DADABI Capstone Project: Motor Collision Analysis – An end-to-end project

Agash Uthayasuriyan
Ram Prashanth Rao G

Contents:

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

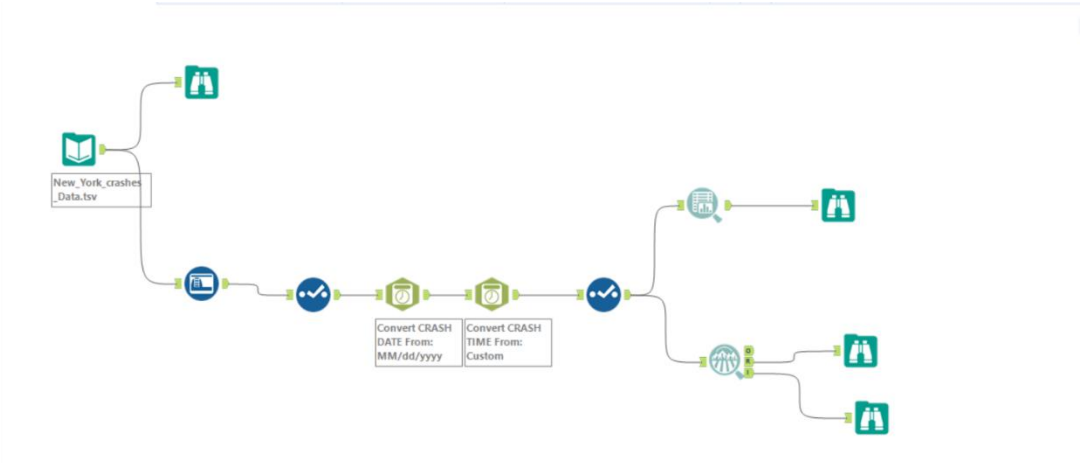
3. New York City Dataset:
   No. of Columns: 29
   No. of Rows: 2139048
   Source: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data

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:

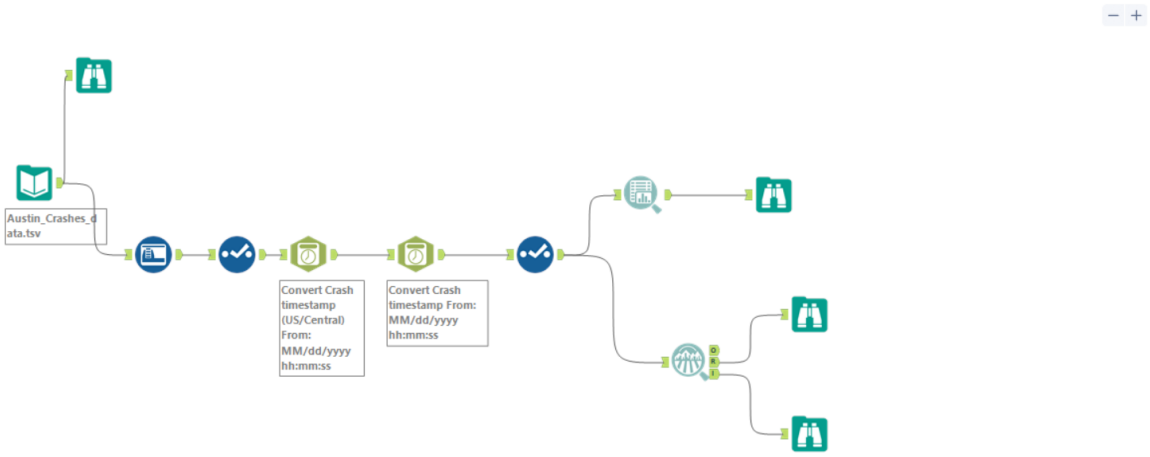


Key Findings

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.

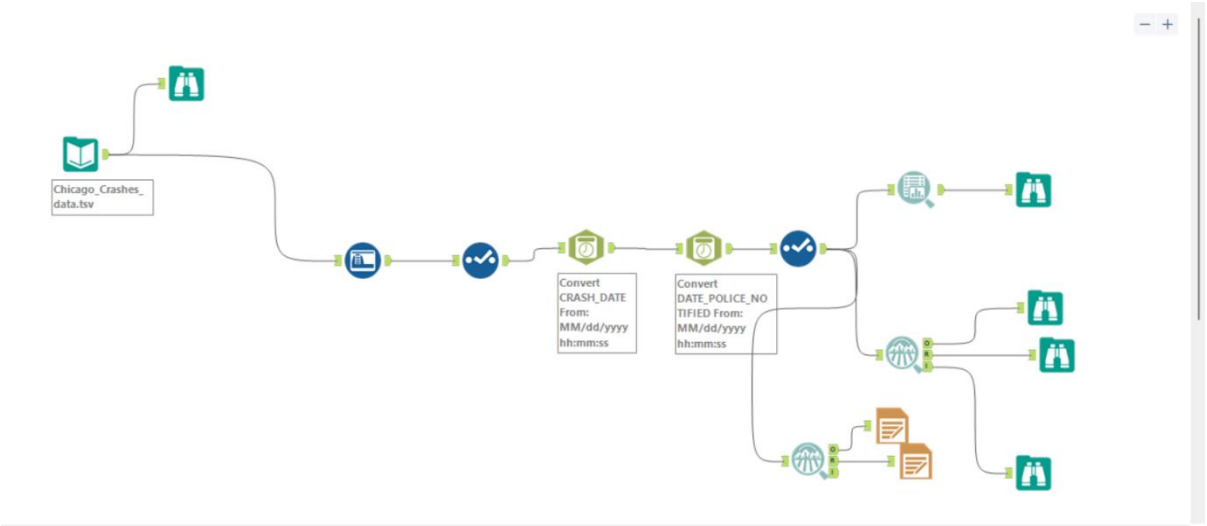
Austin Dataset:



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.

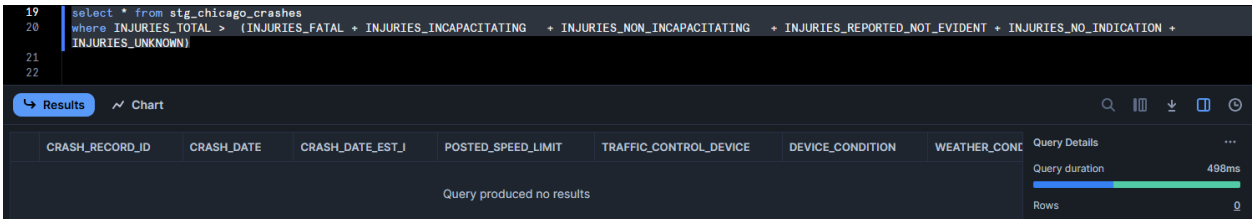
Chicago Dataset:



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.



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.

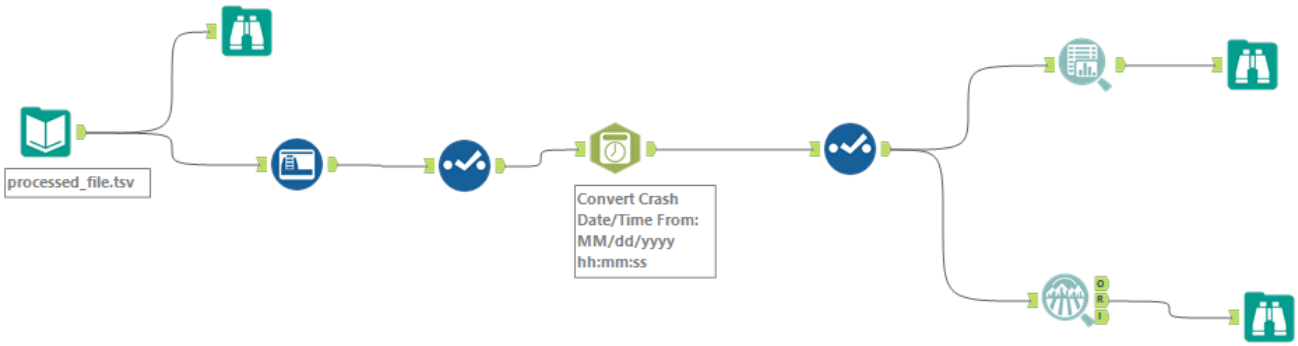
```
select * from stg_chicago_crashes
where INJURIES_TOTAL != (INJURIES_FATAL + INJURIES_INCAPACITATING   + INJURIES_NON_INCAPACITATING   + INJURIES_REPORTED_NOT_EVIDENT +
INJURIES_NO_INDICATION +    INJURIES_UNKNOWN)
```

Results | Chart

	CRASH_RECORD_ID	CRASH_DATE	CRASH_DATE_EST_I	POSTED
1	5dd7fd7f8283655c01858337f00d486bccecdc19bd06df49db543ff615d9446ba	03/24/2023 03:00:00 PM	null	30
2	5dd80e7f610f0252c4d0ddc4b1dc10d9253cc691d9c426f9cb965d3052f35c7e	01/20/2024 09:47:00 PM	null	40

Query Details
Query duration 51ms
Rows 876.7K

4. Montgomery Profile:



Key Observations:

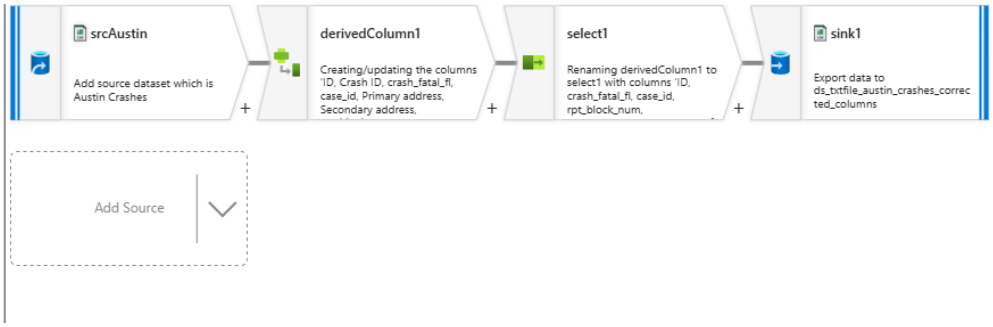
1. There are few records that have \n in between the values, causing a record to breakdown into multiple records. This must be addressed.

Staging-As-Is Tables Using ADF:

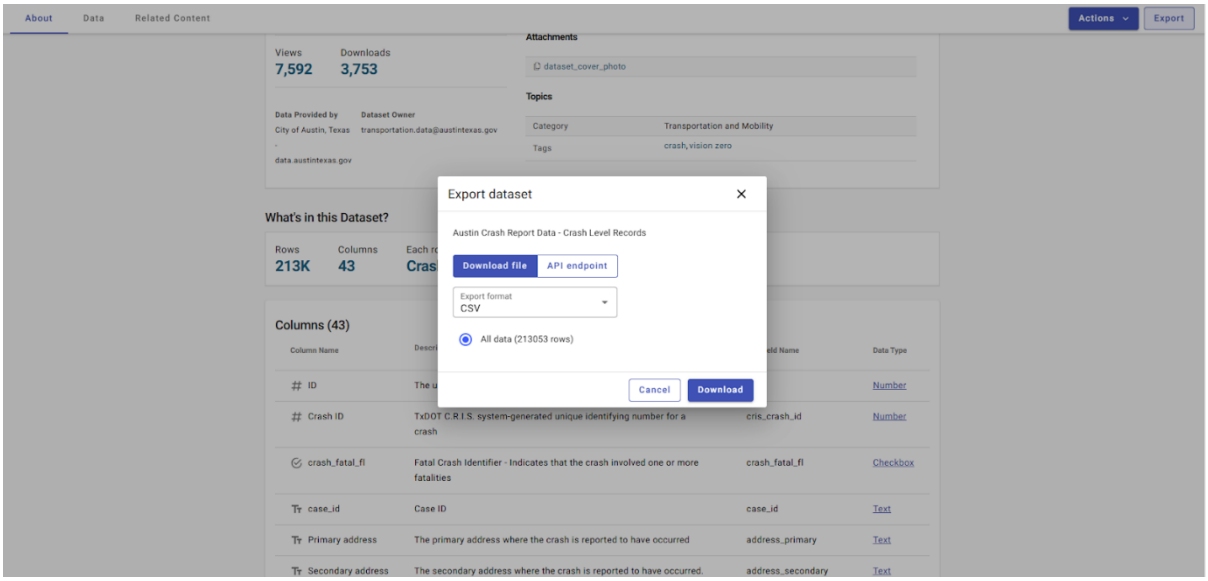
ADF was used to create pipelines for the initial staging-as-is tables. The "Medallion Architecture" was implemented to separate operations. The "bronze" layer was utilized to stage all source files, and the "silver" layer was used to stage all Parquet files in the first step of the pipeline. Source files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

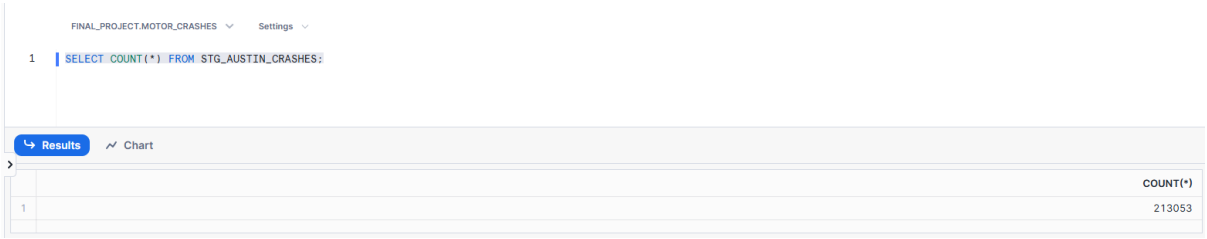
For Austin:

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.

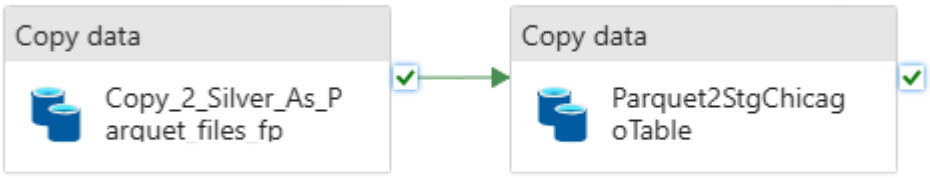


Counts of the as-is Staging Tables and the dataset:

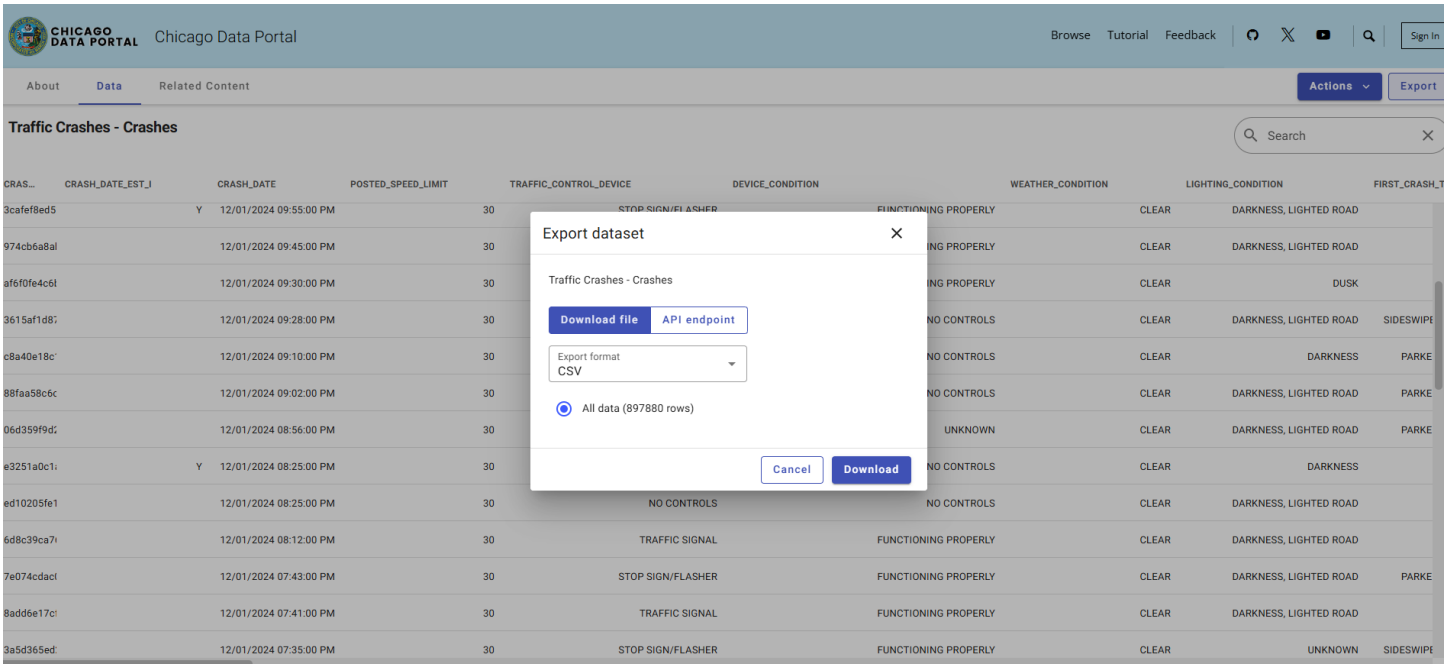


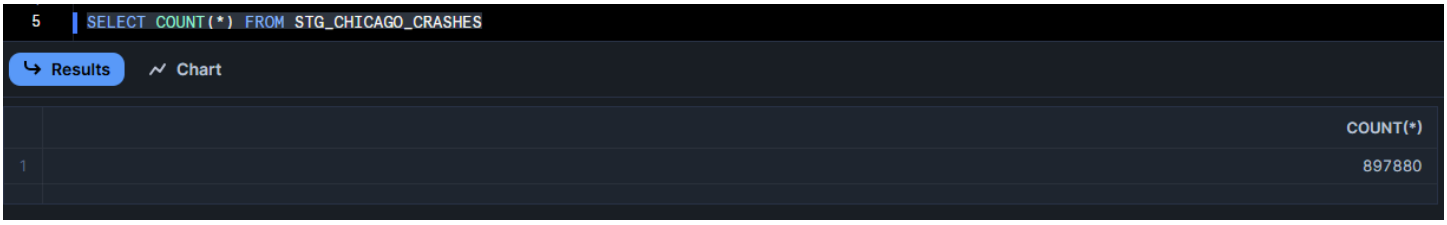


For Chicago:



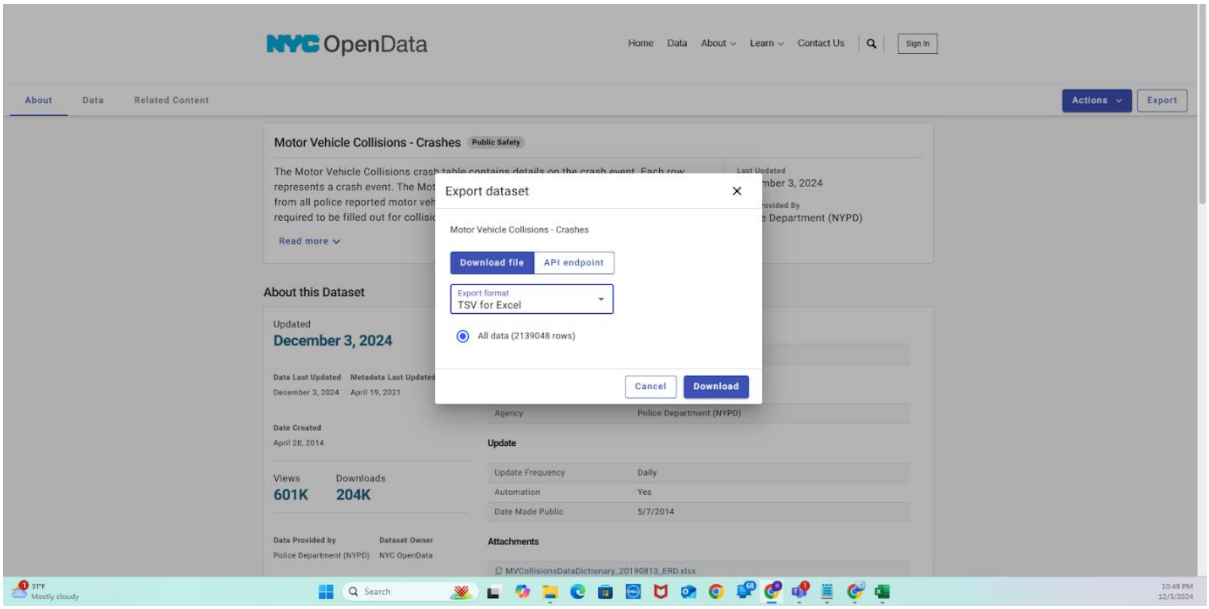
Chicago's dataset contains valid column names hence we didn't use dataflow to correct the column names.





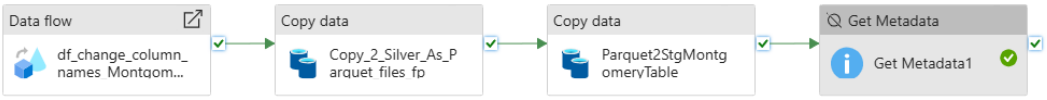
3. For NYC:

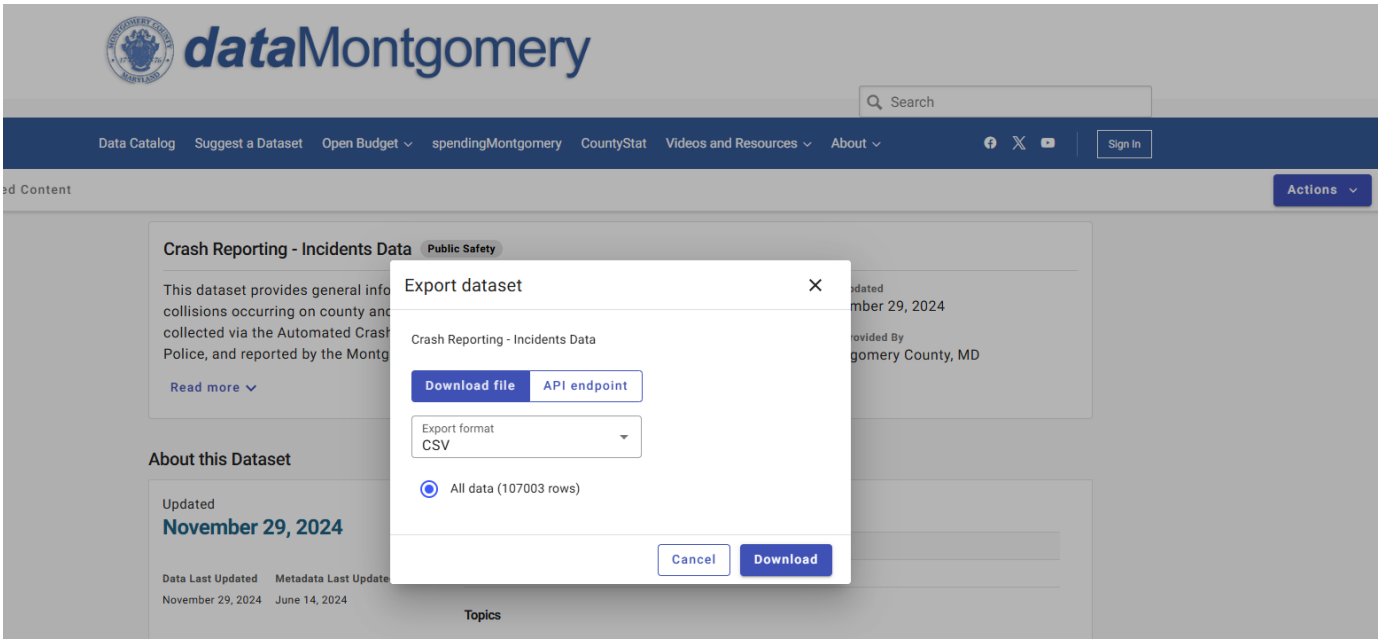
Like Austin we used Dataflows to modify the column names.

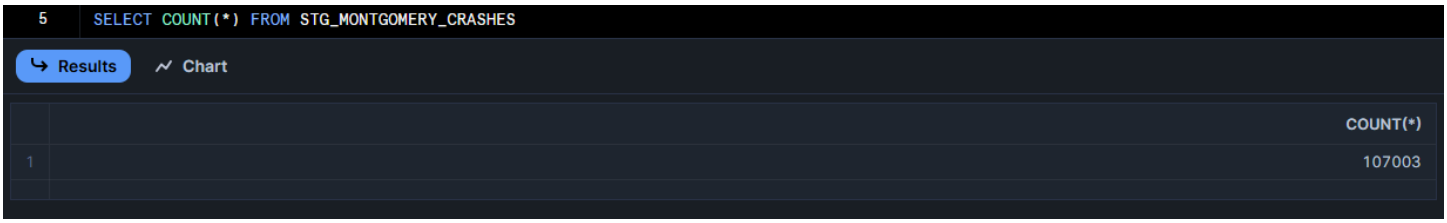


4. For Montgomery:

Similarly for Montgomery.







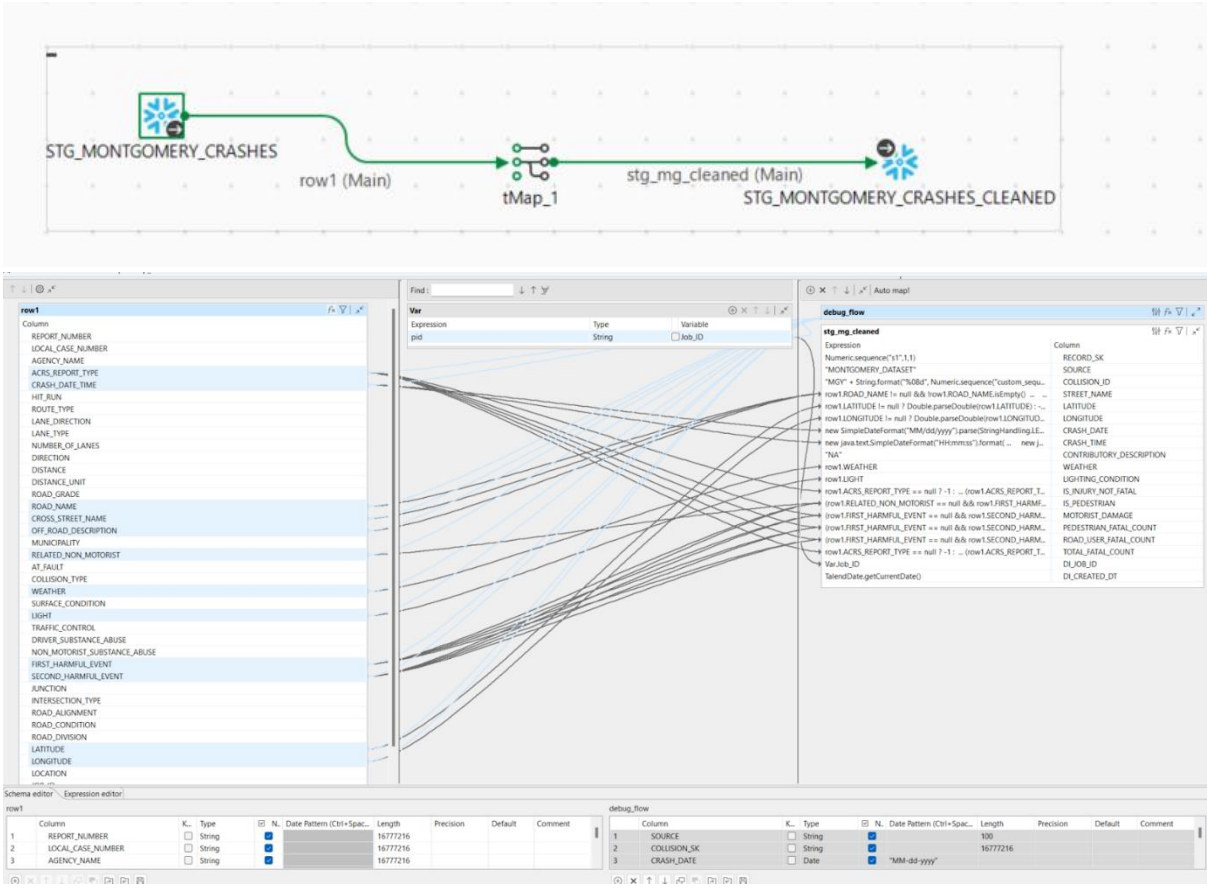
Mapping Document:

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

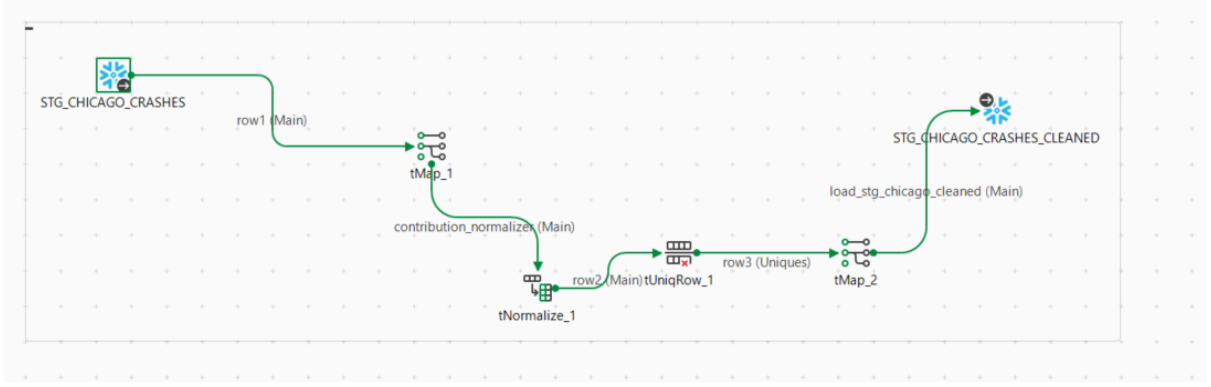
After staging the tables as-is in the dataset, Talend was used to load the cleaned data, incorporating both cloud and local pipelines. The columns in each dataset were finalized for cleaning and transformation based on the specified business requirements to be achieved.

For Montgomery:

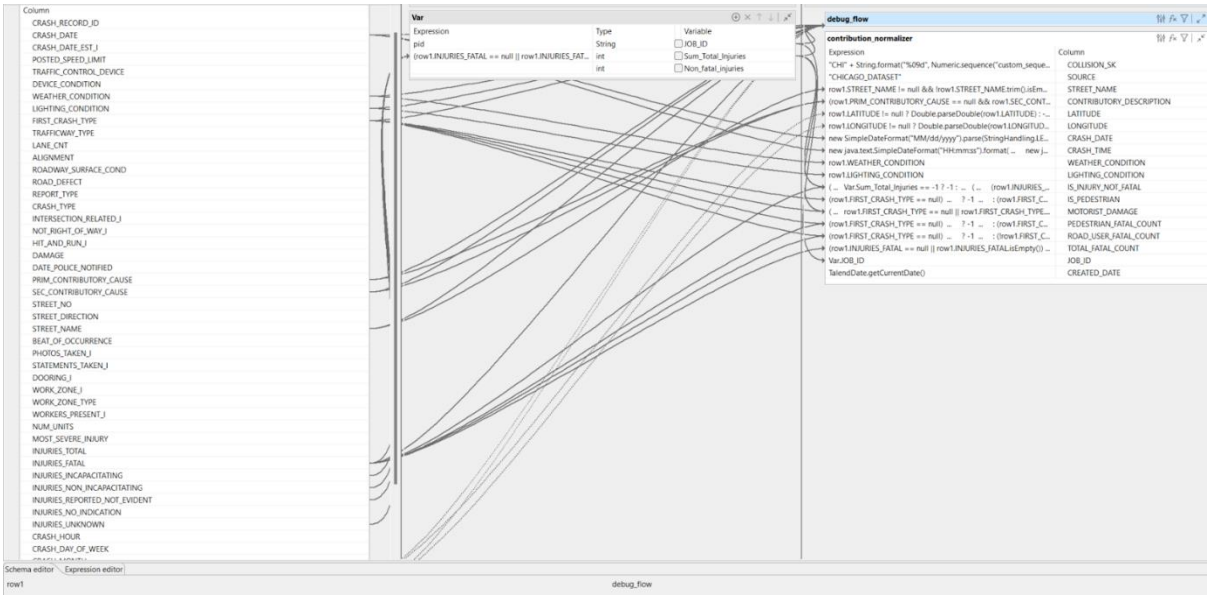


Transformation logic was added in the TMap component of Talend to address null values, clean improper data, and implement logic to meet business requirements by finalizing the columns from the actual dataset.

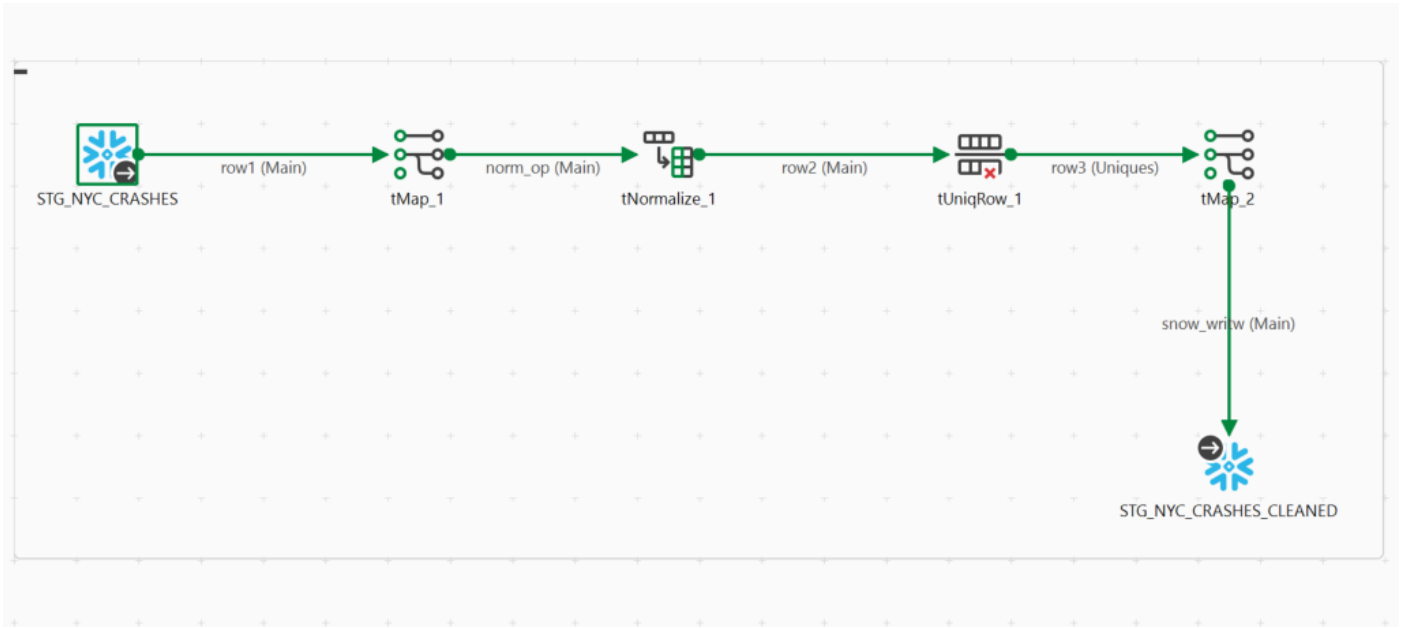
For Chicago:

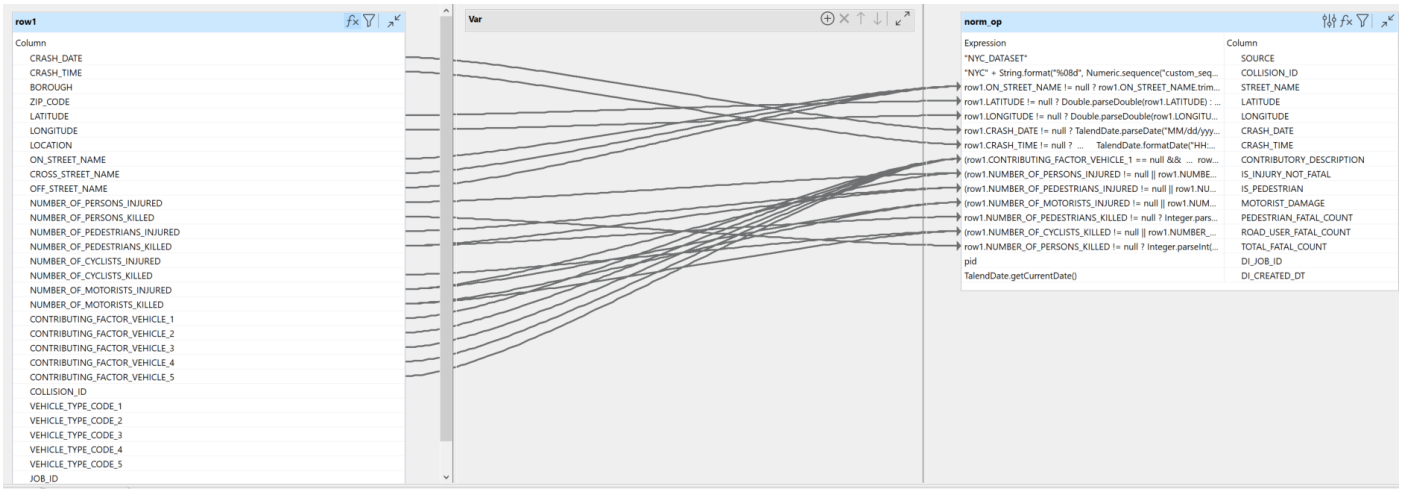


The Chicago dataset required additional transformations, including the application of the tNormalize component in Talend to address the contribution factors of accidents. As per the CR in phase-2, a transformation was implemented to assign proper codes to the contribution factors. The tUniq component was used to filter unique records and write them to the target table in Snowflake.

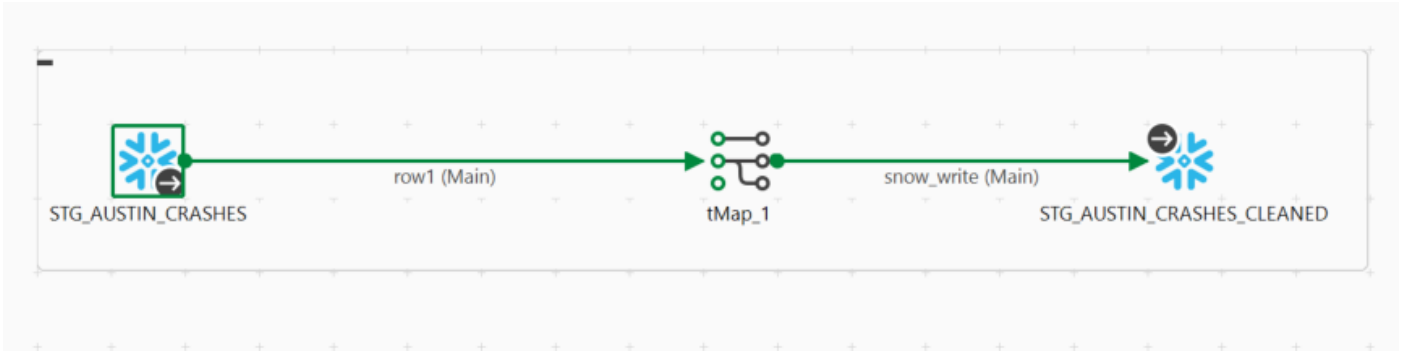


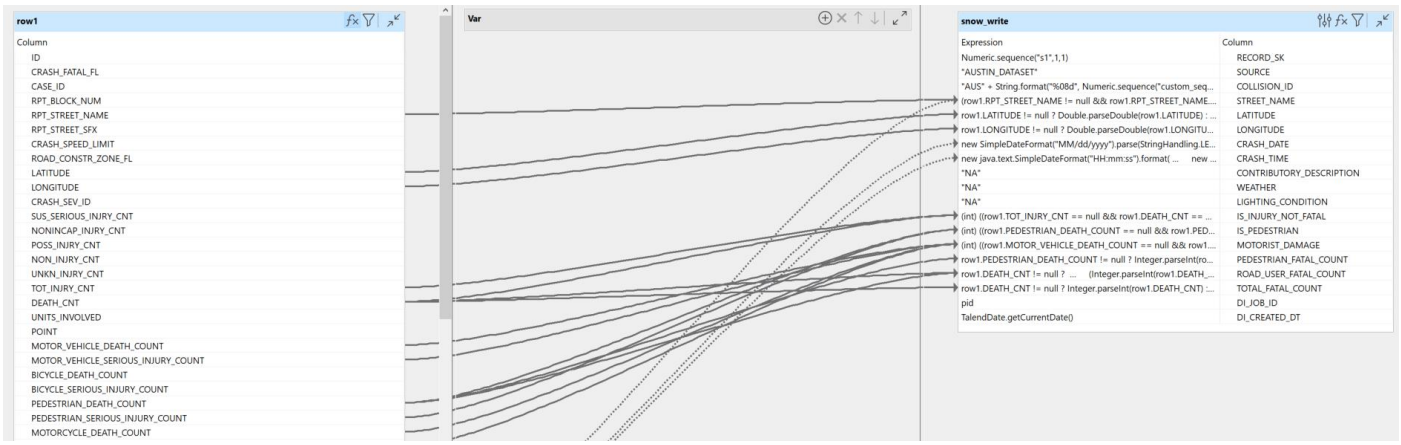
For NYC:





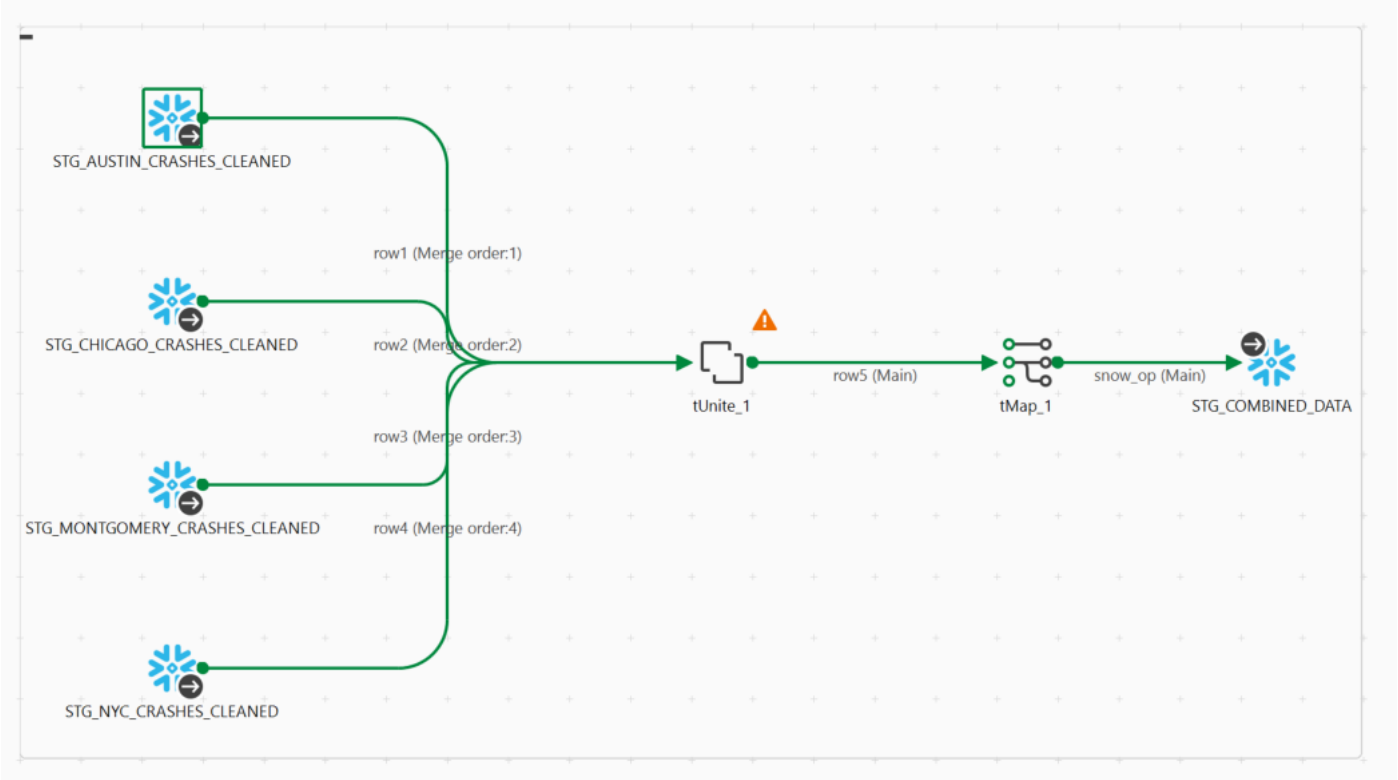
For Austin:





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
```

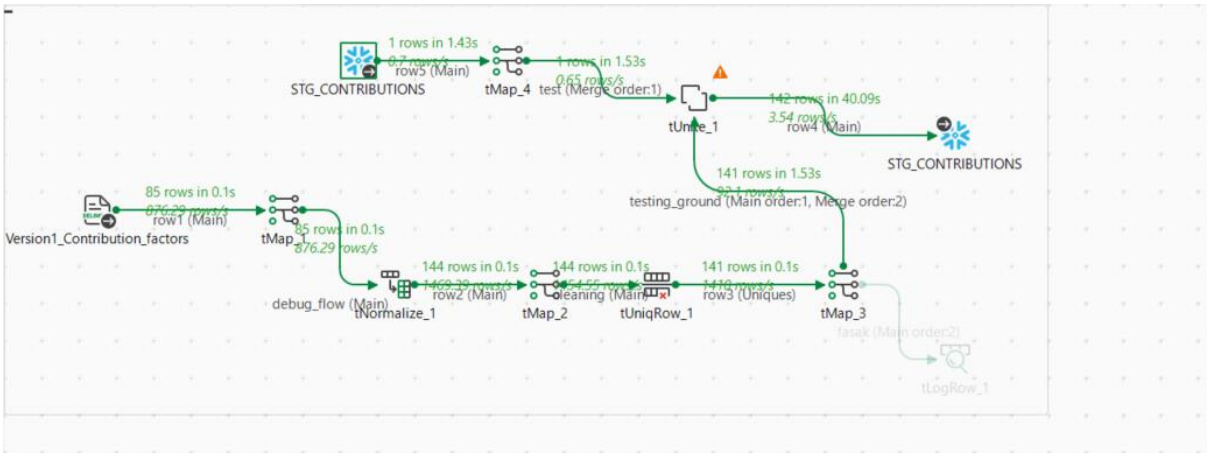
Results | Chart

	COUNT(*)
1	5077174

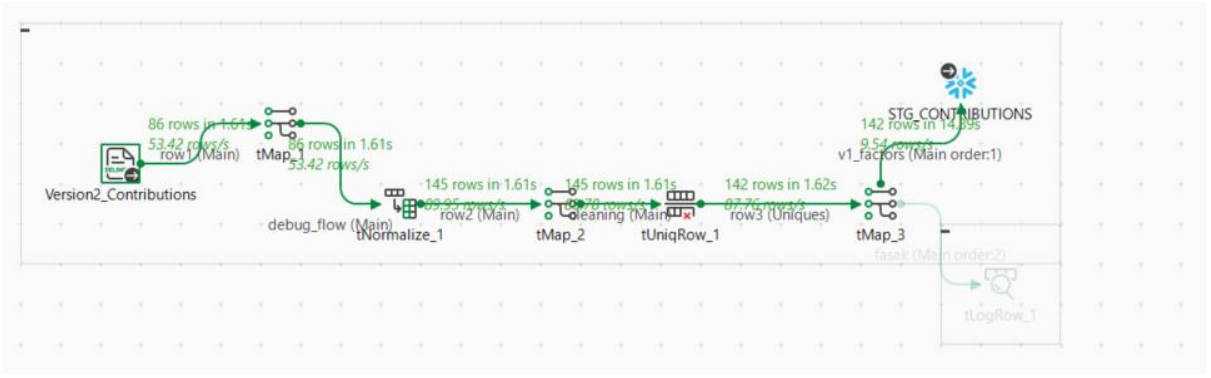
SCD Type-2 Implementation

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.



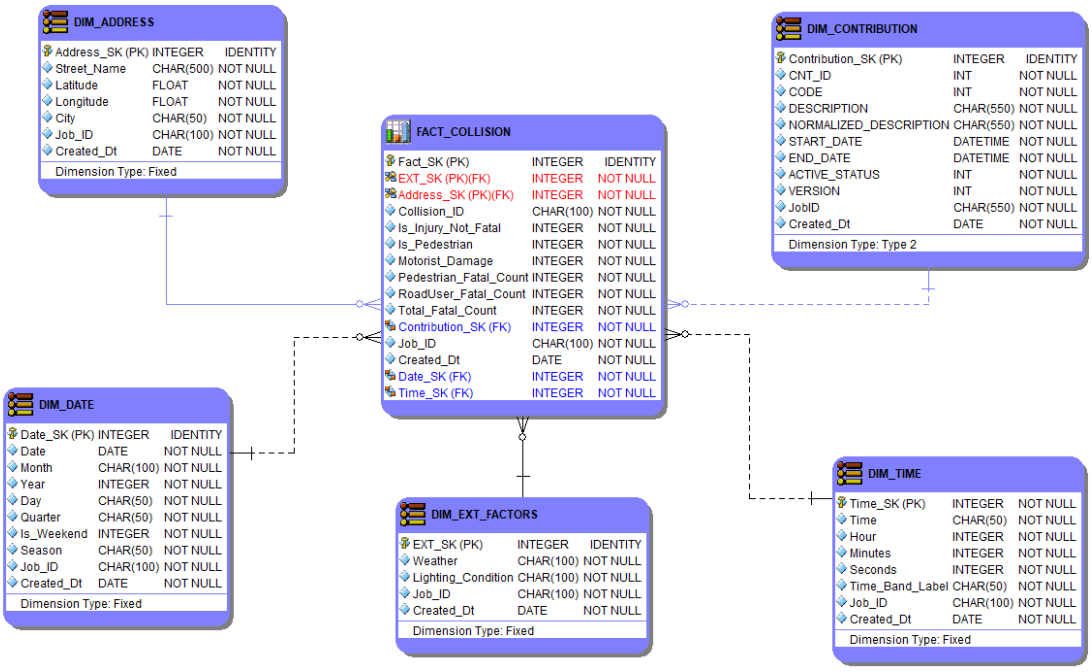
The tJDBCELT component in Talend was used to implement Slowly Changing Dimensions (SCD) for the datasets. Another pipeline was created to upsert any changes based on the version-2 file, targeting the STG_CONTRIBUTIONS table.



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

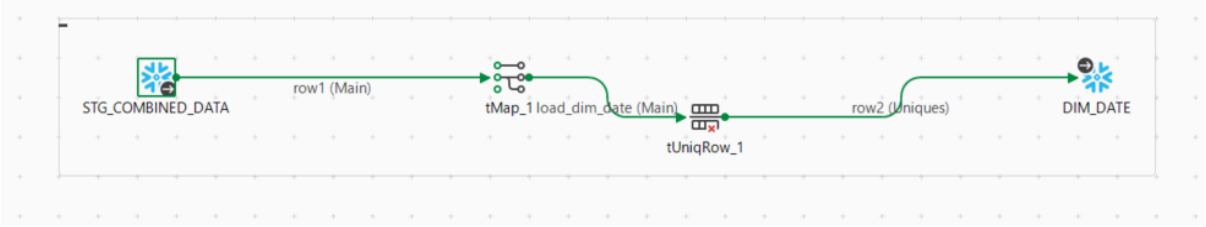
Using ER Studio, the dimensional model was designed, and DDL scripts for dimension tables and the fact table were set up.



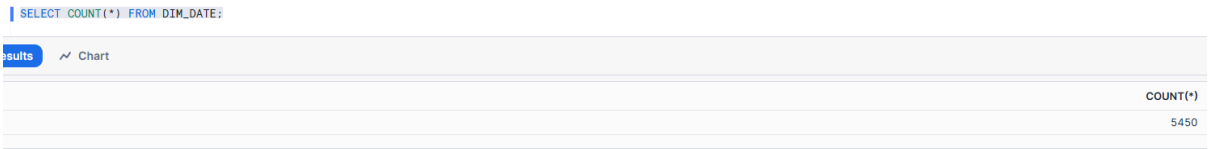
The relationships between entities have been assigned appropriately.

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.

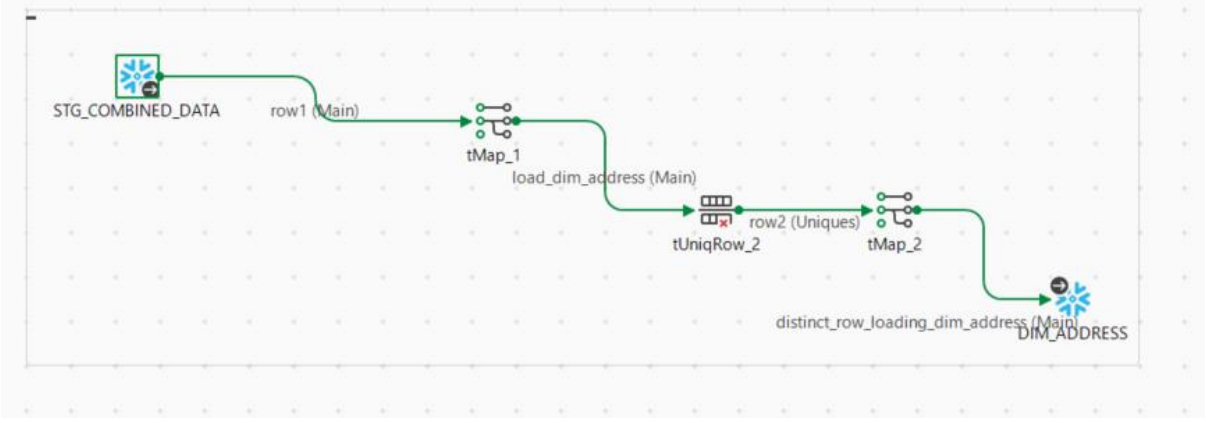
Pipeline to load DIM_DATE:



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



Pipeline to load DIM_ADDRESS:



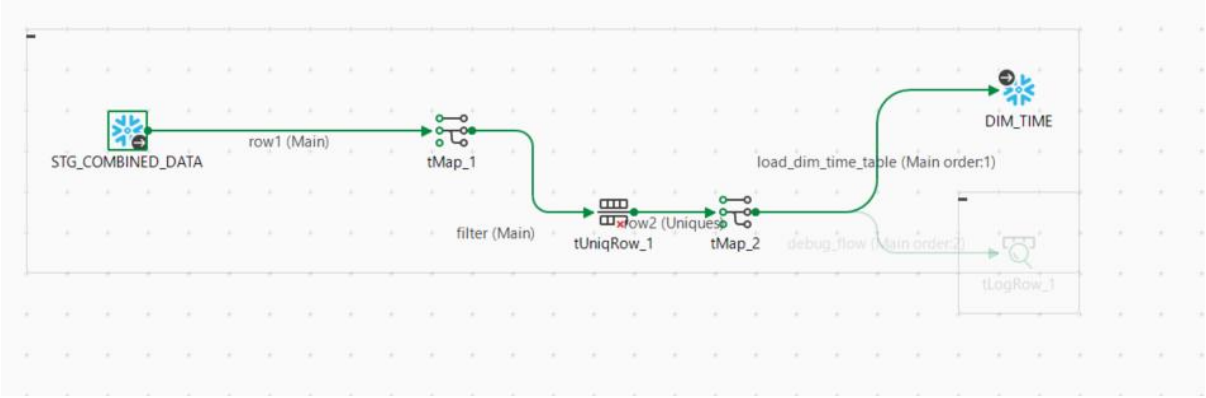
```
SELECT COUNT(*) FROM DIM_ADDRESS;
```

ults | 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.

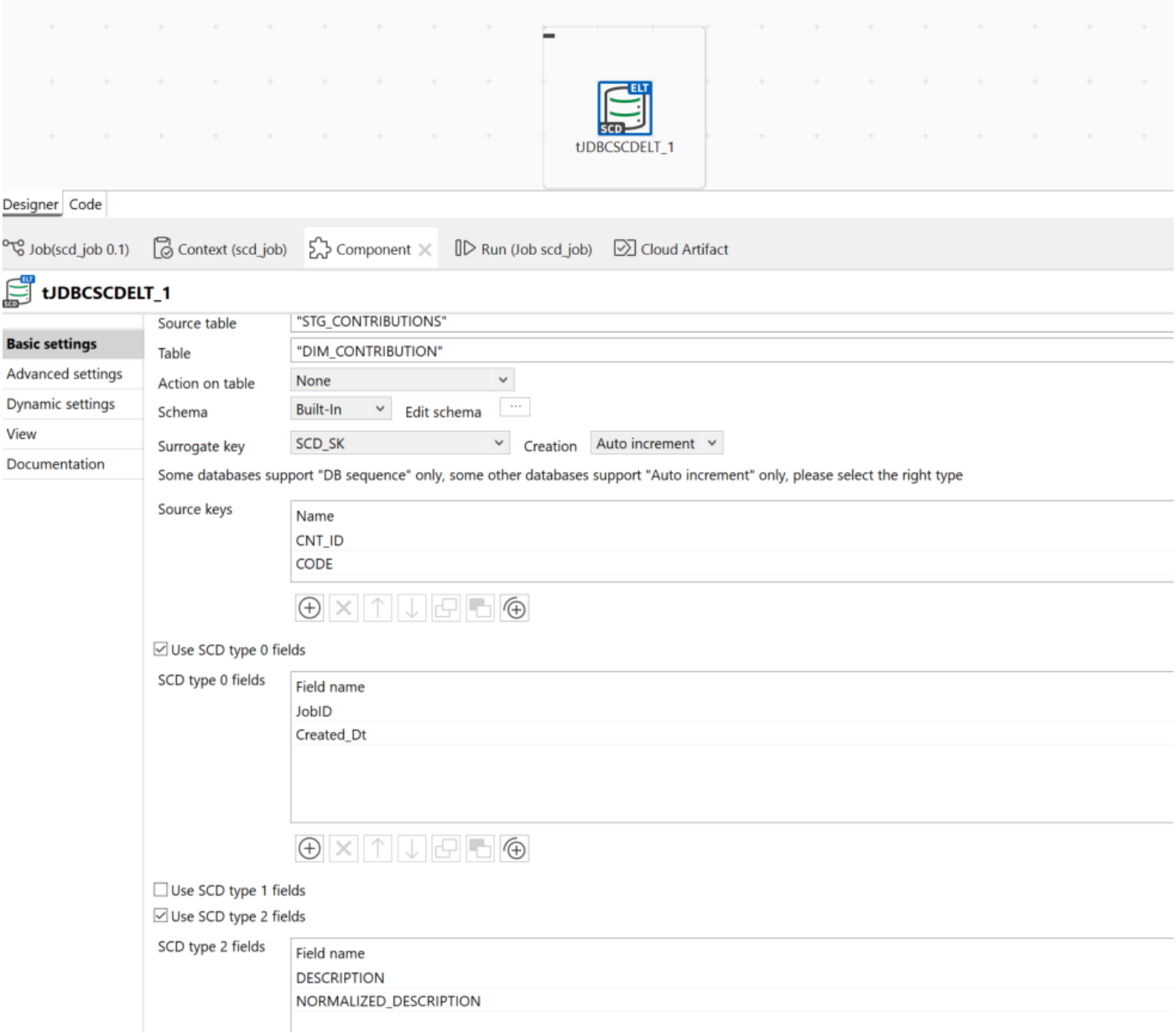



```
SELECT COUNT(*) FROM DIM_TIME;
```

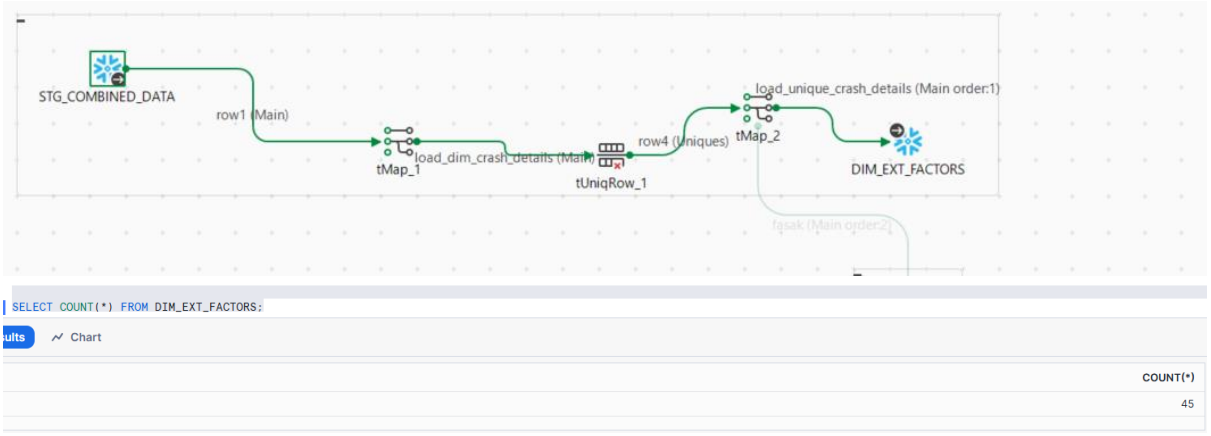
esults | Chart

COUNT(*)
1440

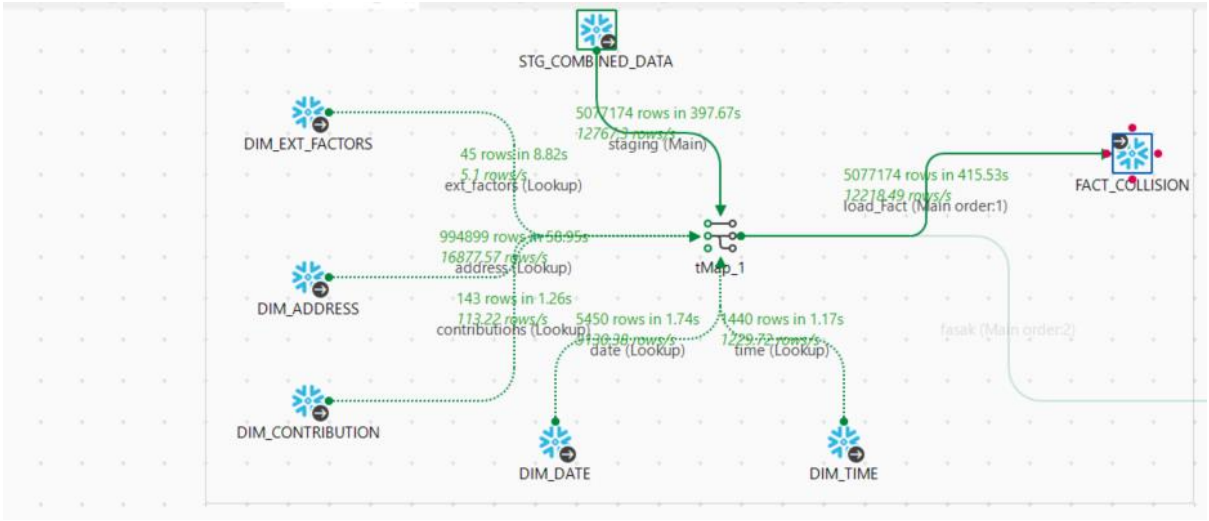
Pipeline to load DIM_CONTRIBUTIONS:

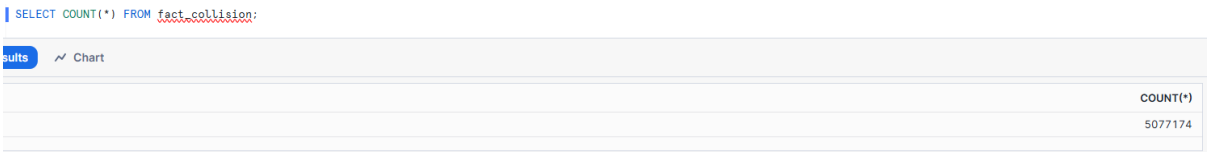


Pipeline to load DIM_EXTERNAL_FACTORS:



Pipeline to load Fact_Table:





The counts match the number of records in stg_combined_table.

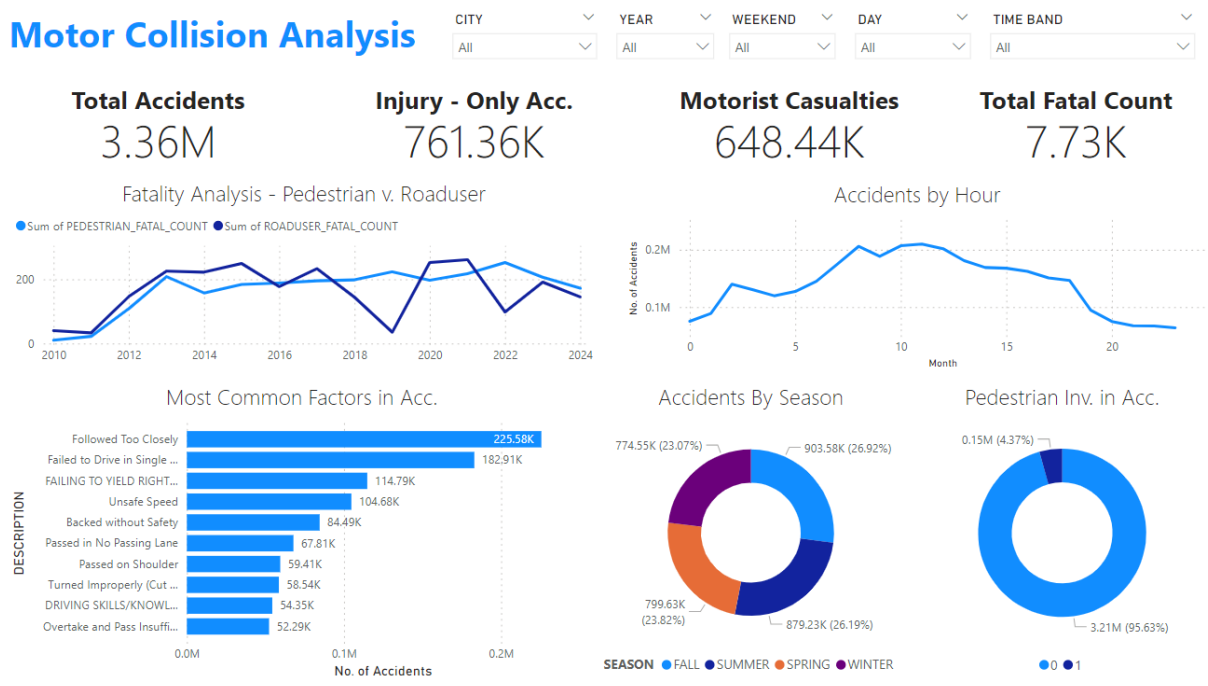
Visualizations:

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:



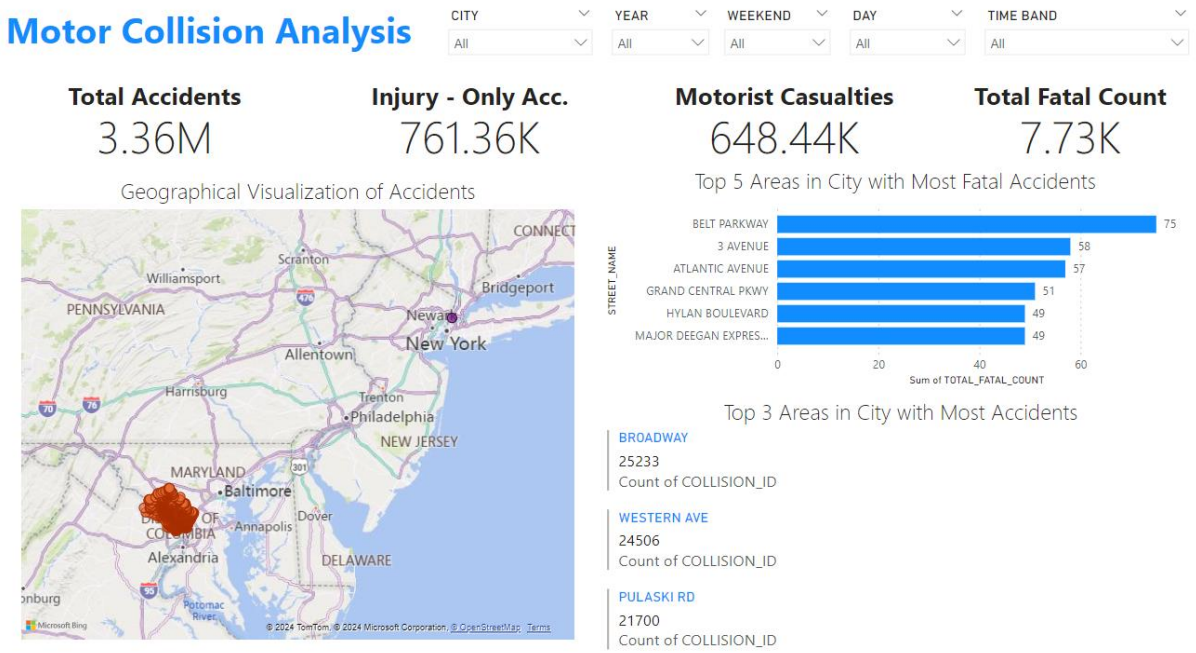
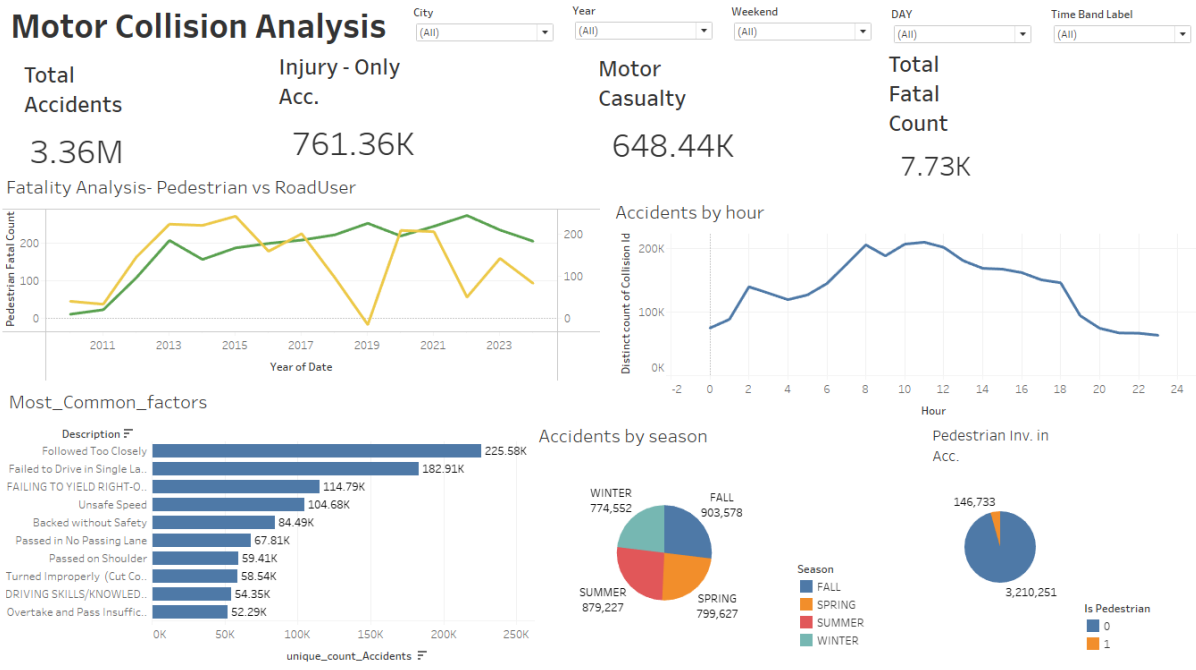

Motor Collision Analysis
CITY
All
YEAR
All
WEEKEND
All
DAY
All
TIME BAND
All
Total Accidents
3.36M
Injury - Only Acc.
761.36K
Motorist Casualties
648.44K
Total Fatal Count
7.73K
Geographical Visualization of Accidents
Top 5 Areas in City with Most Fatal Accidents
BELT PARKWAY 75
3 AVENUE 58
ATLANTIC AVENUE 57
GRAND CENTRAL PKWY 51
HYLAN BOULEVARD 49
MAJOR DEEGAN EXPRES... 49
STREET_NAME
Sum of TOTAL_FATAL_COUNT
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:

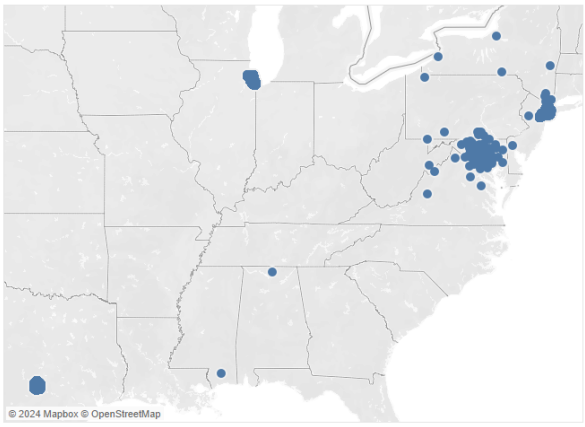

Motor Collision Analysis
City
(All)
Year
(All)
Weekend
(All)
DAY
(All)
Time Band Label
(All)
Total Accidents
3.36M
Injury - Only Acc.
761.36K
Motor Casualty
648.44K
Total Fatal Count
7.73K
Fatality Analysis- Pedestrian vs RoadUser
Pedestrian Fatal Count
Year of Date
Accidents by hour
Distinct count of Collision Id
Hour
Most_Common_factors
Description
Followed Too Closely 225.58K
Failed to Drive in Single La.. 182.91K
FAILING TO YIELD RIGHT-O.. 114.79K
Unsafe Speed 104.68K
Backed without Safety 84.49K
Passed in No Passing Lane 67.81K
Passed on Shoulder 59.41K
Turned Improperly (Cut Co.. 58.54K
DRIVING SKILLS/KNOWLED.. 54.35K
Overtake and Pass Insuffic.. 52.29K
unique_count_Accidents
Accidents by season
WINTER 774,552
FALL 903,578
SUMMER 879,227
SPRING 799,627
Season
FALL
SPRING
SUMMER
WINTER
Pedestrian Inv. in Acc.
146,733
3,210,251
Is Pedestrian
0
1

Motor Collision Analysis

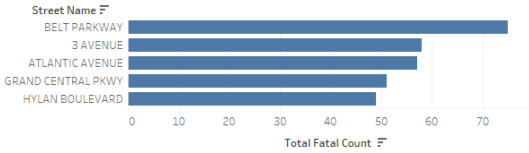
City: (All) | Year: (All) | Weekend: (All) | DAY: (All) | Time Band Label: (All)

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