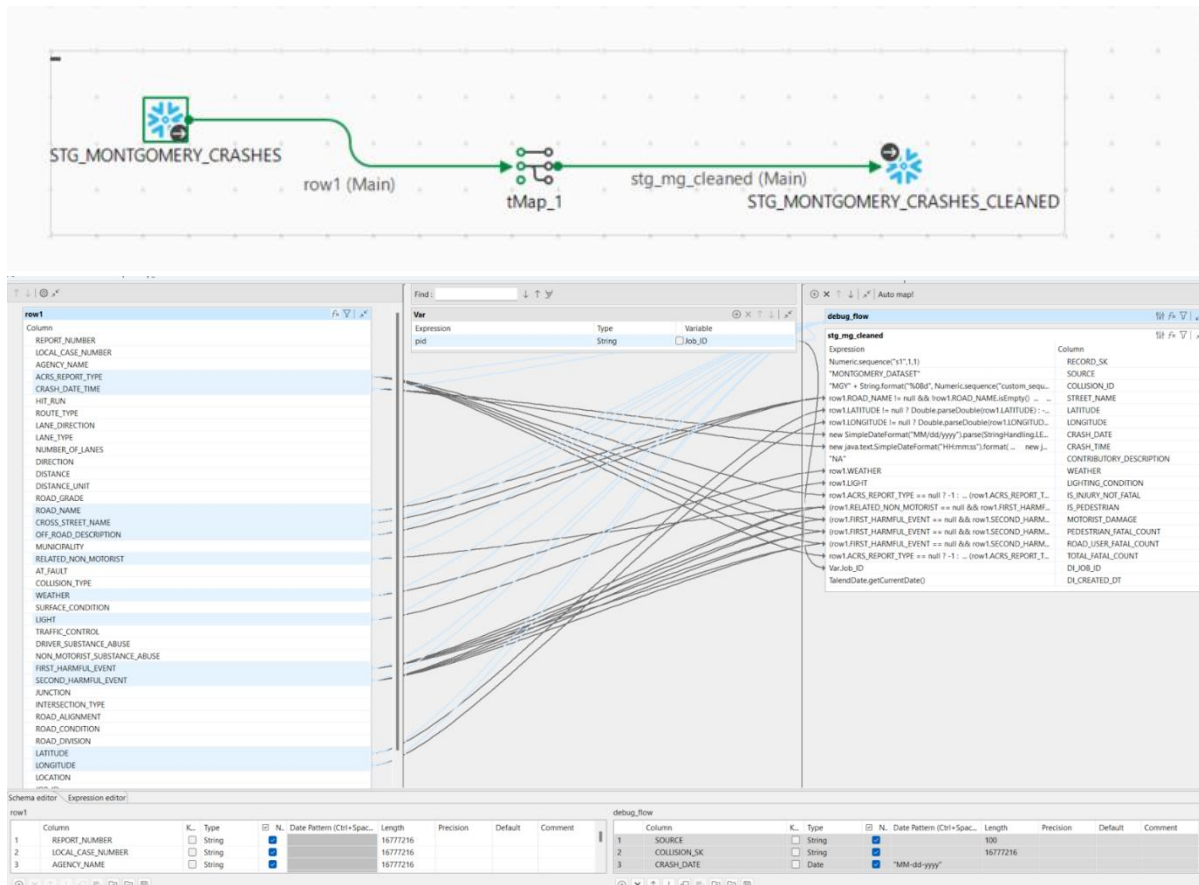
Mapping Document:

A Mapping Document was created, where all the columns in all the datasets were examined at an overall level. Based on the business requirements, the necessary data present in columns of each of the datasets were observed and the final mapping document was created, that is in a separate file.

ETL Pipelines Using Talend for Cleaning and Transforming the Intermediate Tables

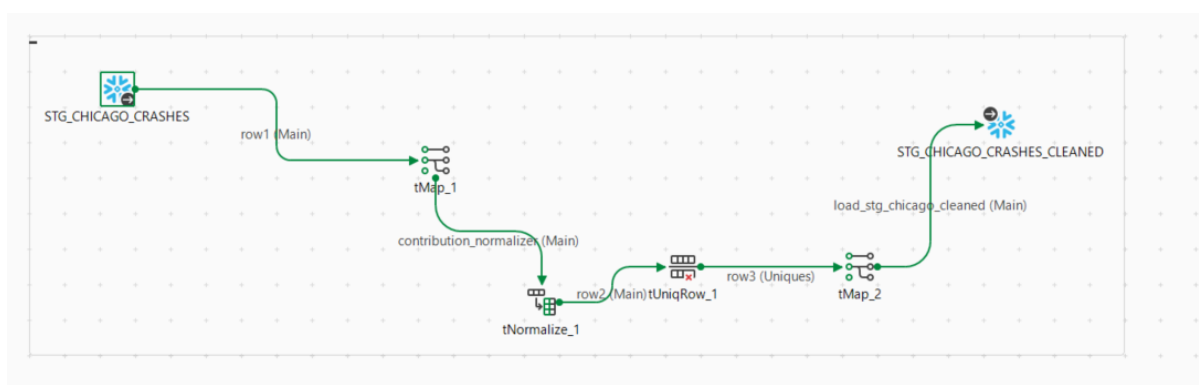
After staging the tables as-is in the dataset, Talend was used to load the cleaned data, incorporating both cloud and local pipelines. The columns in each dataset were finalized for cleaning and transformation based on the specified business requirements to be achieved.

For Montgomery:



Transformation logic was added in the TMap component of Talend to address null values, clean improper data, and implement logic to meet business requirements by finalizing the columns from the actual dataset.

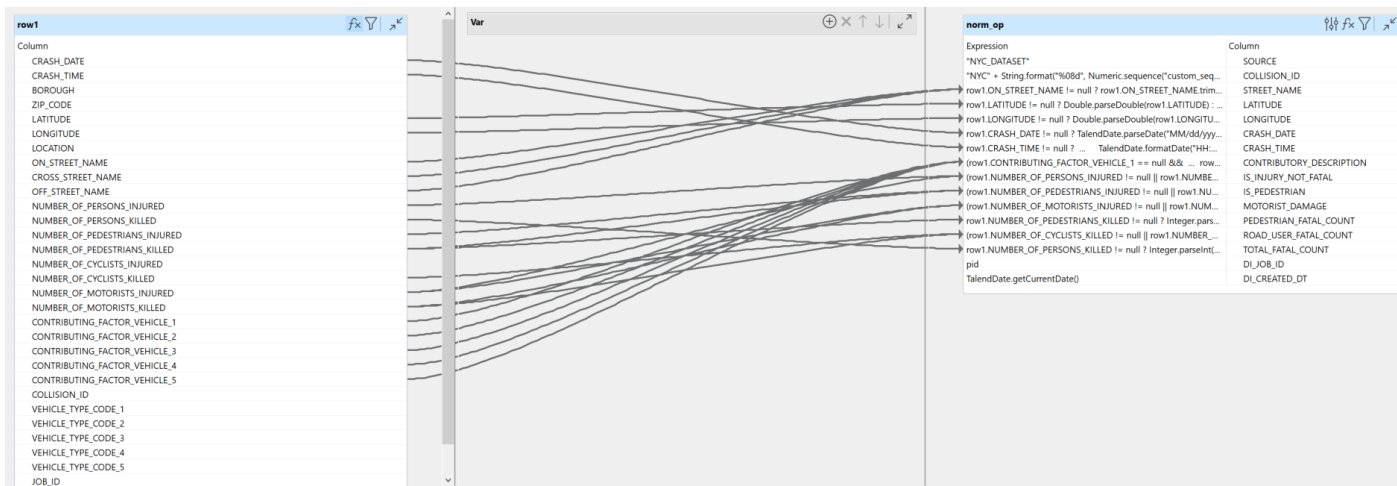
For Chicago:



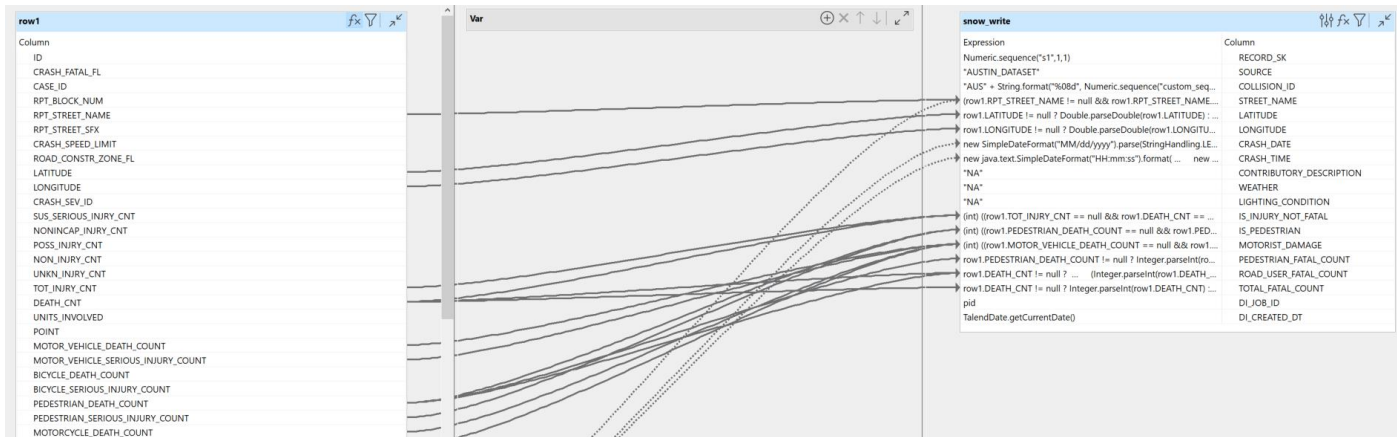
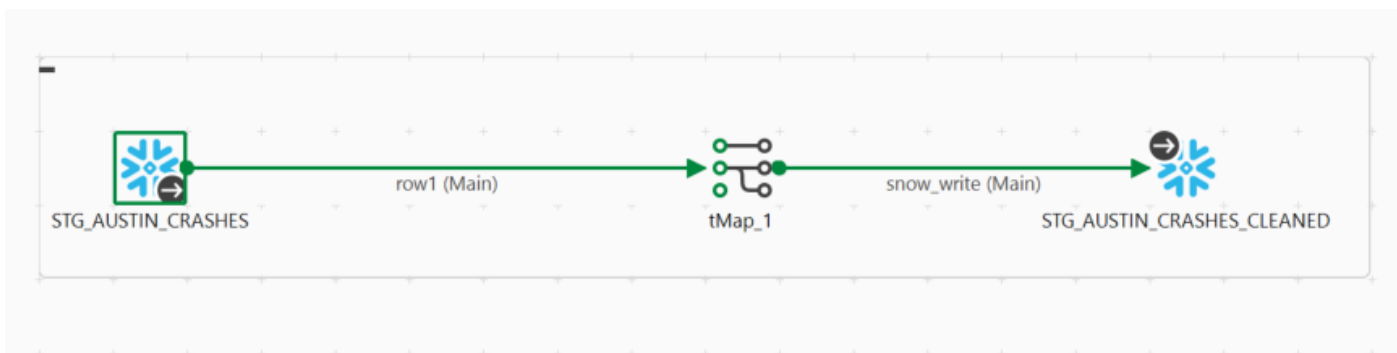
[illegible]

```

graph LR
    A[STG_NYC_CRASHES] -- row1 (Main) --> B[tMap_1]
    B -- norm_op (Main) --> C[tNormalize_1]
    C -- row2 (Main) --> D[tUniqRow_1]
    D -- row3 (Uniques) --> E[tMap_2]
    E -- snow_write (Main) --> F[STG_NYC_CRASHES_CLEANED]
  
```

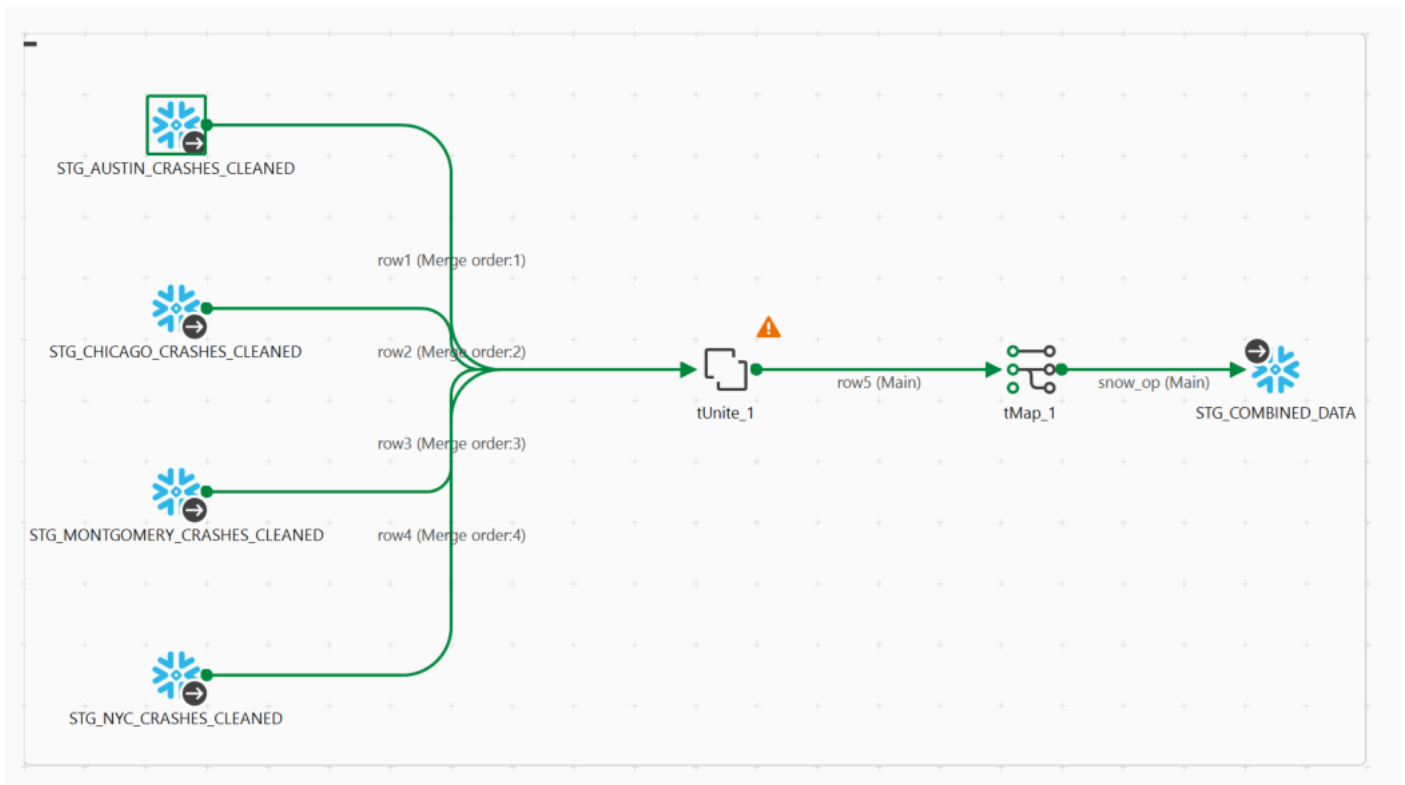


For Austin:



For the final Merged Staging table:

Since all the column names, datatypes are same across all datasets, they are be directly union-ed.



Number of rows:

20 | `SELECT COUNT(*) FROM STG_COMBINED_DATA`

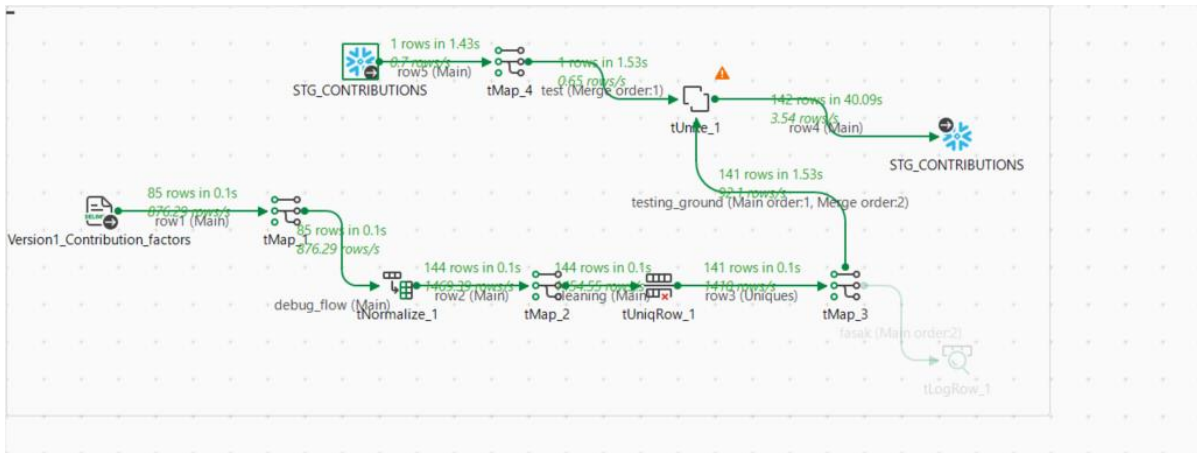
Results
Chart

	COUNT(*)
1	5077174

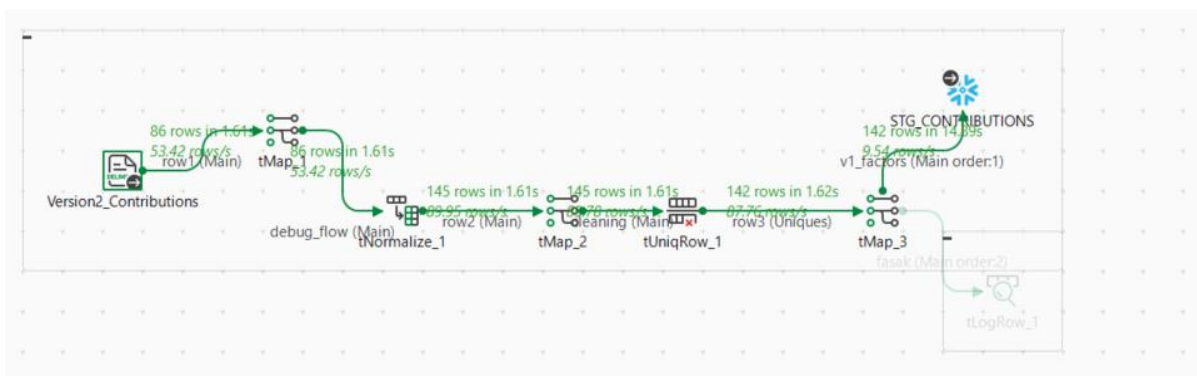
SCD Type-2 Implementation

As per the Change Request in Phase-2 of the project, logic was implemented to capture data changes and maintain a history. Talend was used to create pipelines for this purpose. Two files, version-1 and version-2, were provided to track changes in the code and description of four datasets.

A pipeline was created to load version-1 changes into the STG_CONTRIBUTIONS table and assign proper codes based on the description.



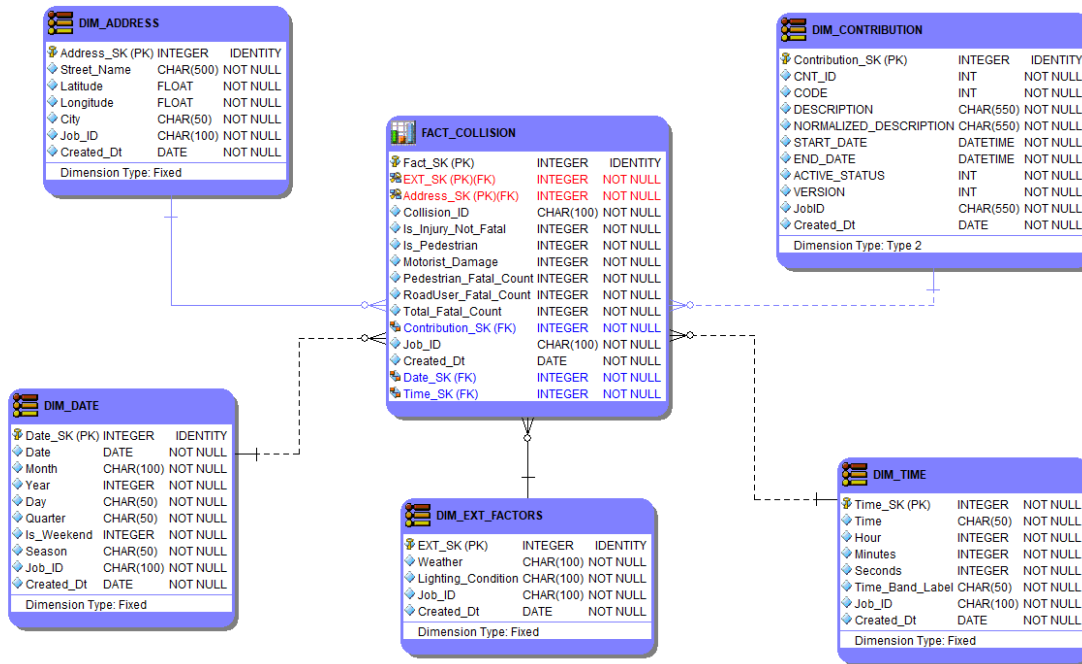
The tJDBCEL component in Talend was used to implement Slowly Changing Dimensions (SCD) for the datasets. Another pipeline was created to upsert any changes based on the version-2 file, targeting the STG_CONTRIBUTIONS table.



This is another pipeline where it will upsert any changes as per the version-2 file. The target is STG_CONTRIBUTIONS table.

DIMENSIONAL MODELING & LOADING THE FACT AND DIMENSION TABLES

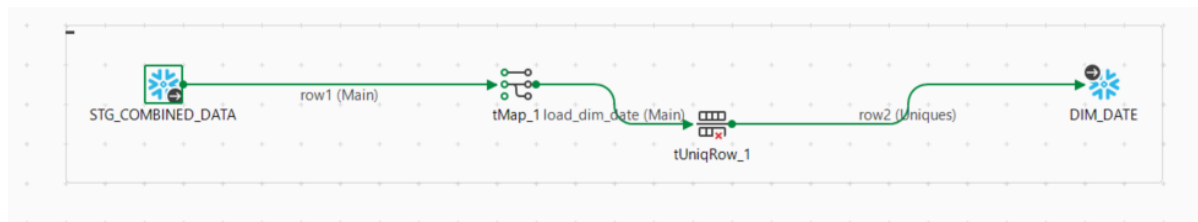
Using ER Studio, the dimensional model was designed, and DDL scripts for dimension tables and the fact table were set up.



The relationships between entities have been assigned appropriately.

With the dimensional model design finalized, the DDL scripts provided by ER Studio were used to create the respective tables in Snowflake. Pipelines were then created to load these tables and apply any necessary transformations.

Pipeline to load DIM_DATE:

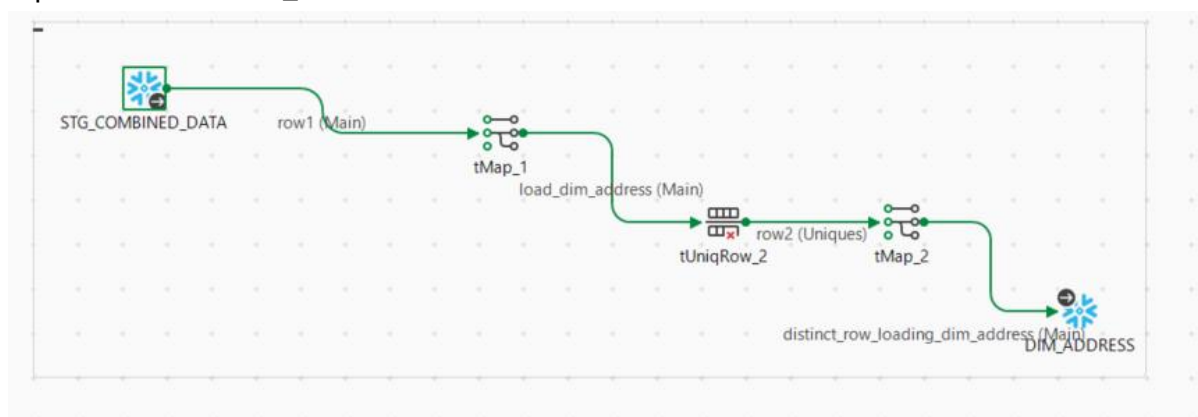


The stg_combined_data will be acting as a source to load all the dimension and fact tables.

```
SELECT COUNT(*) FROM DIM_DATE;
```

COUNT(*)
5450

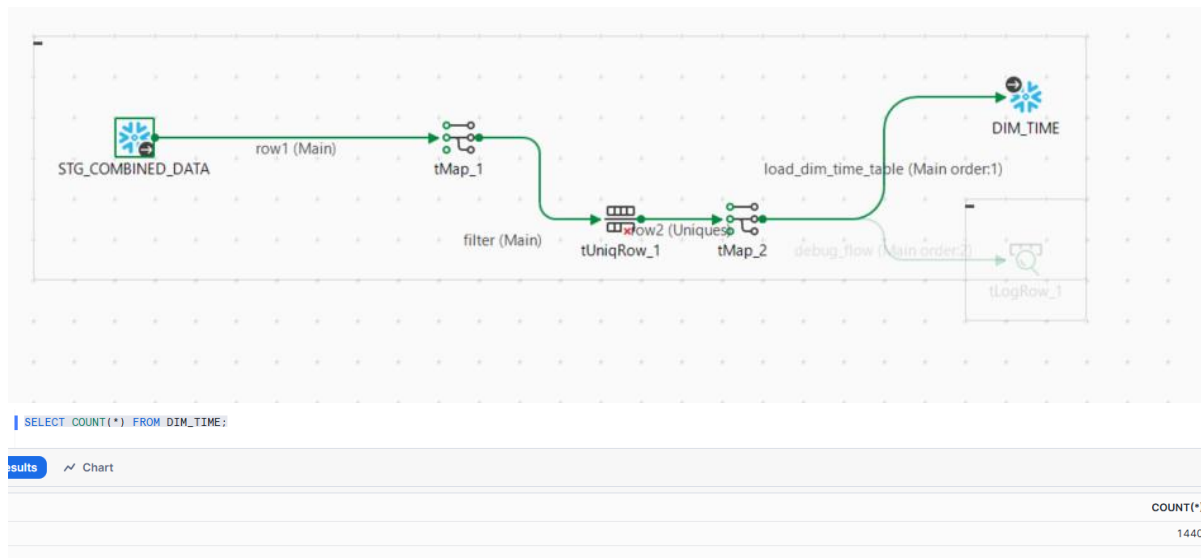
Pipeline to load DIM_ADDRESS:



SELECT COUNT(*) FROM DIM_ADDRESS;	
results	Chart
	COUNT(*)
	994899

Pipeline to load DIM_TIME:

This DIM has only 1440 records – satisfying the logic of 24 hours * 60 minutes in a day = 1440 unique time.



Pipeline to load DIM_CONTRIBUTIONS:

UDBCSCDELT_1

Designer Code

Job(scd_job 0.1) Context (scd_job) Component X Run (Job scd_job) Cloud Artifact

UDBCSCDELT_1

Basic settings

Advanced settings

Dynamic settings

View

Documentation

Source table: "STG_CONTRIBUTIONS"
Table: "DIM_CONTRIBUTION"
Action on table: None
Schema: Built-In Edit schema
Surrogate key: SCD_SK Creation Auto increment
Some databases support "DB sequence" only, some other databases support "Auto increment" only, please select the right type

Source keys

Name
CNT_ID
CODE

+ × ↑ ↓ ↺ ↻

☒ Use SCD type 0 fields

SCD type 0 fields

Field name
JobID
Created_Dt

+ × ↑ ↓ ↺ ↻

☐ Use SCD type 1 fields
☒ Use SCD type 2 fields

SCD type 2 fields

Field name
DESCRIPTION
NORMALIZED_DESCRIPTION

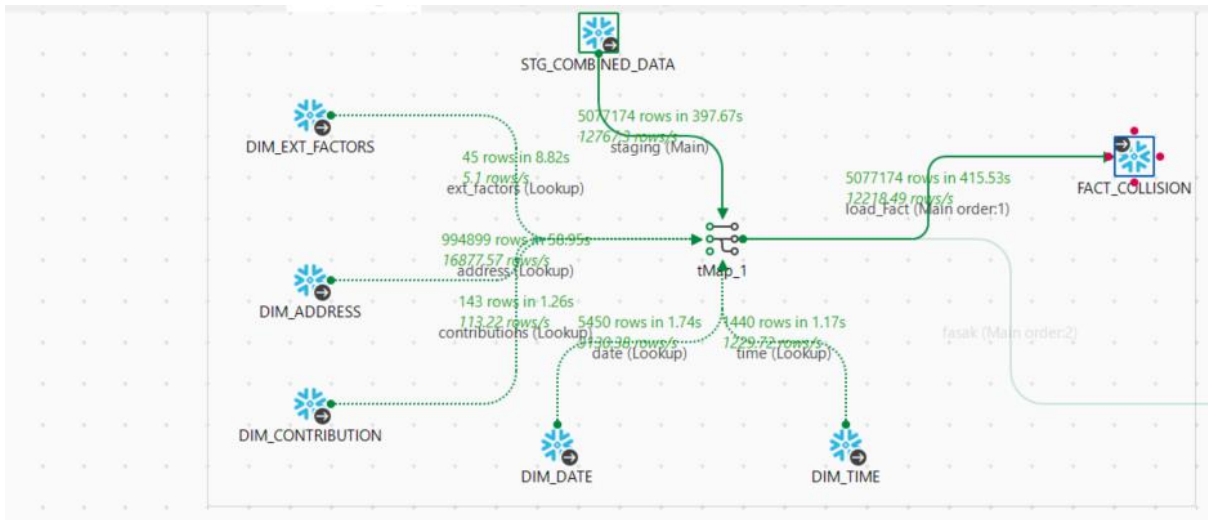
Pipeline to load DIM_EXTERNAL_FACTORS:

SELECT COUNT(*) FROM DIM_EXT_FACTORS;

Results
Chart

	COUNT(*)
	45

Pipeline to load Fact_Table:



```
SELECT COUNT(*) FROM fact_collision;
```

Count	COUNT(*)
5077174	5077174

The counts match the number of records in stg_combined_table.

Visualizations:

Design Rationale:

The design strategy was to position the title and filters at the top, with the most important KPIs placed directly below them. This layout optimizes space for visualizing charts while adhering to best practices. The most important KPI, Total Accidents, was positioned in the top-left corner, aligning with the natural eye movement of the viewer.

Both Power BI and Tableau dashboards were designed to maintain a similar structure for consistency.

Power BI:

Motor Collision Analysis

CITY
YEAR
WEEKEND
DAY
TIME BAND

All
All
All
All
All

Total Accidents

3.36M

Injury - Only Acc.

761.36K

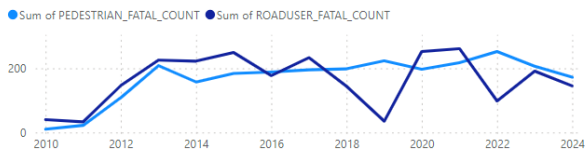
Motorist Casualties

648.44K

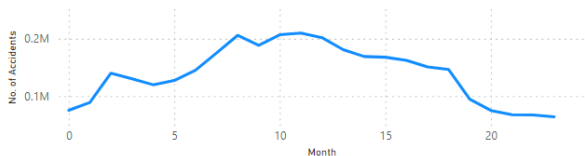
Total Fatal Count

7.73K

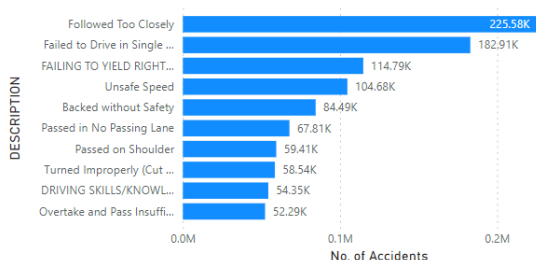
Fatality Analysis - Pedestrian v. Roaduser



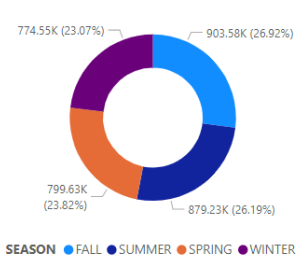
Accidents by Hour



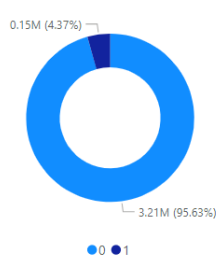
Most Common Factors in Acc.



Accidents By Season



Pedestrian Inv. in Acc.



Motor Collision Analysis

CITY:
 YEAR:
 WEEKEND:
 DAY:
 TIME BAND:

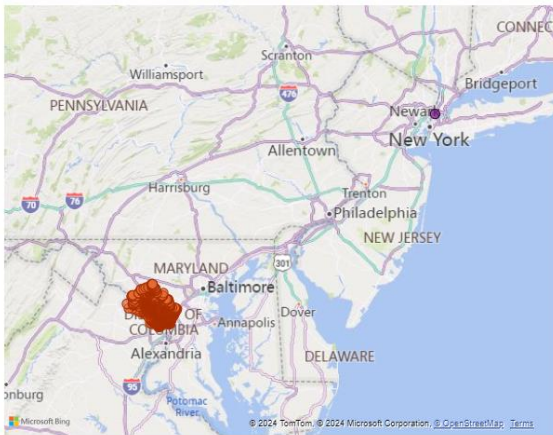
Total Accidents
3.36M

Injury - Only Acc.
761.36K

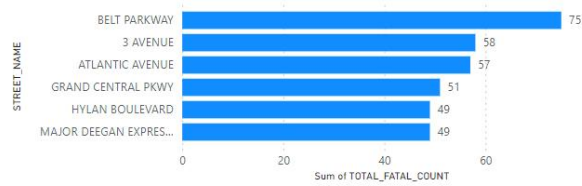
Motorist Casualties
648.44K

Total Fatal Count
7.73K

Geographical Visualization of Accidents



Top 5 Areas in City with Most Fatal Accidents



Top 3 Areas in City with Most Accidents

BROADWAY	25233	Count of COLLISION_ID
WESTERN AVE	24506	Count of COLLISION_ID
PULASKI RD	21700	Count of COLLISION_ID

Tableau:

Motor Collision Analysis

City:
 Year:
 Weekend:
 DAY:
 Time Band Label:

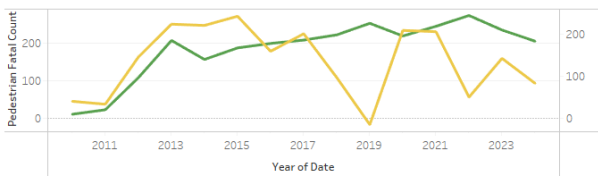
Total Accidents
3.36M

Injury - Only Acc.
761.36K

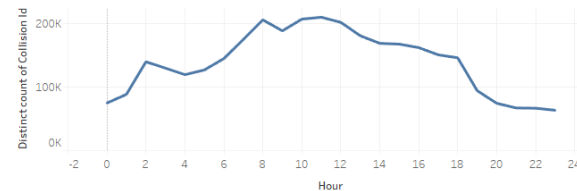
Motor Casualty
648.44K

Total Fatal Count
7.73K

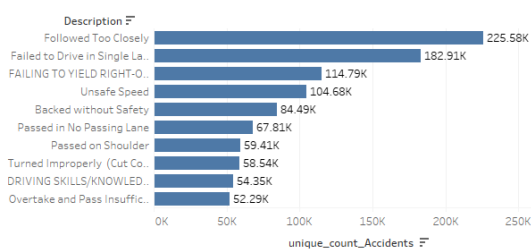
Fatality Analysis- Pedestrian vs RoadUser



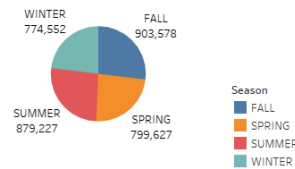
Accidents by hour



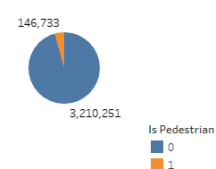
Most_Common_factors



Accidents by season



Pedestrian Inv. in Acc.



Motor Collision Analysis

City

(All)

Year

(All)

Weekend

(All)

DAY

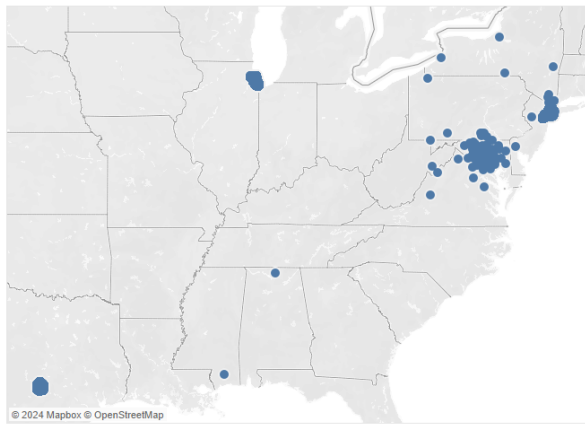
(All)

Time Band Label

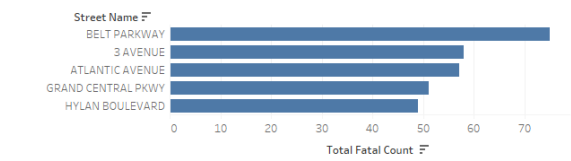
(All)

Total Accidents	Injury - Only Acc.	Motor Casualty	Total Fatal Count
3.36M	761.36K	648.44K	7.73K

Geographical Visualization of Accidents



Top 5 areas in City with most Fatal Accident



Top 3 Areas in City with Most Accidents

