Final Team Project

Motor Vehicle Collision/Crash Report in Chicago, New York, and Austin

Group 6

- Sathyavarthan Balachandar
- Praveen Jagadishan
- Mithali Manjunath
- Manish Choudhary

Table of Contents

1. Introduction

1.1. Project Overview

This assignment involves the analysis of vehicle collision/crash data in the cities of New York, Austin, and Chicago from the government data of the cities. This documentation provides a comprehensive overview of the project's key components, data sources, and deliverables.

This project involves the flat file data download. Staging the data using Talend and performing the data profiling using Y data profiling. Cleaning the data using Talend and carrying out the dimensional modeling, dimensional loading, and fact loading operations. Performing the visualizations from all the dimensions and facts.

1.2. Data Source Description

Chicago -

Crash data shows information about each traffic crash on city streets within the City of Chicago limits and under the jurisdiction of the Chicago Police Department (CPD). Data are shown as is from the electronic crash reporting system (E-Crash) at CPD, excluding any personally identifiable information. Records are added to the data portal when a crash report is finalized or when amendments are made to an existing report in E-Crash. Data from E-Crash are available for some police districts in 2015, but citywide data are not available until September 2017. About half of all crash reports, mostly minor crashes, are self-reported at the police district by the driver(s) involved, and the other half are recorded at the scene by the police officer responding to the crash. Many of the crash parameters, including street condition data, weather conditions, and posted speed limits, are recorded by the reporting officer based on the best available information at the time, but many of these may disagree with posted information or other assessments on road conditions. If any new or updated information on a crash is received, the reporting officer may amend the crash report later. A traffic crash within the city limits for which CPD is not the responding police agency typically crashes on interstate highways, freeway ramps, and local roads along the City boundary, are excluded from this dataset.

New York –

The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police-reported motor vehicle collisions in NYC. The police report (MV104-AN) is required to be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage.

Austin–

Crash data is obtained from the Texas Department of Transportation (TXDOT) Crash Record Information System (CRIS) database, which is populated by reports submitted by Texas Peace Officers throughout the state, including the Austin Police Department (APD), and maintained by TXDOT. This dataset contains crash-level records for crashes that have occurred in the last ten years. Crash data may take several days or weeks to be initially provided and finalized as it is furnished to the Austin Transportation & Public Works Department, therefore a two-week delay is implemented to help ensure more accurate and complete results.

2. Data Preparation

2.1 Obtain Data Files from the source

The flat file data is downloaded from the following link.

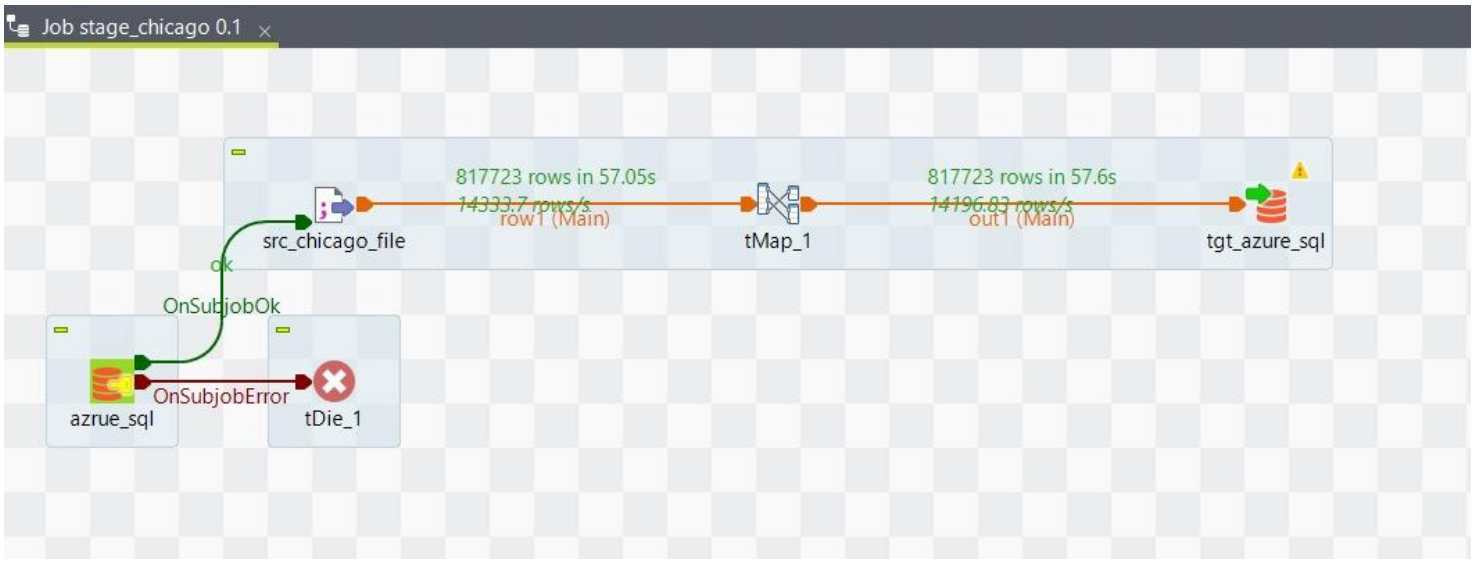
Chicago: https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data

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

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

2.2 Staging

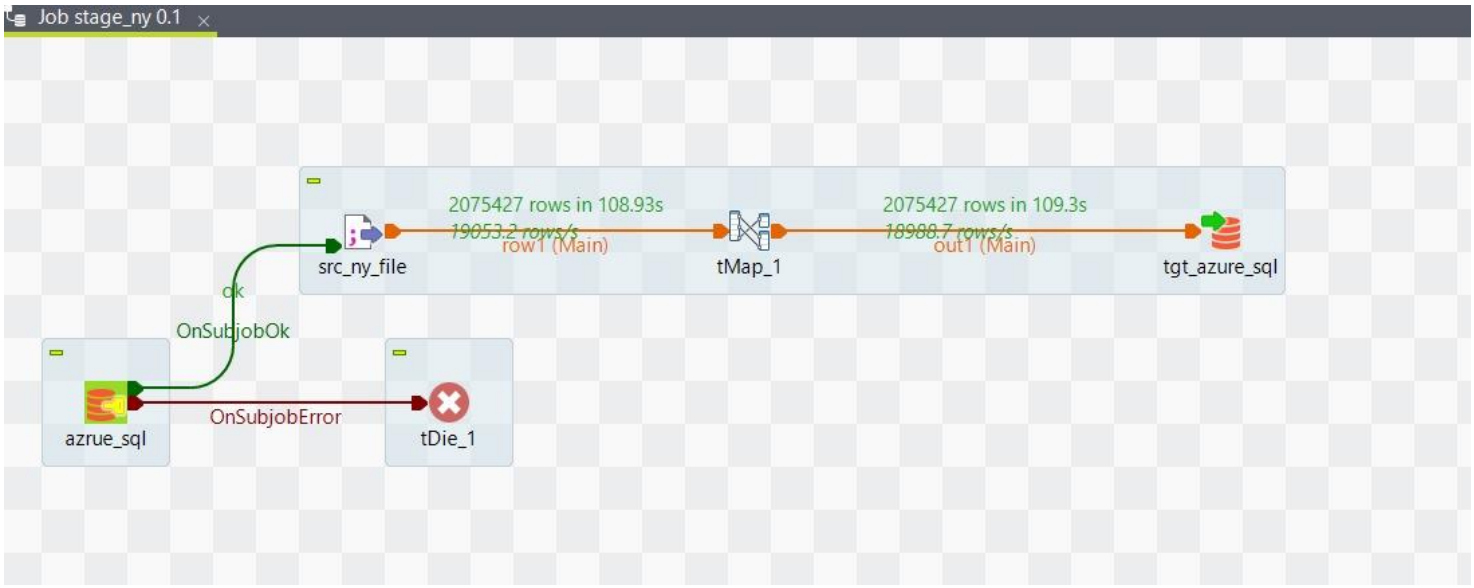
Chicago:



Here we are importing the data from the source TSV file and performing data cleansing, date conversion, the addition of Create_Date column, and selecting only specific columns operations before staging to as a staging entity in the database. Additionally, an output of the stage table in the form of a CSV is also taken.

List the total time your job took to complete: 5.6 seconds

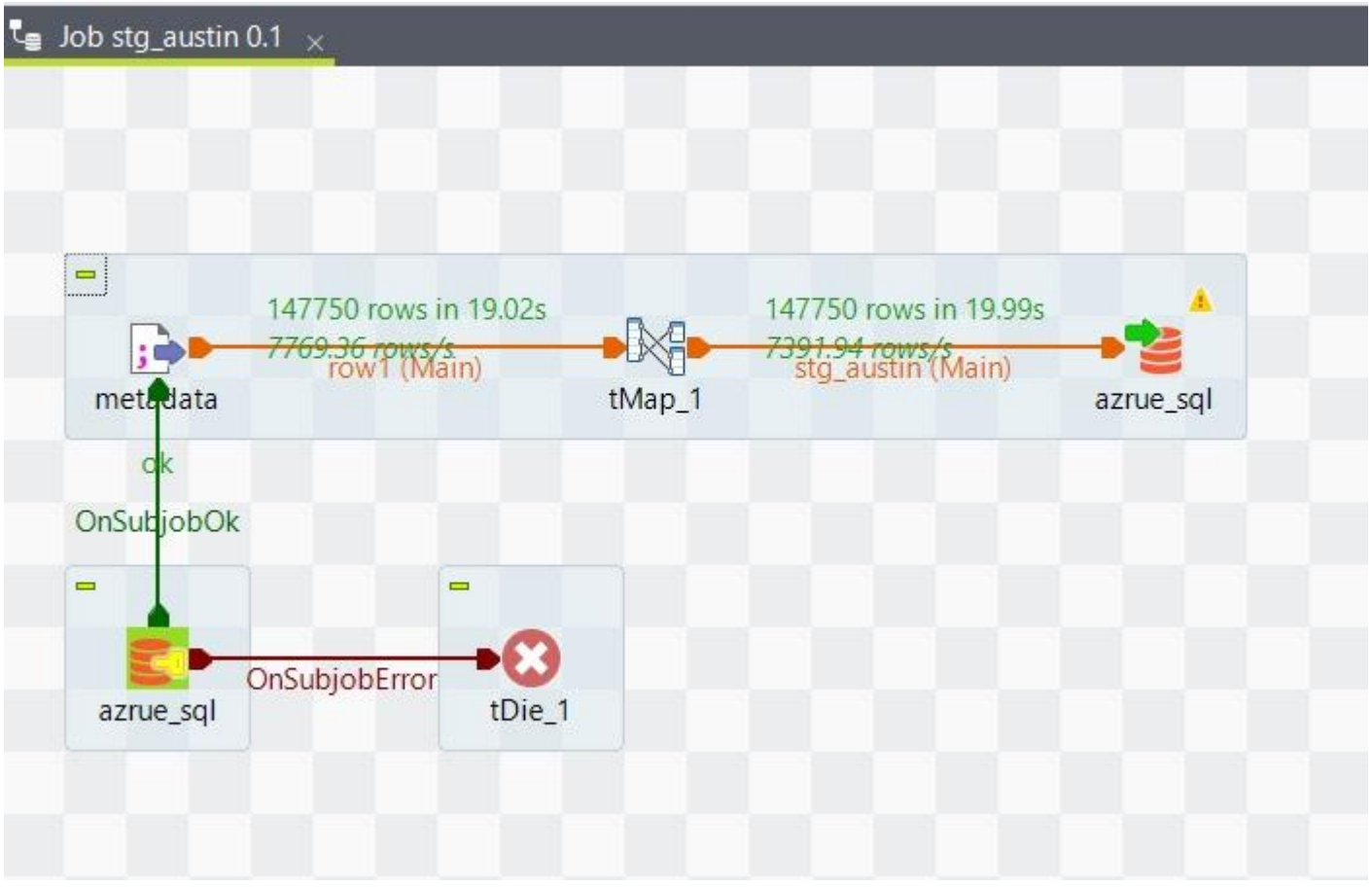
New York:



Here we are importing the data from the source TSV file and performing date conversion, the addition of the Create_Date column, and selecting only specific column operations before staging to as a staging entity in the database.

List the total time your job took to complete: 109.3s

Austin:



Here we are importing the data from the source TSV file and performing date conversion, the addition of the Create_Date column, and selecting only specific column operations before staging to as a staging entity in the database.

List the total time your job took to complete: 19.99 seconds

2.2 Y Data Profiling

For Chicago:

The Y data profiling is performed for the Chicago data set using the Y data profiling and can be accessed using the HTML file attached to the compressed file

CRASH_RECORD_ID has no null values and is 100% distinct

For New York:

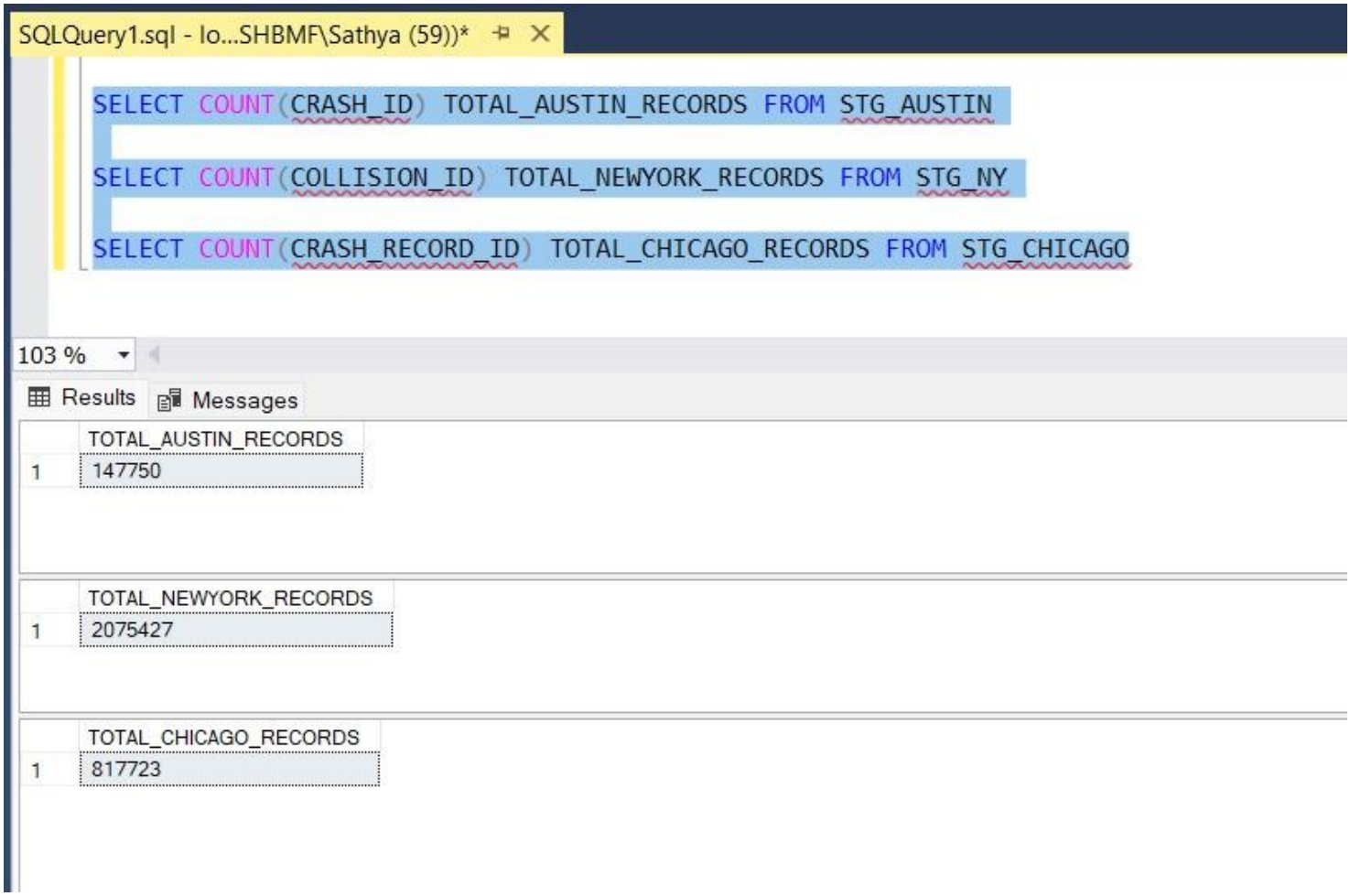
The Y data profiling is performed for the New York data set using the Y data profiling and can be accessed using the HTML file attached to the compressed file

COLLISION_ID has no null values and is 100% distinct

For Austin:

crash_id has no null values and is 100% distinct

2.3 Data Validation



3. Dimensional Data Modelling

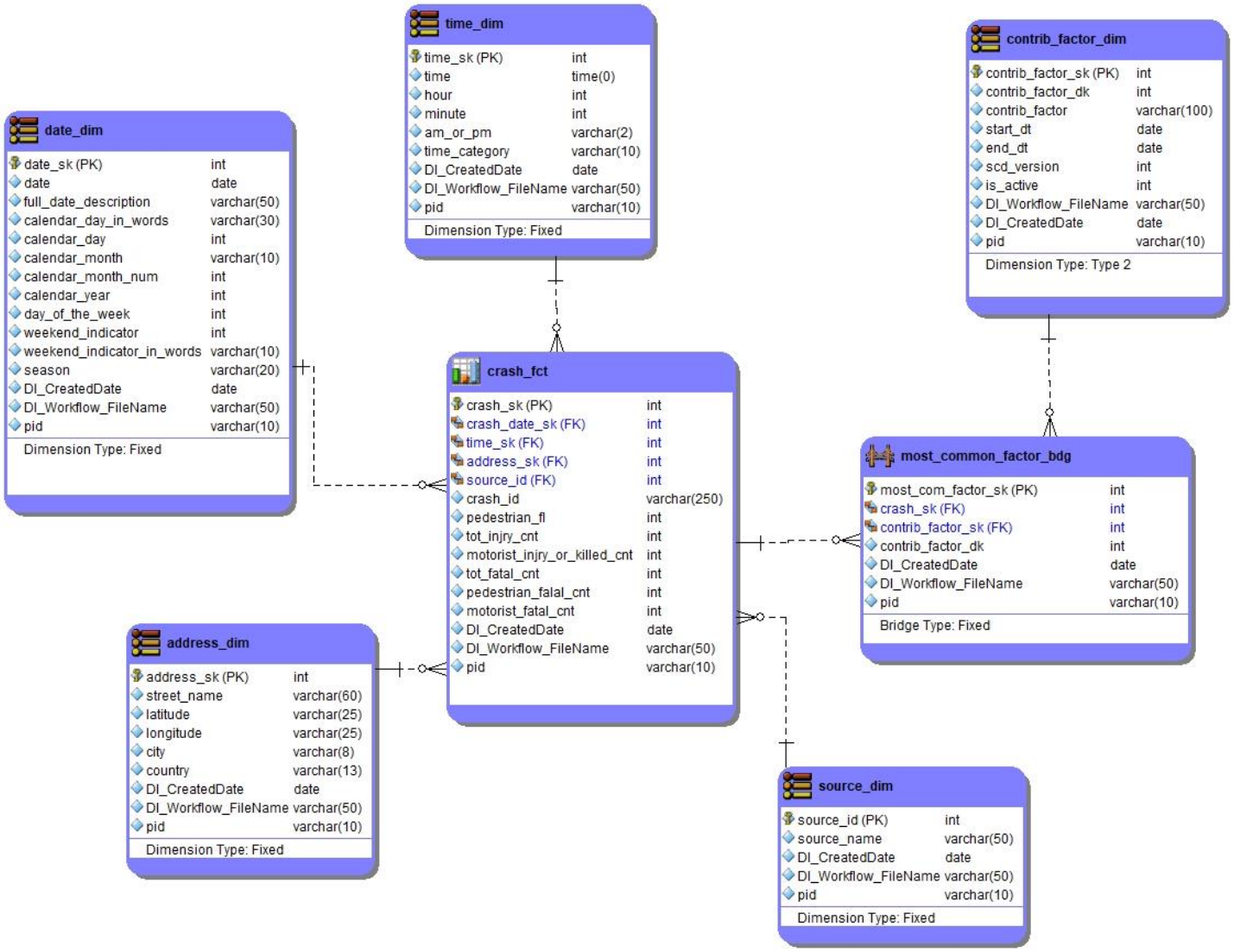
3.1. Identification of Facts and Dimensions

The Dimensional Data model created has the dimensions time_dim, contrib_factor_dim, source_dim, date_dim, and address_dim, bridge table as most_common_factor_bdg and we take the fact table crash_fct.

3.2. Grain Definition

The grains are of Transaction level.

3.3 ER/Studio Data Model



3.4 Handling Data Inconsistencies

Chicago:

Street_No and Street_Name is concatenated, where Street_No has no null values but Street_Name has 1 missing value Street_No has zeros and Street_Name has values called unknown and unknown avenue this is addressed in the further staging table.

Latitude and Longitude have some null values and also have some spaces which are replaced by 0

Country/ City and columns were not there so it is populated manually

PRIM_CONTRIBUTORY_CAUSE and SEC_CONTRIBUTORY_CAUSE have no nulls

Crash_Time is a derived column extracted from crash_date by splitting the time and date and inserting only the time in the crash_time column.

New York:

On_Street_Name has some null values which are replaced by the Off_Street_Name on the same index as the null value in On_Street_Name

Latitude and Longitude have some null values and also have some spaces which are replaced by 0

Country/ City and columns were not there so it is populated manually

CONTRIBUTING_FACTOR_VEHICLE_1, CONTRIBUTING_FACTOR_VEHICLE_2, CONTRIBUTING_FACTOR_VEHICLE_3, CONTRIBUTING_FACTOR_VEHICLE_4, and CONTRIBUTING_FACTOR_VEHICLE_5 has blank spaces which is replaced by the nulls

Austin:

Street_Nbr and Street_Name are concatenated, where Street_Nbr has no null values but Street_Name has 1 missing value Street_Nbr has zeros and Street_Name has values called unknown and unknown avenue this is addressed in the further staging table.

Latitude and Longitude have some null values and also have some spaces which are replaced by 0

Country/ City and columns were not there so it is populated manually

contrib_factr_p1_id, and contrib_factr_p2_id have blank spaces which are replaced by the nulls

3.5 DDL Scripts of the dimensional and fact table

```
CREATE TABLE address_dim(
    address_sk              int             IDENTITY(1,1),
    street_name             varchar(60)     NOT NULL,
    latitude                varchar(25)     NULL,
    longitude               varchar(25)     NULL,
    city                    varchar(8)      NOT NULL,
    country                 varchar(13)     NOT NULL,
    DI_CreatedDate          date            NULL,
    DI_Workflow_FileName    varchar(50)     NULL,
    pid                     varchar(10)     NULL,
    CONSTRAINT PK5 PRIMARY KEY NONCLUSTERED (address_sk)
)

go


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

/*
 * TABLE: contrib_factor_dim
 */

CREATE TABLE contrib_factor_dim(
    contrib_factor_sk       int             IDENTITY(1,1),
    contrib_factor_dk       int             NOT NULL,
    contrib_factor          varchar(100)    NULL,
    start_dt                date            NULL,
    end_dt                  date            NULL,
    scd_version             int             NULL,
    is_active               int             NULL,
    DI_Workflow_FileName    varchar(50)     NULL,
    DI_CreatedDate          date            NULL,
    pid                     varchar(10)     NULL,
    CONSTRAINT PK2 PRIMARY KEY NONCLUSTERED (contrib_factor_sk)
)

go


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

/*
 * TABLE: crash_fct
 */

CREATE TABLE crash_fct(
    crash_sk                        int             IDENTITY(1,1),
    crash_date_sk                   int             NOT NULL,
```

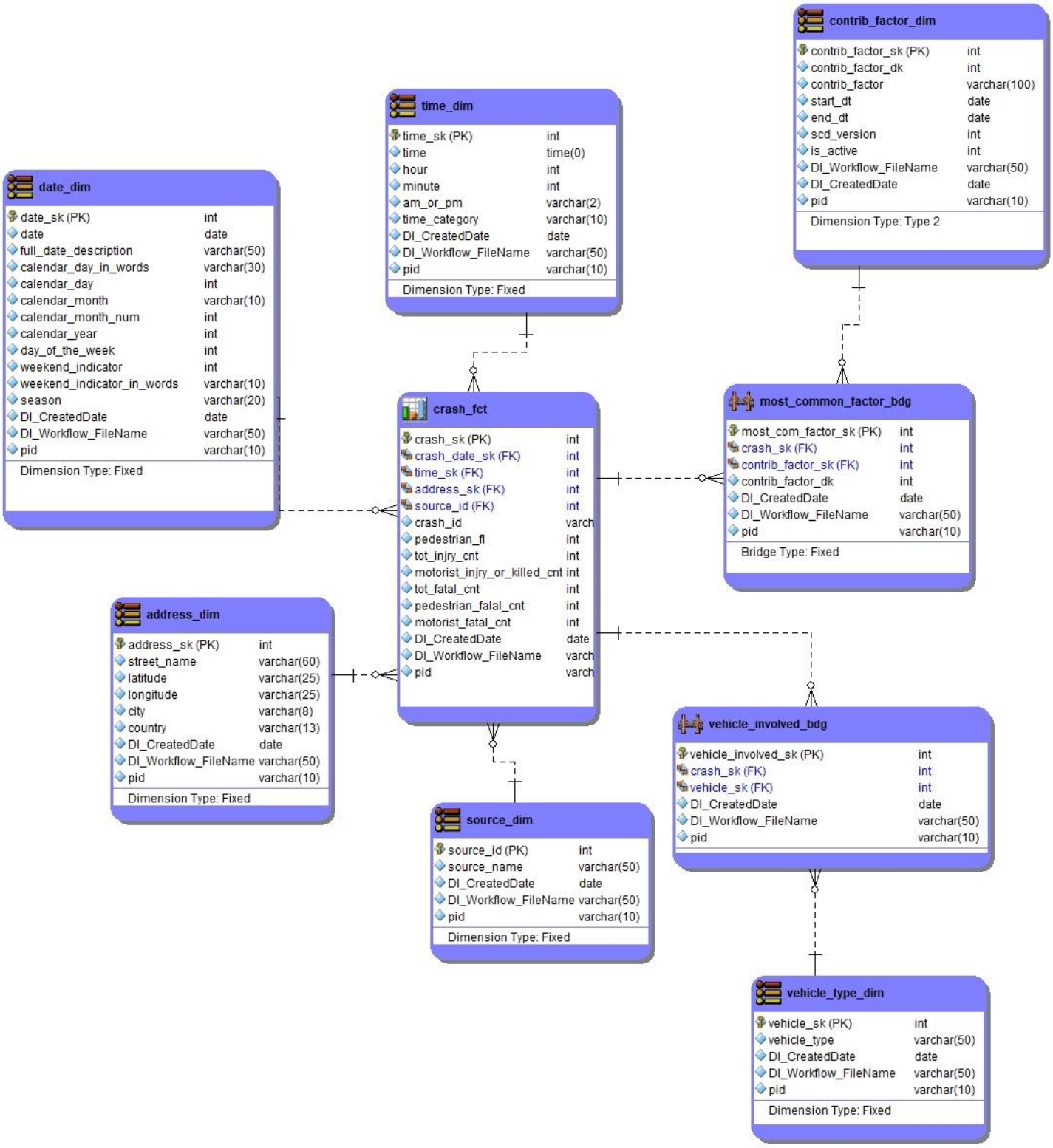
```
    time_sk                         int             NOT NULL,
    address_sk                      int             NOT NULL,
    source_id                       int             NOT NULL,
    crash_id                        varchar(250)    NOT NULL,
    pedestrian_fl                   int             NOT NULL,
    tot_injry_cnt                   int             NOT NULL,
    motorist_injry_or_killed_cnt    int             NOT NULL,
    tot_fatal_cnt                   int             NOT NULL,
    pedestrian_falal_cnt            int             NOT NULL,
    motorist_fatal_cnt              int             NOT NULL,
    DI_CreatedDate                  date            NULL,
    DI_Workflow_FileName            varchar(50)     NULL,
    pid                             varchar(10)     NULL,
    CONSTRAINT PK4 PRIMARY KEY NONCLUSTERED (crash_sk)
)

go


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

/*
 * TABLE: date_dim
 */

CREATE TABLE date_dim(
    date_sk                     int             NOT NULL,
    date                        date            NULL,
    full_date_description       varchar(50)     NULL,
    calendar_day_in_words       varchar(30)     NULL,
    calendar_day                int             NULL,
    calendar_month              varchar(10)     NULL,
    calendar_month_num          int             NULL,
    calendar_year               int             NULL,
    day_of_the_week             int             NULL,
    weekend_indicator           int             NULL,
    weekend_indicator_in_words  varchar(10)     NULL,
    season                      varchar(20)     NULL,
    DI_CreatedDate              date            NULL,
    DI_Workflow_FileName        varchar(50)     NULL,
    pid                         varchar(10)     NULL,
    CONSTRAINT PK2_1 PRIMARY KEY NONCLUSTERED (date_sk)
)

go


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

/*
 * TABLE: most_common_factor_bdg
 */

CREATE TABLE most_common_factor_bdg(
    most_com_factor_sk      int             IDENTITY(1,1),
    crash_sk                int             NOT NULL,
    contrib_factor_sk       int             NOT NULL,
    contrib_factor_dk       int             NOT NULL,
```

```
    DI_CreatedDate          date            NULL,
    DI_Workflow_FileName    varchar(50)     NULL,
    pid                     varchar(10)     NULL,
    CONSTRAINT PK3 PRIMARY KEY NONCLUSTERED (most_com_factor_sk)
)

go


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

/*
 * TABLE: source_dim
 */

CREATE TABLE source_dim(
    source_id               int             NOT NULL,
    source_name             varchar(50)     NULL,
    DI_CreatedDate          date            NULL,
    DI_Workflow_FileName    varchar(50)     NULL,
    pid                     varchar(10)     NULL,
    CONSTRAINT PK8 PRIMARY KEY NONCLUSTERED (source_id)
)

go


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

/*
 * TABLE: time_dim
 */

CREATE TABLE time_dim(
    time_sk                 int             NOT NULL,
    time                    time(0)         NULL,
    hour                    int             NULL,
    minute                  int             NULL,
    am_or_pm                varchar(2)      NULL,
    time_category           varchar(10)     NULL,
    DI_CreatedDate          date            NULL,
    DI_Workflow_FileName    varchar(50)     NULL,
    pid                     varchar(10)     NULL,
    CONSTRAINT PK1 PRIMARY KEY NONCLUSTERED (time_sk)
)

go


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

/*
 * TABLE: crash_fct
 */
```

```
ALTER TABLE crash_fct ADD CONSTRAINT Refdate_dim2
    FOREIGN KEY (crash_date_sk)
    REFERENCES date_dim(date_sk)
go

ALTER TABLE crash_fct ADD CONSTRAINT Reftime_dim8
    FOREIGN KEY (time_sk)
    REFERENCES time_dim(time_sk)
go

ALTER TABLE crash_fct ADD CONSTRAINT Refaddress_dim10
    FOREIGN KEY (address_sk)
    REFERENCES address_dim(address_sk)
go

ALTER TABLE crash_fct ADD CONSTRAINT Refsource_dim19
    FOREIGN KEY (source_id)
    REFERENCES source_dim(source_id)
go


/*
 * TABLE: most_common_factor_bdg
 */

ALTER TABLE most_common_factor_bdg ADD CONSTRAINT Refcrash_fct17
    FOREIGN KEY (crash_sk)
    REFERENCES crash_fct(crash_sk)
go

ALTER TABLE most_common_factor_bdg ADD CONSTRAINT Refcontrib_factor_dim18
    FOREIGN KEY (contrib_factor_sk)
    REFERENCES contrib_factor_dim(contrib_factor_sk)
Go
```

4. Dimensional Data Modelling

4.1 Altered Dimensional Model



4.2 DDL Scripts of the dimensional and fact table

```
CREATE TABLE address_dim(
    address_sk              int            IDENTITY(1,1),
    street_name             varchar(60)    NOT NULL,
    latitude                varchar(25)    NULL,
    longitude               varchar(25)    NULL,
    city                    varchar(8)     NOT NULL,
    country                 varchar(13)    NOT NULL,
    DI_CreatedDate          date           NULL,
    DI_Workflow_FileName    varchar(50)    NULL,
    pid                     varchar(10)    NULL,
    CONSTRAINT PK5 PRIMARY KEY NONCLUSTERED (address_sk)
)

go



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

/*
 * TABLE: contrib_factor_dim
 */

CREATE TABLE contrib_factor_dim(
    contrib_factor_sk       int             IDENTITY(1,1),
    contrib_factor_dk       int             NOT NULL,
    contrib_factor          varchar(100)    NULL,
    start_dt                date            NULL,
    end_dt                  date            NULL,
    scd_version             int             NULL,
    is_active               int             NULL,
```

```
    DI_Workflow_FileName    varchar(50)     NULL,
    DI_CreatedDate          date            NULL,
    pid                     varchar(10)     NULL,
    CONSTRAINT PK2 PRIMARY KEY NONCLUSTERED (contrib_factor_sk)
)

go


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

/*
 * TABLE: crash_fct
 */

CREATE TABLE crash_fct(
    crash_sk                        int             IDENTITY(1,1),
    crash_date_sk                   int             NOT NULL,
    time_sk                         int             NOT NULL,
    address_sk                      int             NOT NULL,
    source_id                       int             NOT NULL,
    crash_id                        varchar(250)    NOT NULL,
    pedestrian_fl                   int             NOT NULL,
    tot_injry_cnt                   int             NOT NULL,
    motorist_injry_or_killed_cnt    int             NOT NULL,
    tot_fatal_cnt                   int             NOT NULL,
    pedestrian_falal_cnt            int             NOT NULL,
    motorist_fatal_cnt              int             NOT NULL,
    DI_CreatedDate                  date            NULL,
    DI_Workflow_FileName            varchar(50)     NULL,
    pid                             varchar(10)     NULL,
    CONSTRAINT PK4 PRIMARY KEY NONCLUSTERED (crash_sk)
)
```

```
go


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

/*
 * TABLE: date_dim
 */

CREATE TABLE date_dim(
    date_sk                      int            NOT NULL,
    date                         date           NULL,
    full_date_description        varchar(50)    NULL,
    calendar_day_in_words        varchar(30)    NULL,
    calendar_day                 int            NULL,
    calendar_month               varchar(10)    NULL,
    calendar_month_num           int            NULL,
    calendar_year                int            NULL,
    day_of_the_week              int            NULL,
    weekend_indicator            int            NULL,
    weekend_indicator_in_words   varchar(10)    NULL,
    season                       varchar(20)    NULL,
    DI_CreatedDate               date           NULL,
    DI_Workflow_FileName         varchar(50)    NULL,
    pid                          varchar(10)    NULL,
    CONSTRAINT PK2_1 PRIMARY KEY NONCLUSTERED (date_sk)
)

go


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

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

/*
 * TABLE: most_common_factor_bdg
 */

CREATE TABLE most_common_factor_bdg(
    most_com_factor_sk      int             IDENTITY(1,1),
    crash_sk                int             NOT NULL,
    contrib_factor_sk       int             NOT NULL,
    contrib_factor_dk       int             NOT NULL,
    DI_CreatedDate          date            NULL,
    DI_Workflow_FileName    varchar(50)     NULL,
    pid                     varchar(10)     NULL,
    CONSTRAINT PK3 PRIMARY KEY NONCLUSTERED (most_com_factor_sk)
)

go


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

/*
 * TABLE: source_dim
 */

CREATE TABLE source_dim(
    source_id               int             IDENTITY(1,1),
    source_name             varchar(50)     NULL,
    DI_CreatedDate          date            NULL,
```

```
    DI_Workflow_FileName    varchar(50)    NULL,
    pid                     varchar(10)    NULL,
    CONSTRAINT PK8 PRIMARY KEY NONCLUSTERED (source_id)
)

go


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

/*
 * TABLE: time_dim
 */

CREATE TABLE time_dim(
    time_sk                 int            NOT NULL,
    time                    time(0)        NULL,
    hour                    int            NULL,
    minute                  int            NULL,
    am_or_pm                varchar(2)     NULL,
    time_category           varchar(10)    NULL,
    DI_CreatedDate          date           NULL,
    DI_Workflow_FileName    varchar(50)    NULL,
    pid                     varchar(10)    NULL,
    CONSTRAINT PK1 PRIMARY KEY NONCLUSTERED (time_sk)
)

go


IF OBJECT_ID('time_dim') IS NOT NULL
    PRINT '<<< CREATED TABLE time_dim >>>'
ELSE
```

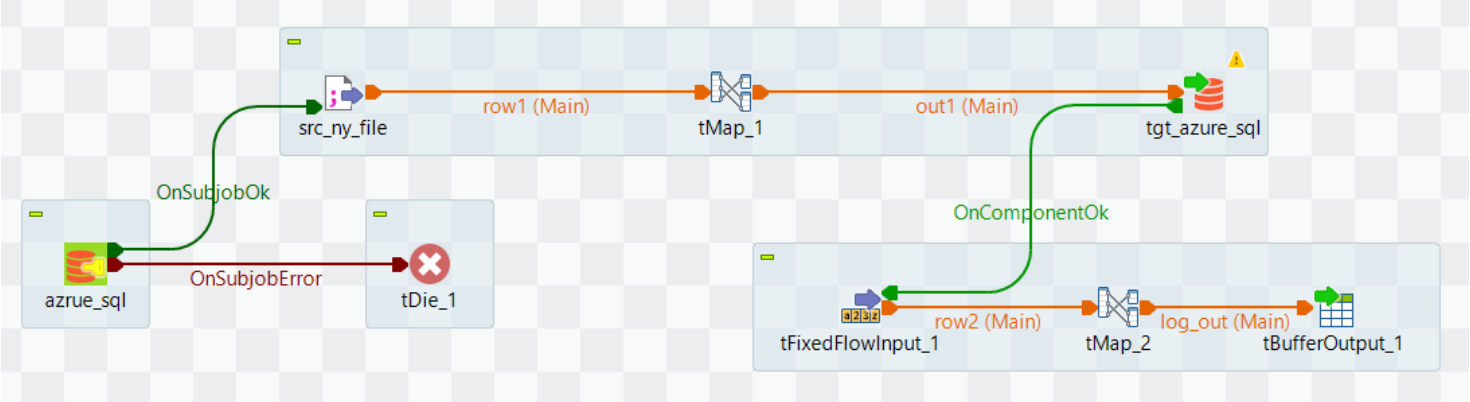
```
    PRINT '<<< FAILED CREATING TABLE time_dim >>>'
go

/*
 * TABLE: vehicle_involved_bdg
 */

CREATE TABLE vehicle_involved_bdg(
    vehicle_involved_sk     int            IDENTITY(1,1),
    crash_sk                int            NOT NULL,
    vehicle_sk              int            NOT NULL,
    DI_CreatedDate          date           NULL,
    DI_Workflow_FileName    varchar(50)    NULL,
    pid                     varchar(10)    NULL,
    CONSTRAINT PK10 PRIMARY KEY NONCLUSTERED (vehicle_involved_sk)
)

go


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

/*
 * TABLE: vehicle_type_dim
 */

CREATE TABLE vehicle_type_dim(
    vehicle_sk              int            IDENTITY(1,1),
    vehicle_type            varchar(50)    NOT NULL,
    DI_CreatedDate          date           NULL,
    DI_Workflow_FileName    varchar(50)    NULL,
    pid                     varchar(10)    NULL,
    CONSTRAINT PK9 PRIMARY KEY NONCLUSTERED (vehicle_sk)
```

```
)

go


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

/*
 * TABLE: crash_fct
 */

ALTER TABLE crash_fct ADD CONSTRAINT Refdate_dim2
    FOREIGN KEY (crash_date_sk)
    REFERENCES date_dim(date_sk)
go

ALTER TABLE crash_fct ADD CONSTRAINT Reftime_dim8
    FOREIGN KEY (time_sk)
    REFERENCES time_dim(time_sk)
go

ALTER TABLE crash_fct ADD CONSTRAINT Refaddress_dim10
    FOREIGN KEY (address_sk)
    REFERENCES address_dim(address_sk)
go

ALTER TABLE crash_fct ADD CONSTRAINT Refsource_dim19
    FOREIGN KEY (source_id)
    REFERENCES source_dim(source_id)
go


/*
```

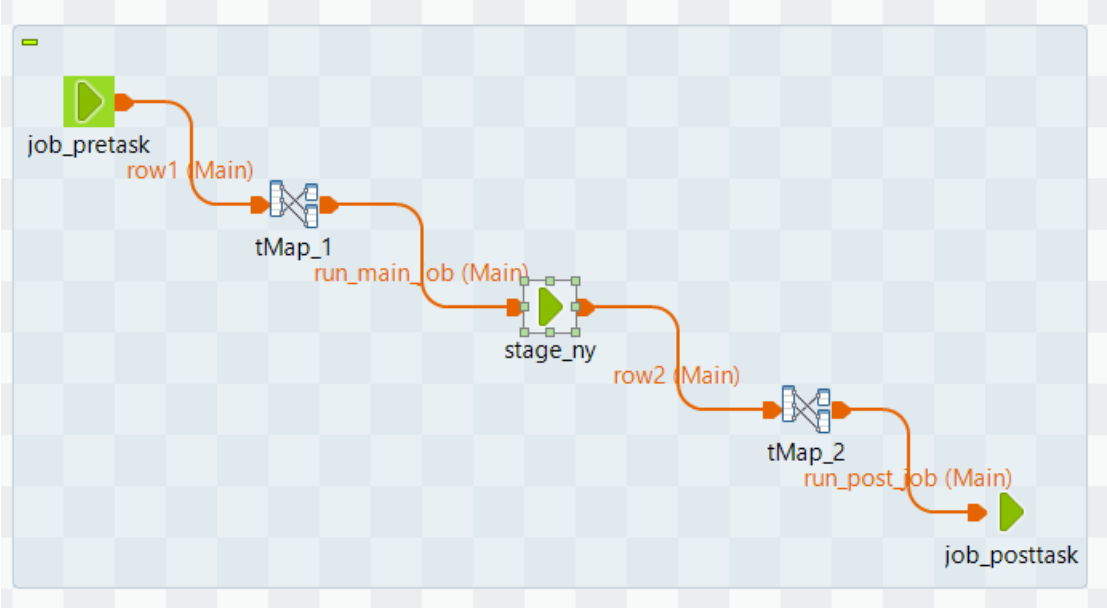
```
 * TABLE: most_common_factor_bdg
 */

ALTER TABLE most_common_factor_bdg ADD CONSTRAINT Refcrash_fct17
    FOREIGN KEY (crash_sk)
    REFERENCES crash_fct(crash_sk)
go

ALTER TABLE most_common_factor_bdg ADD CONSTRAINT Refcontrib_factor_dim18
    FOREIGN KEY (contrib_factor_sk)
    REFERENCES contrib_factor_dim(contrib_factor_sk)
go



/*
 * TABLE: vehicle_involved_bdg
 */

ALTER TABLE vehicle_involved_bdg ADD CONSTRAINT Refcrash_fct20
    FOREIGN KEY (crash_sk)
    REFERENCES crash_fct(crash_sk)
go

ALTER TABLE vehicle_involved_bdg ADD CONSTRAINT Refvehicle_type_dim21
    FOREIGN KEY (vehicle_sk)
    REFERENCES vehicle_type_dim(vehicle_sk)
go
```

5. Data Loading into Integration Schema

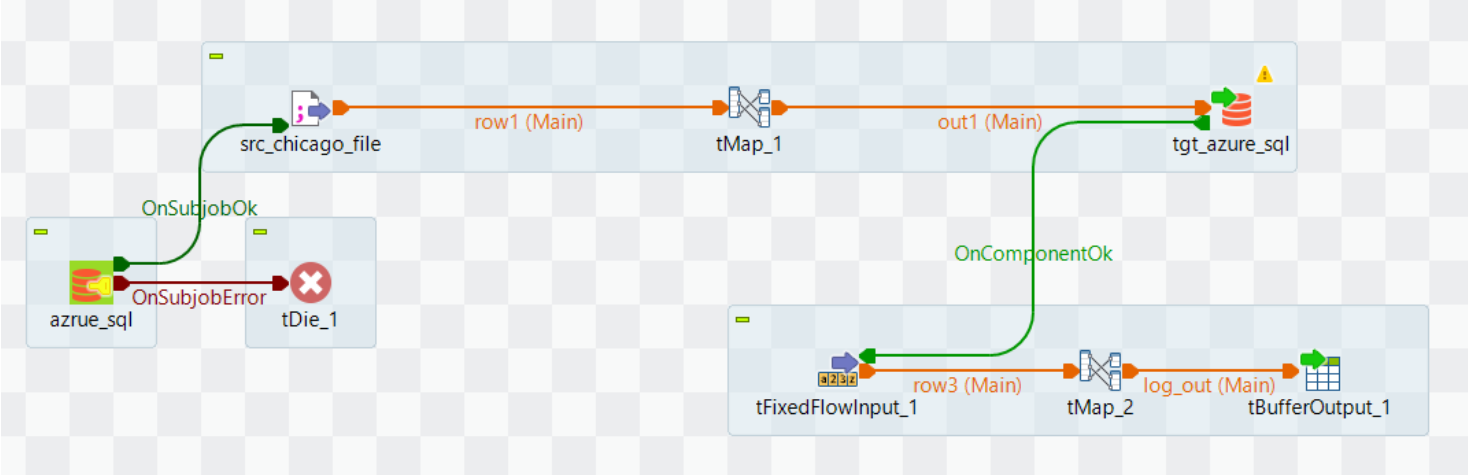
- **stage_ny**



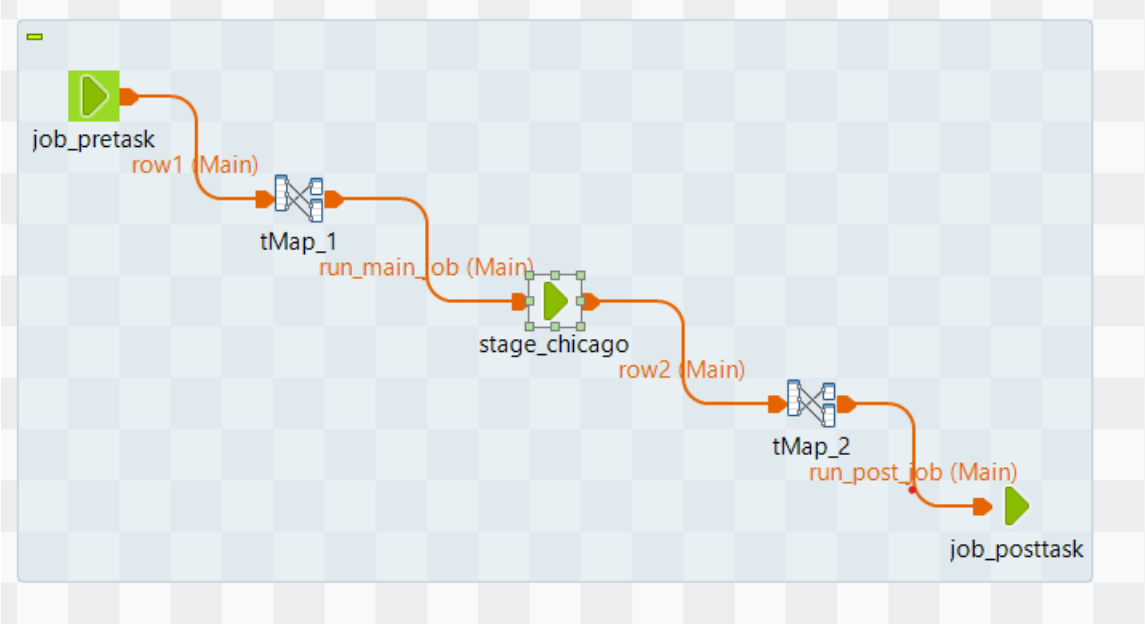
- **audit_stage_ny**



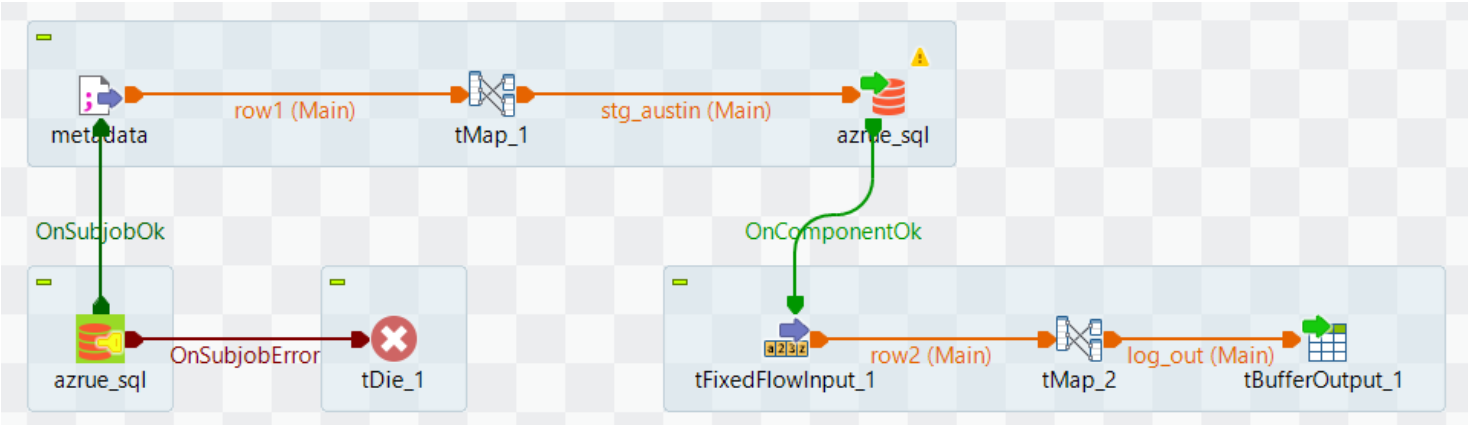
- **stage_chicago**



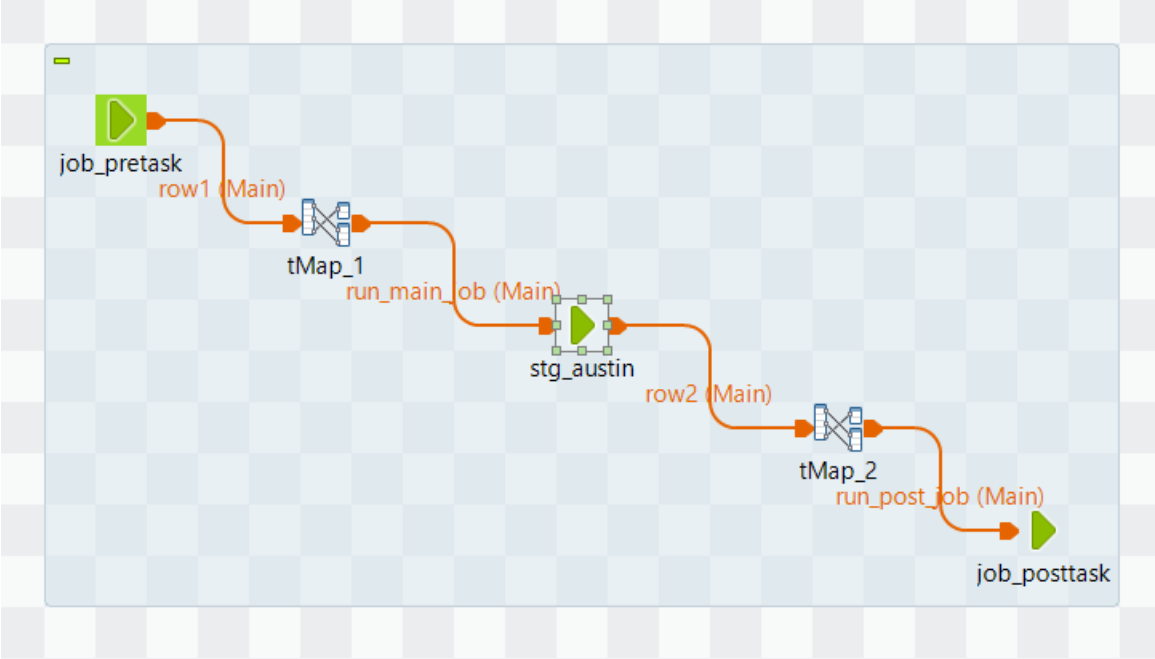
- **audit_stage_chicago**



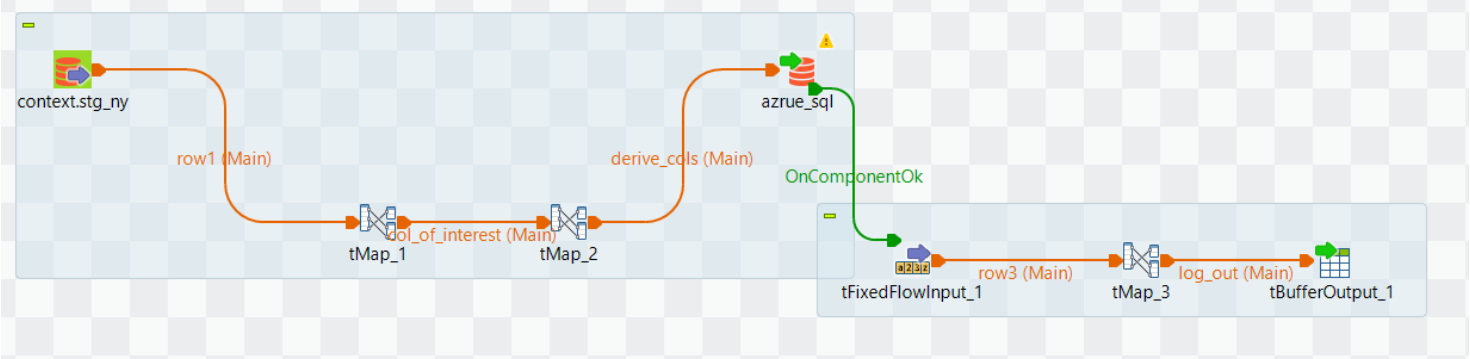
- **stg_austin**



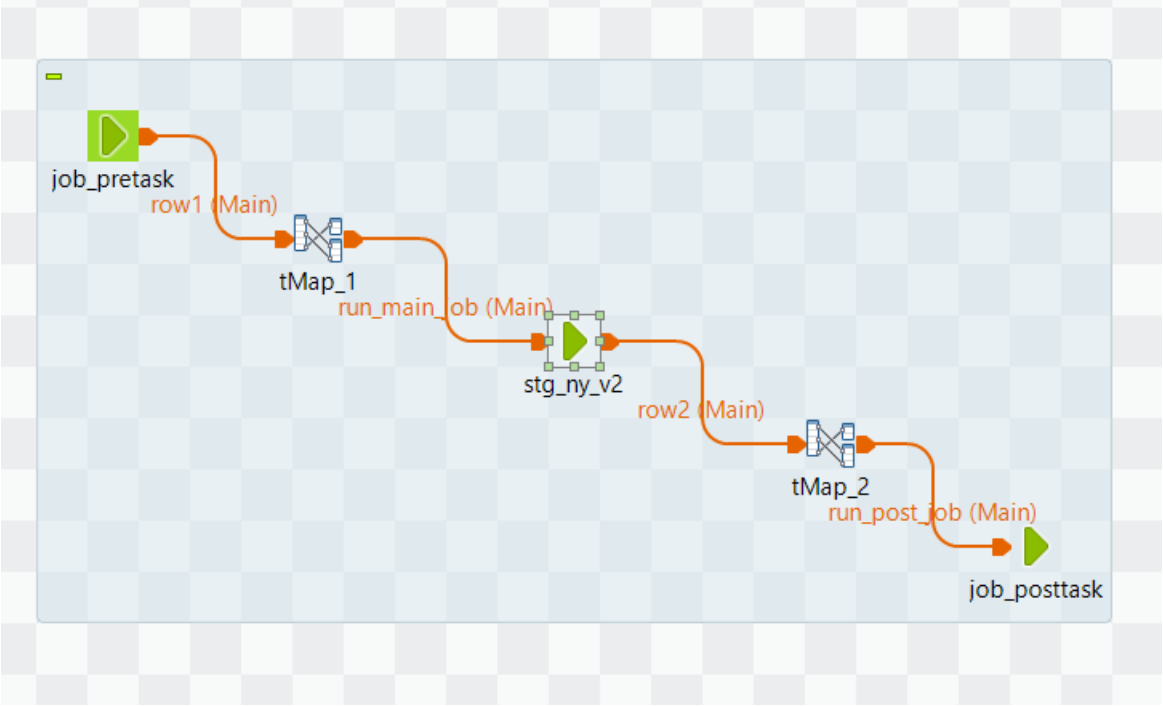
- **audit_stg_austin**



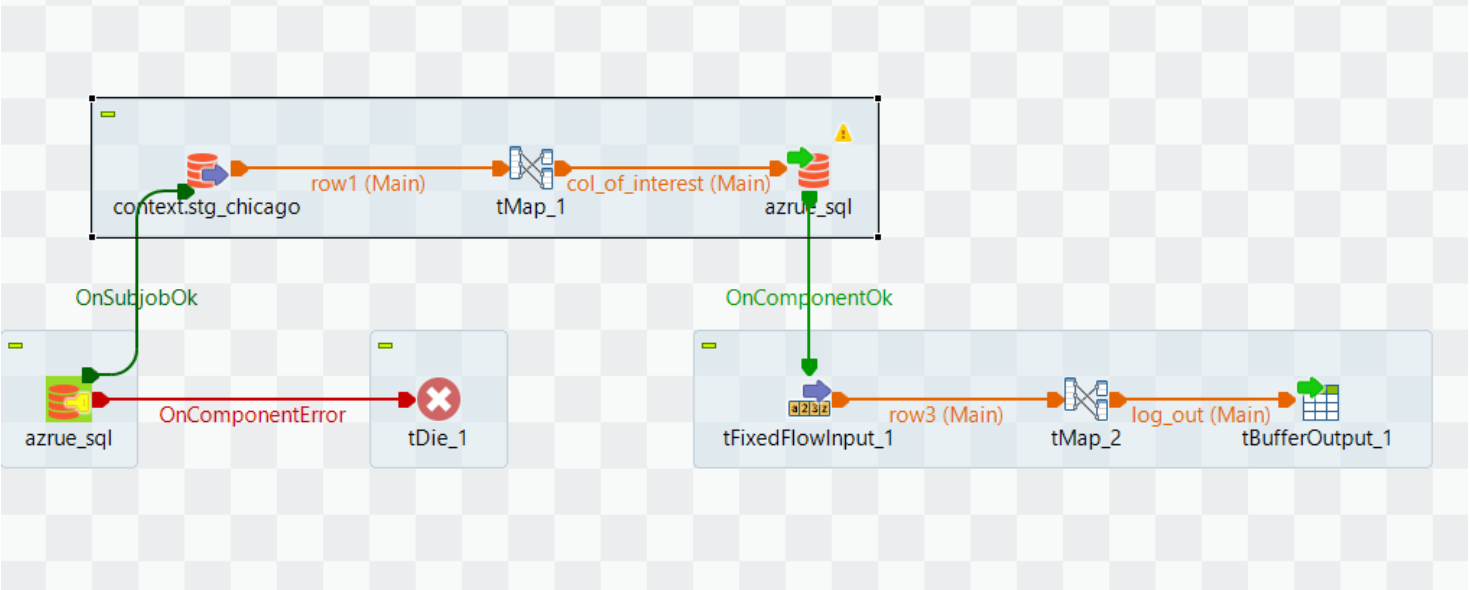
- **stg_ny_v2**



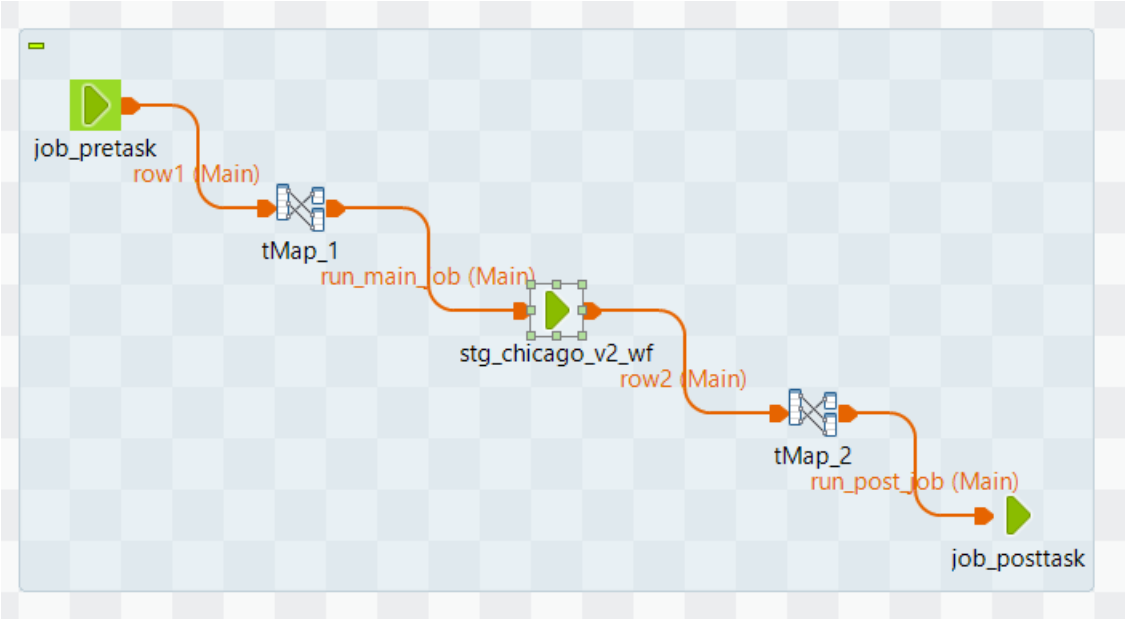
- **audit_stg_ny_v2**



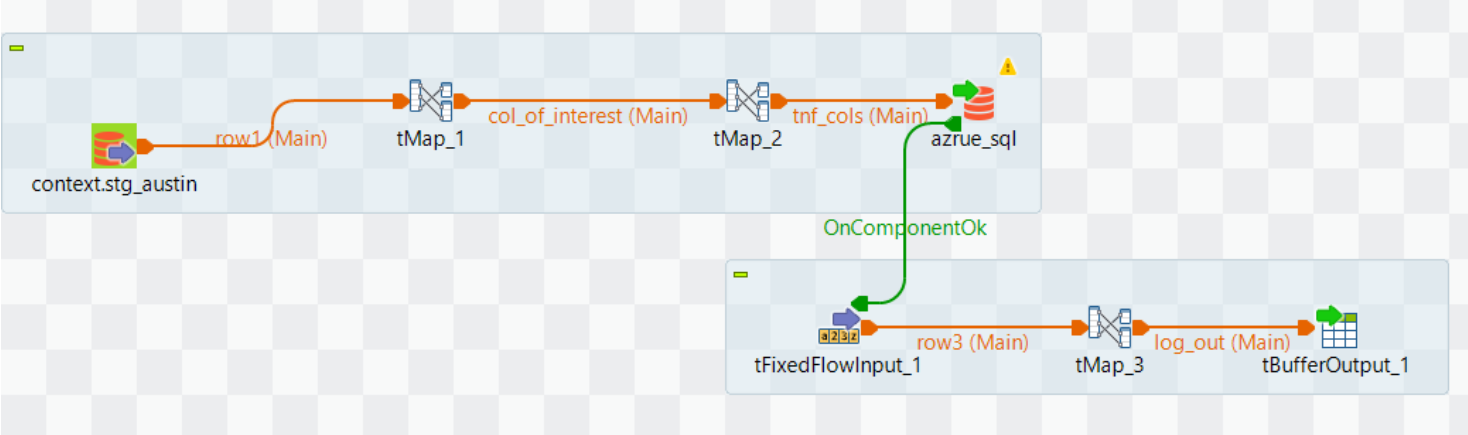
- **stg_chicago_v2**



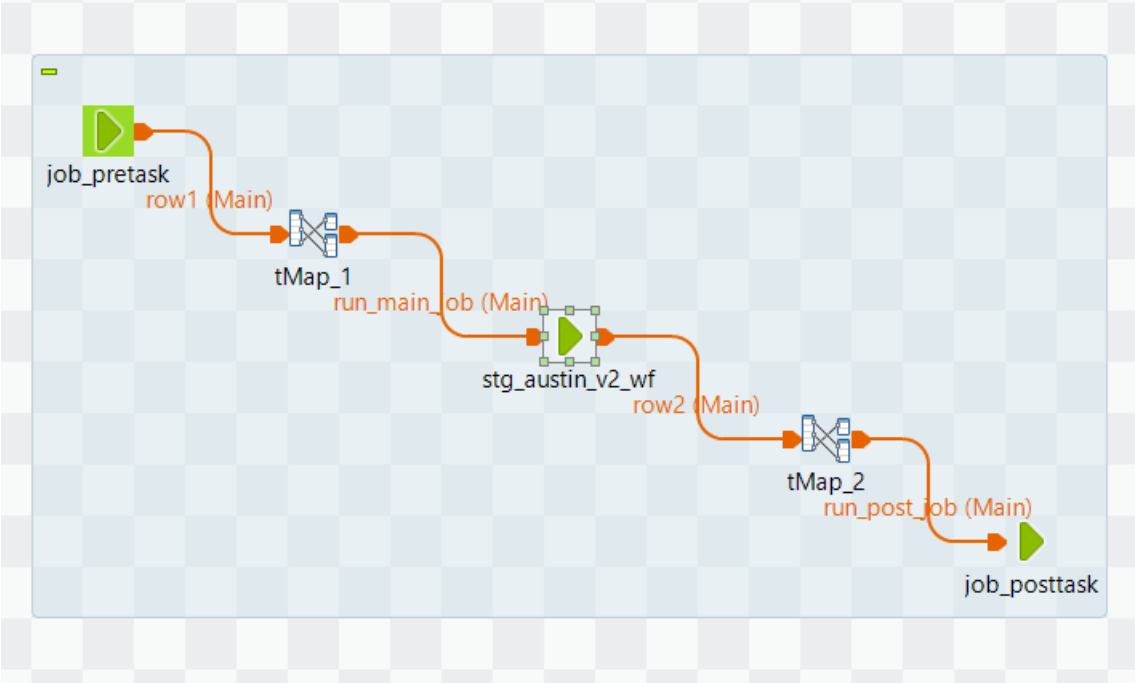
- **audit_stg_chicago_v2**



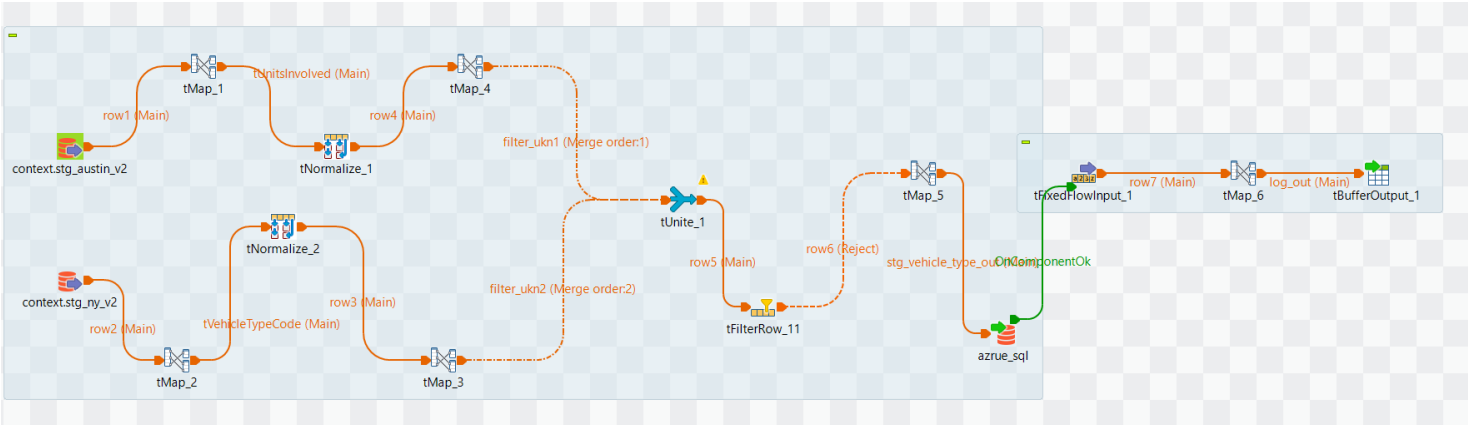
- **stg_austin_v2**



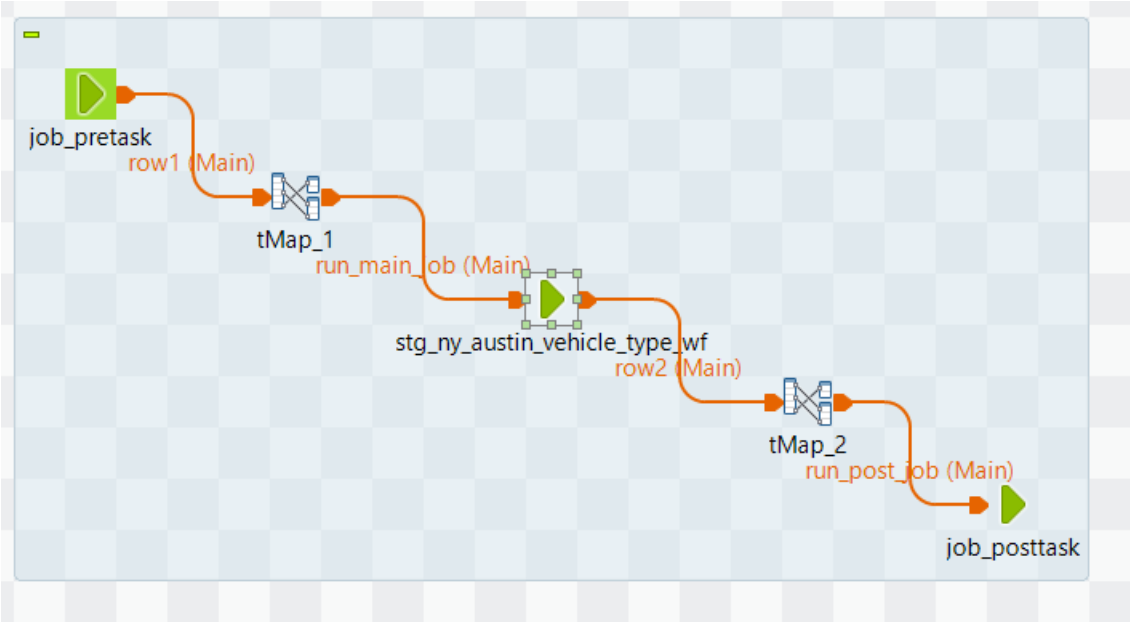
- **audit_stg_austin_v2**



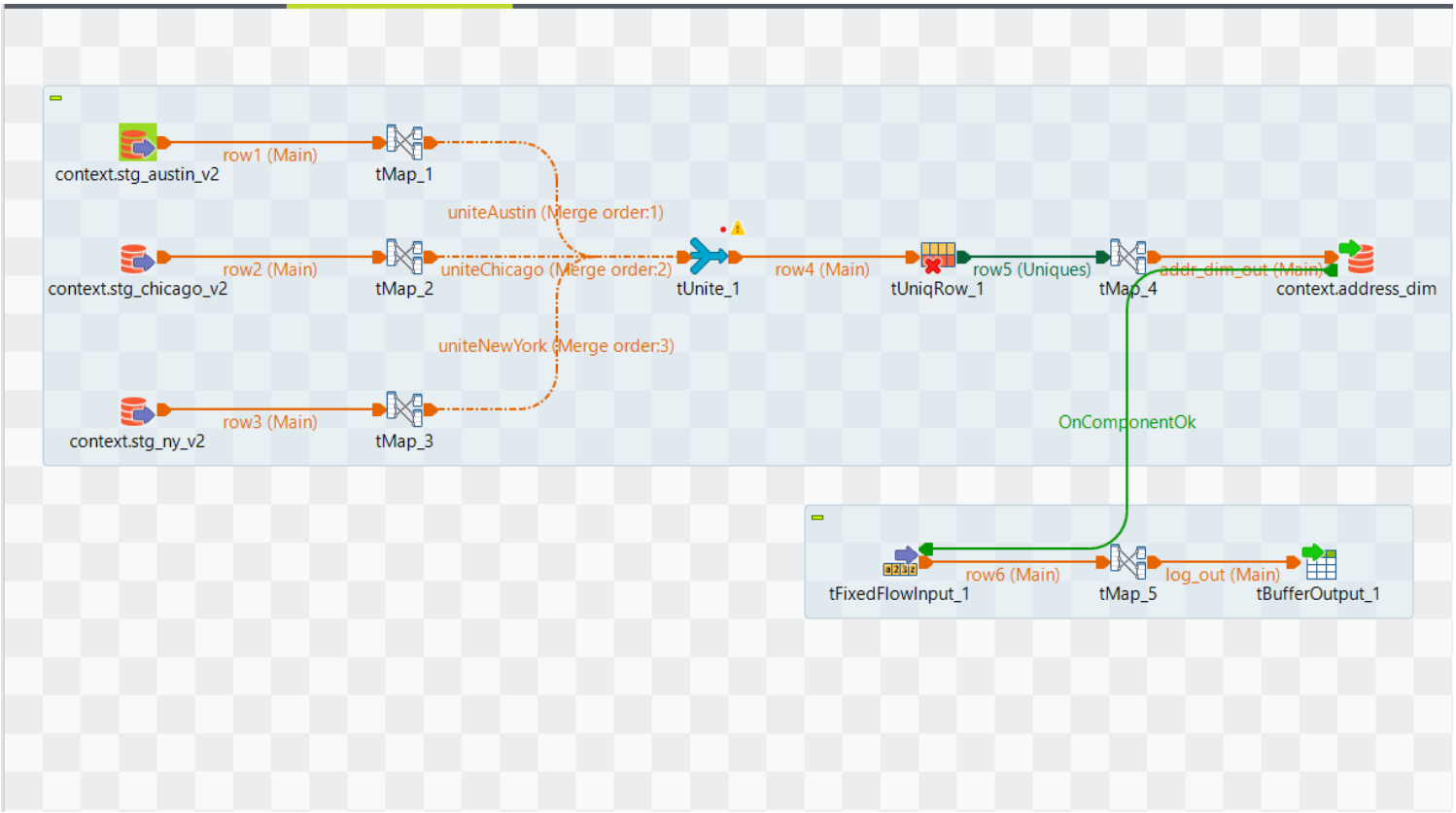
- **stg_ny_austin_vehicle_type_wf**



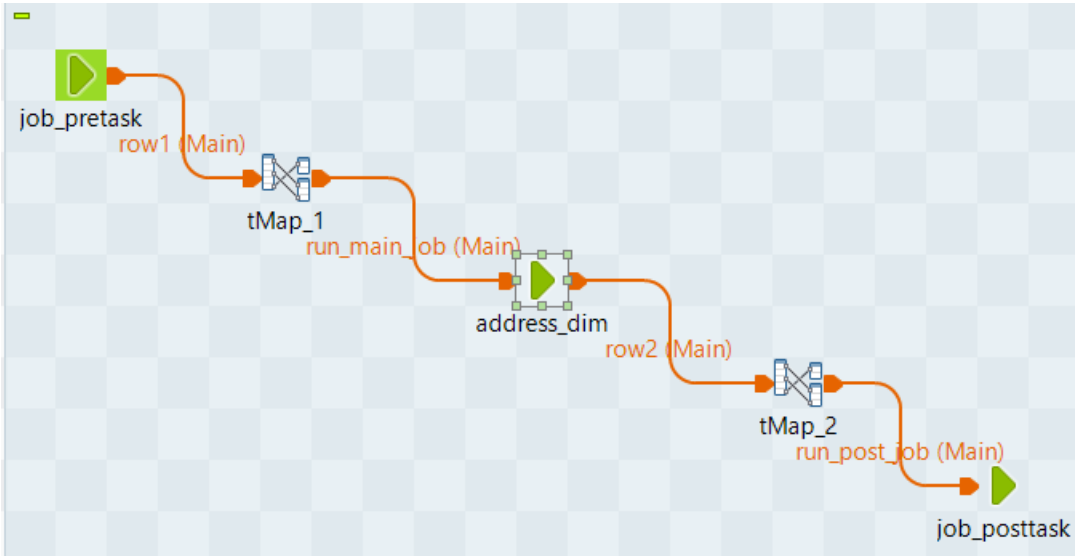
- **audit_stg_ny_austin_vehicle_type_wf**



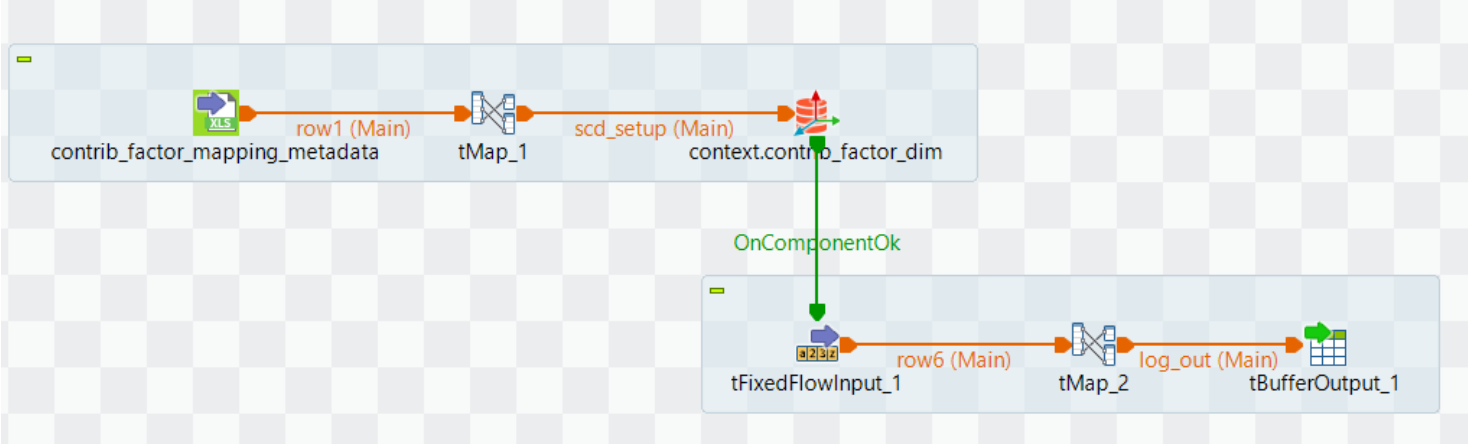
- **address_dim**



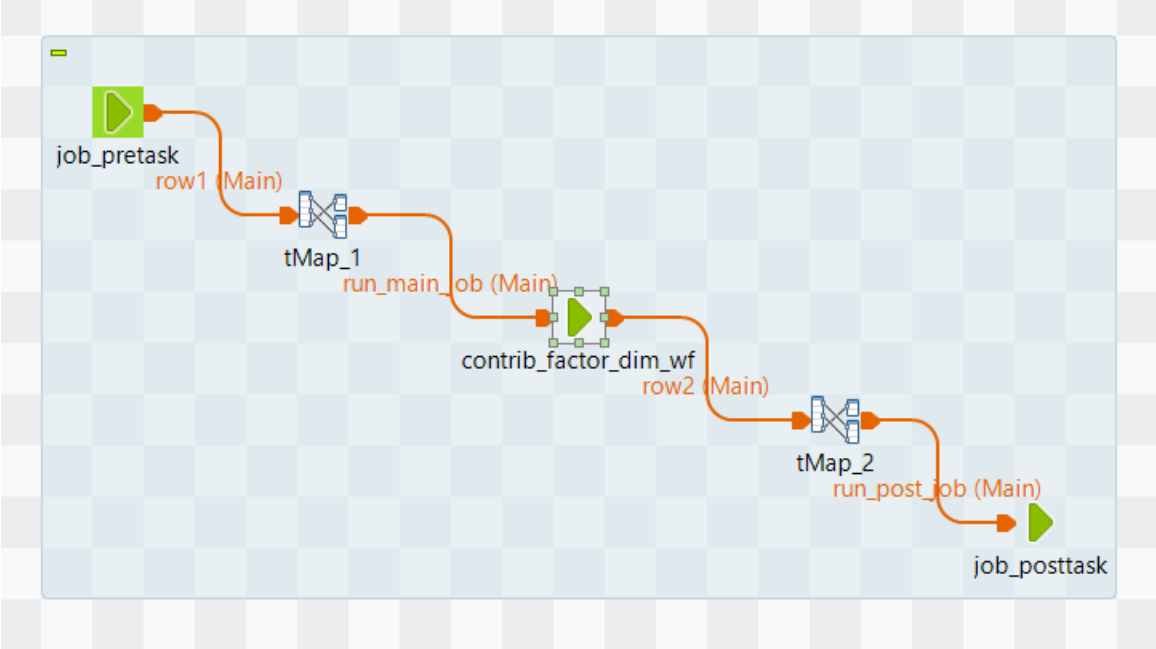
- **audit_address_dim**



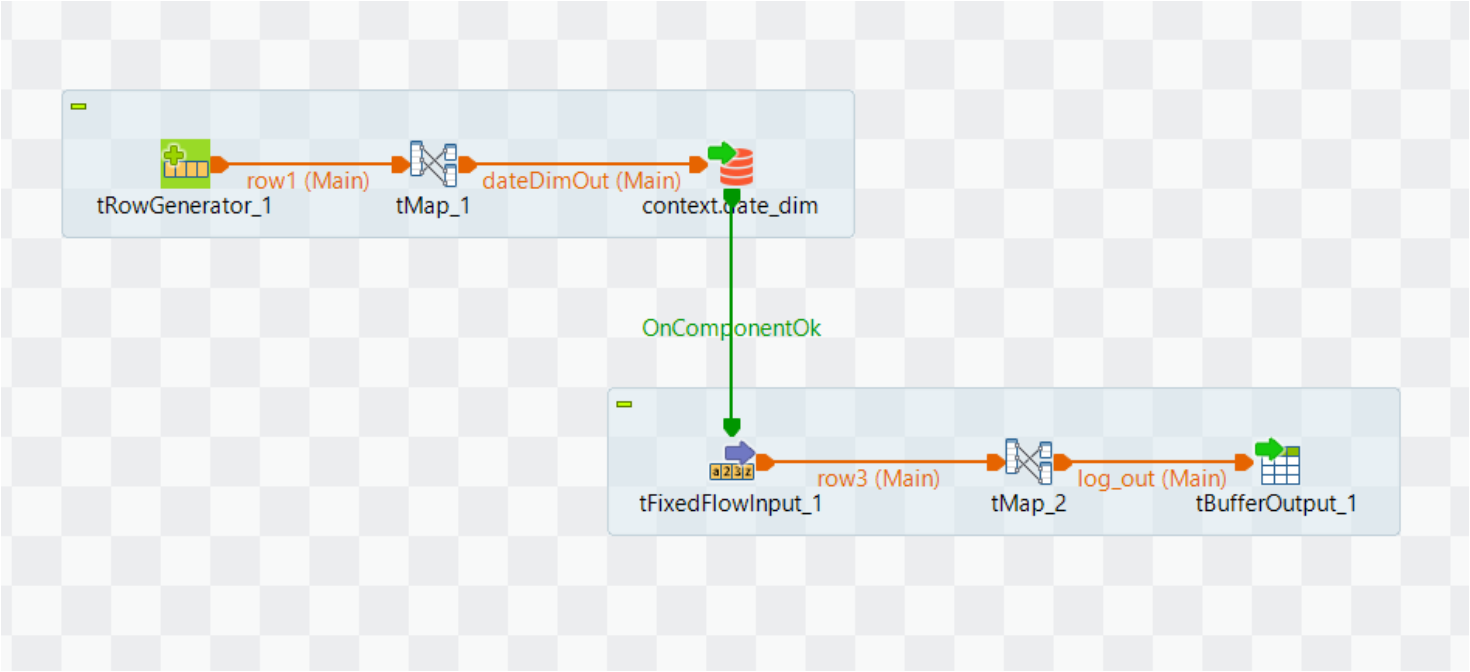
- **contrib_factor_dim_wf**



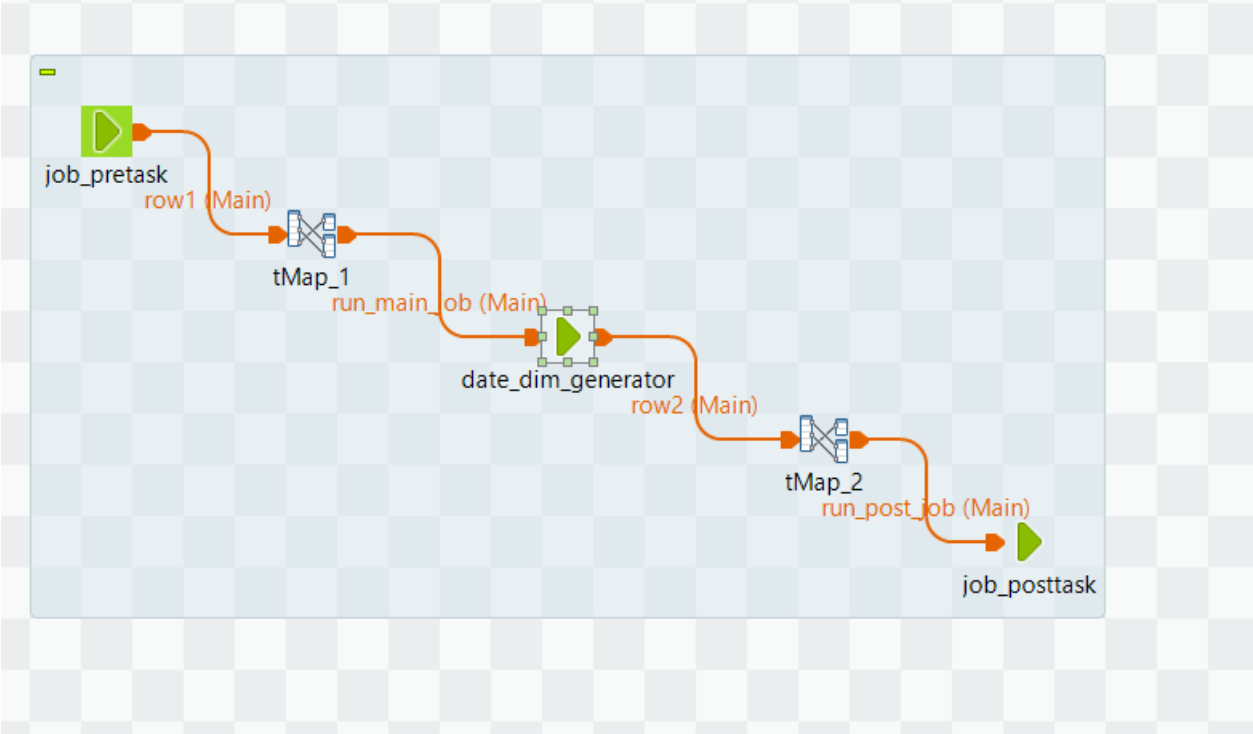
- **audit_contrib_factor_dim _wf**



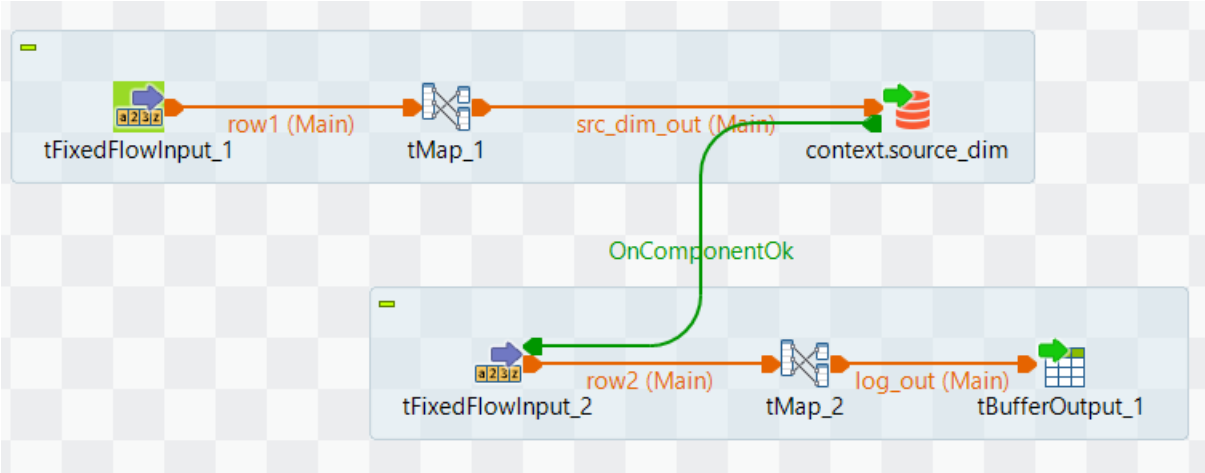
- **date_dim**



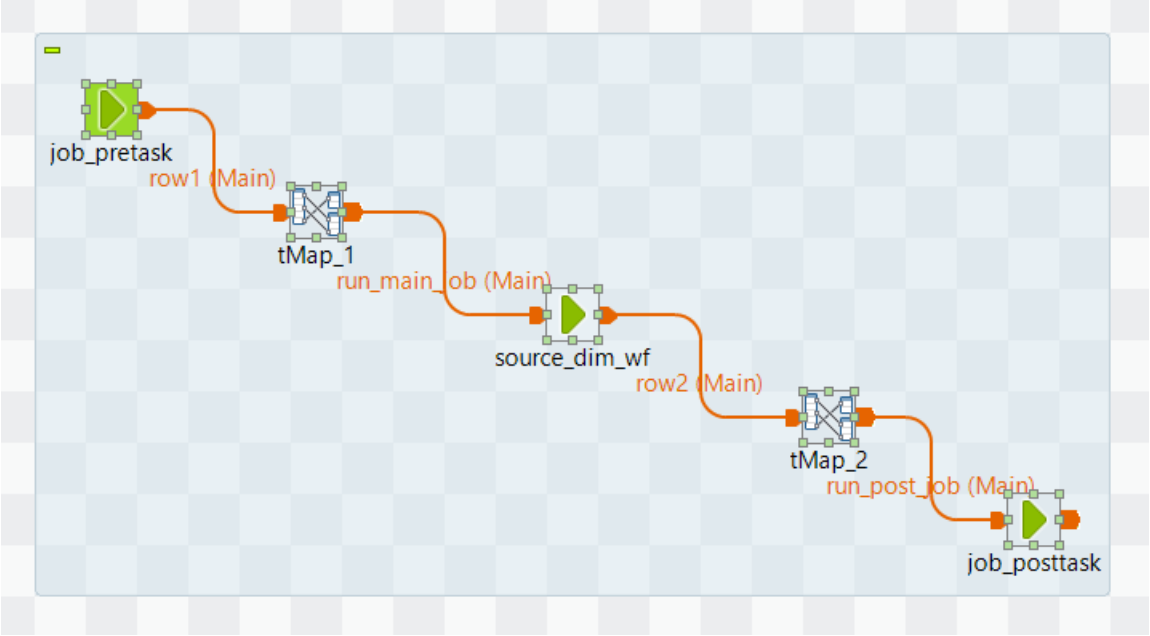
- **audit_date_dim**



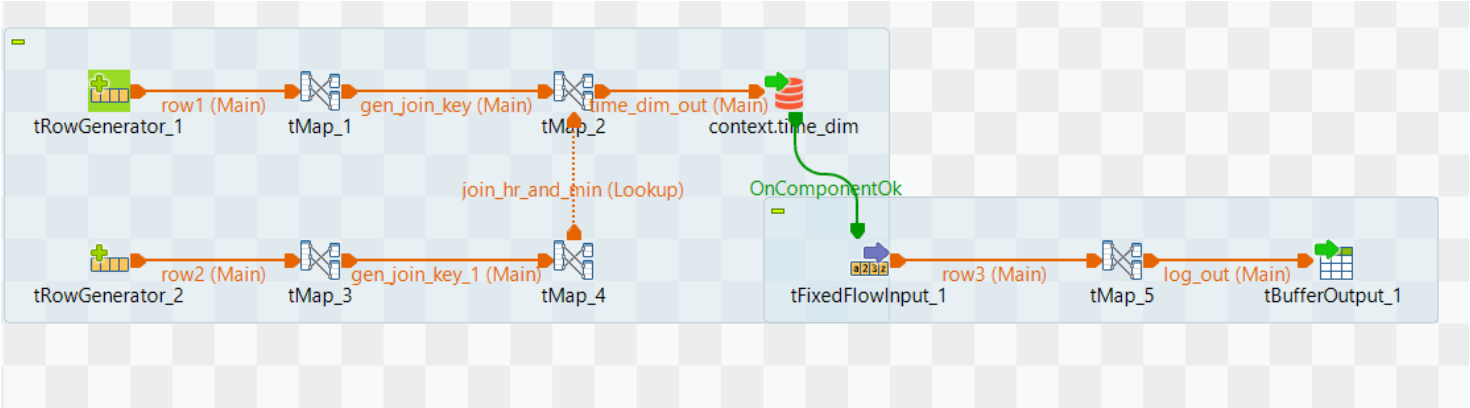
- **source_dim**



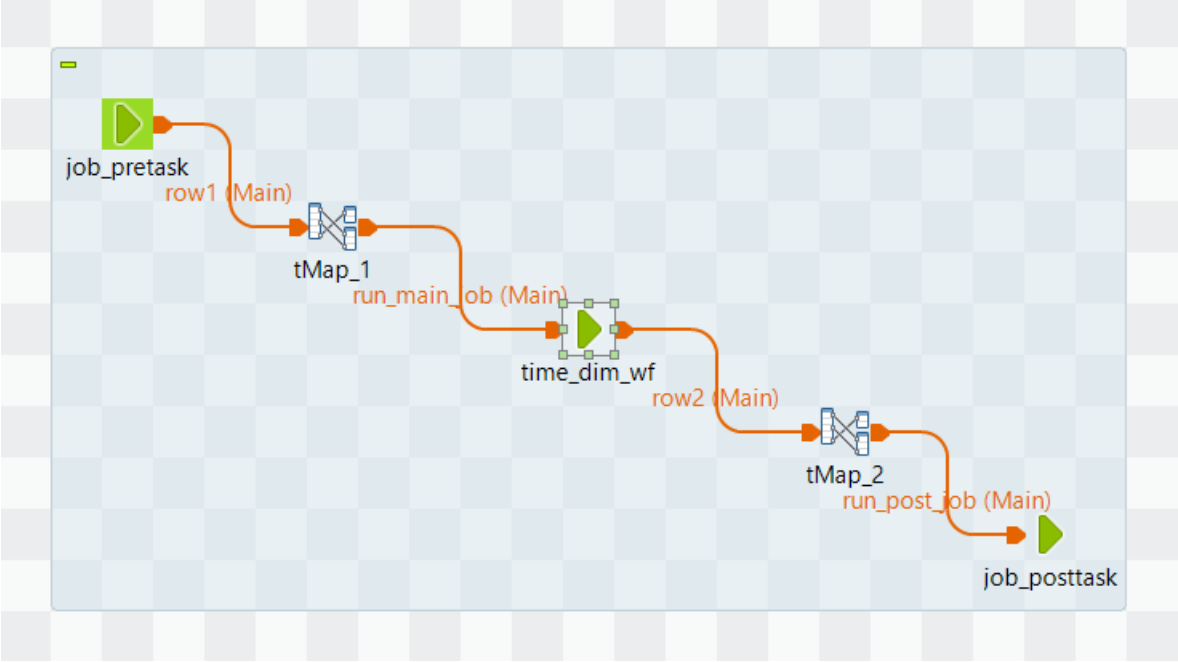
- **audit_source_dim**



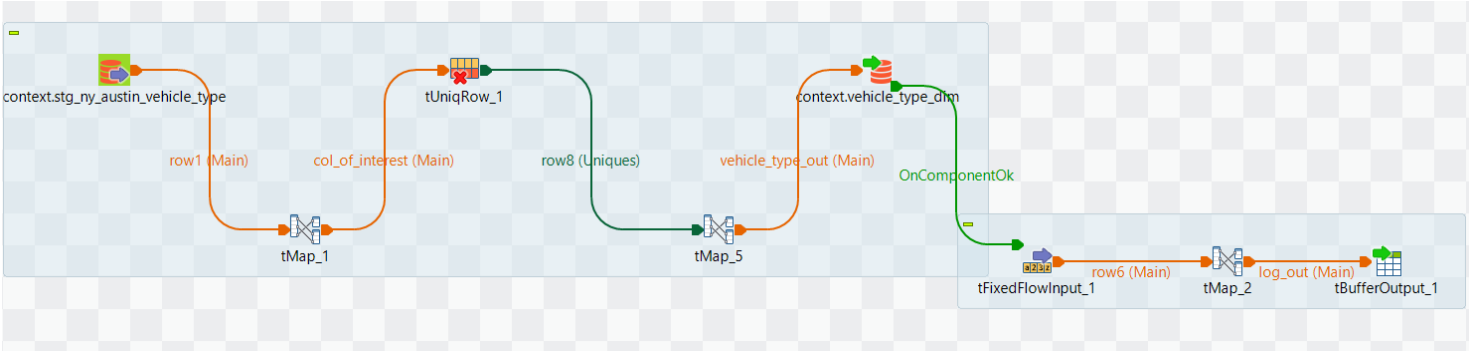
- **time_dim_wf**



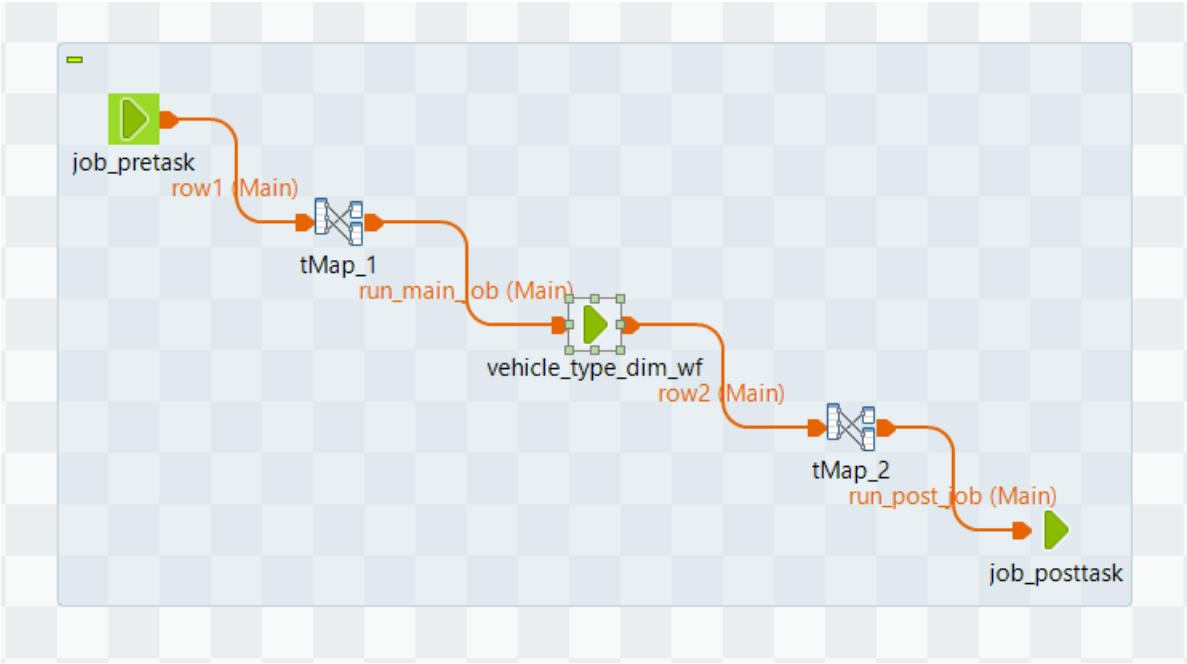
- **audit_time_dim_wf**



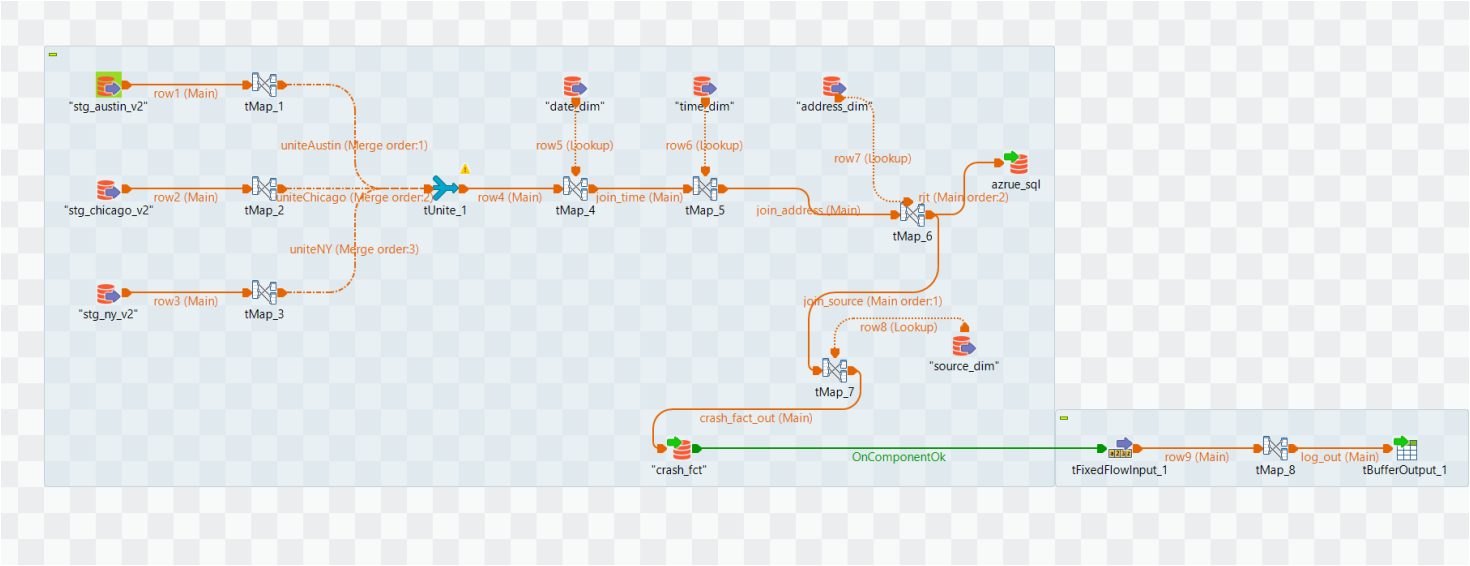
- **vehicle_type_dim_wf**



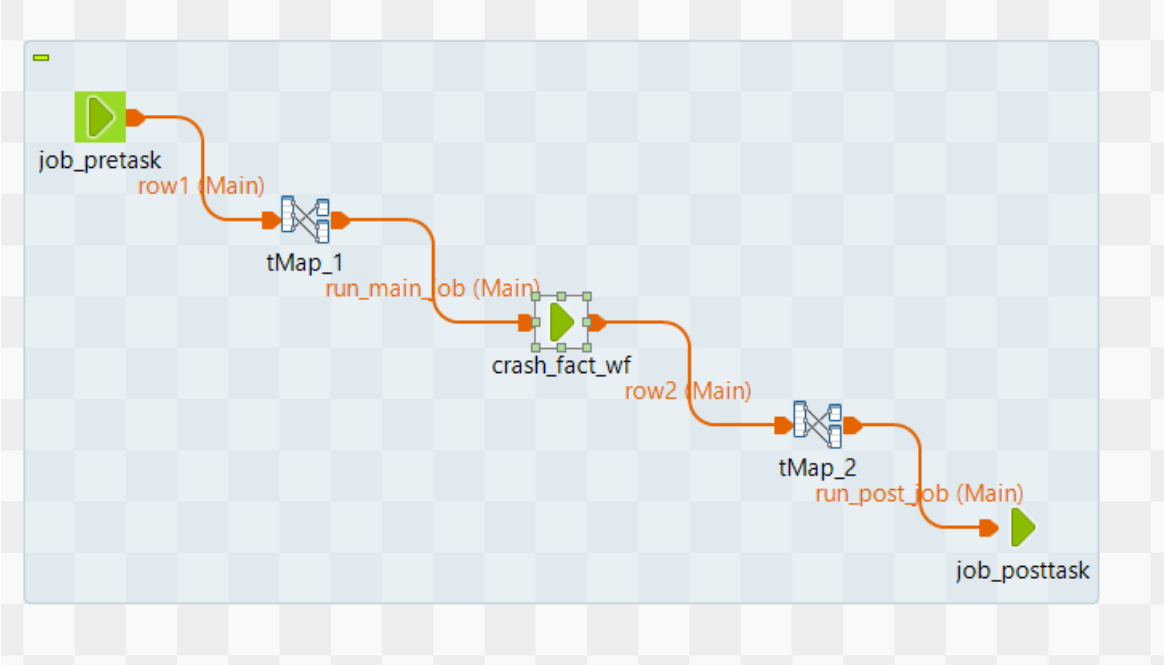
- **audit_vehicle_type_dim_wf**



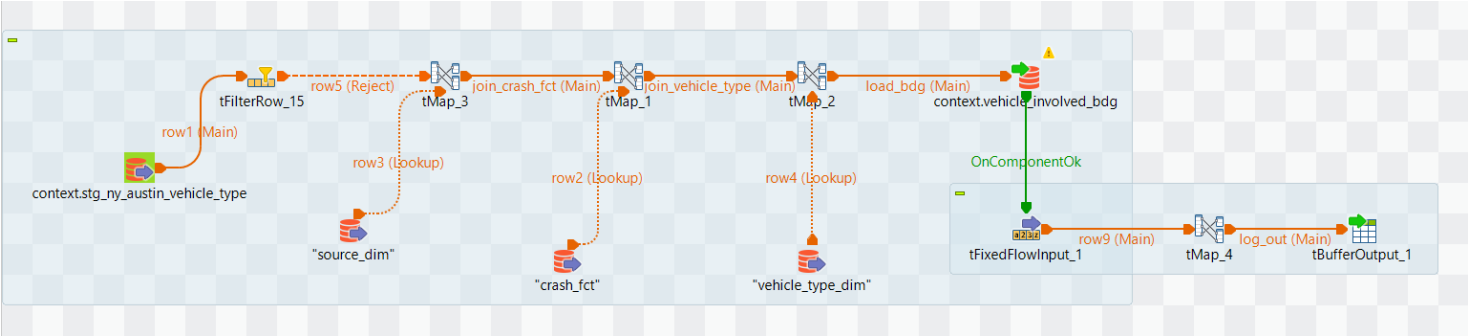
- **crash_fact_wf**



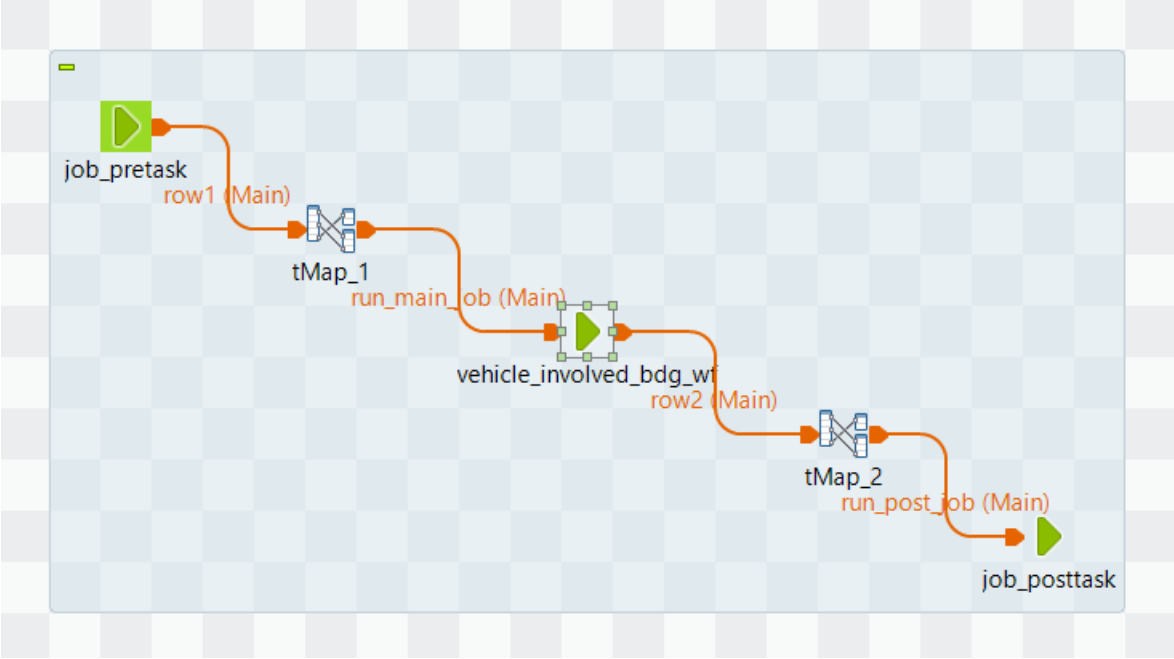
- **audit_crash_fact_wf**



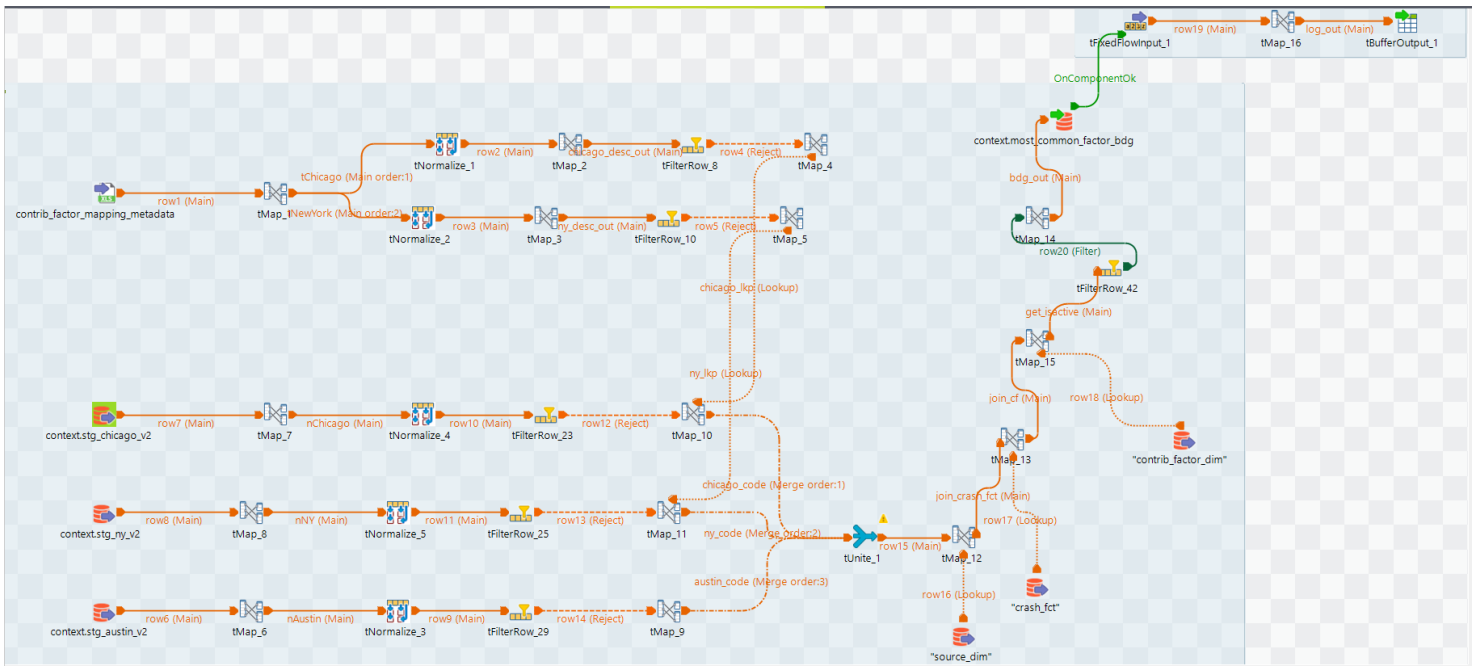
- **vehicle_involved_bdg_wf**



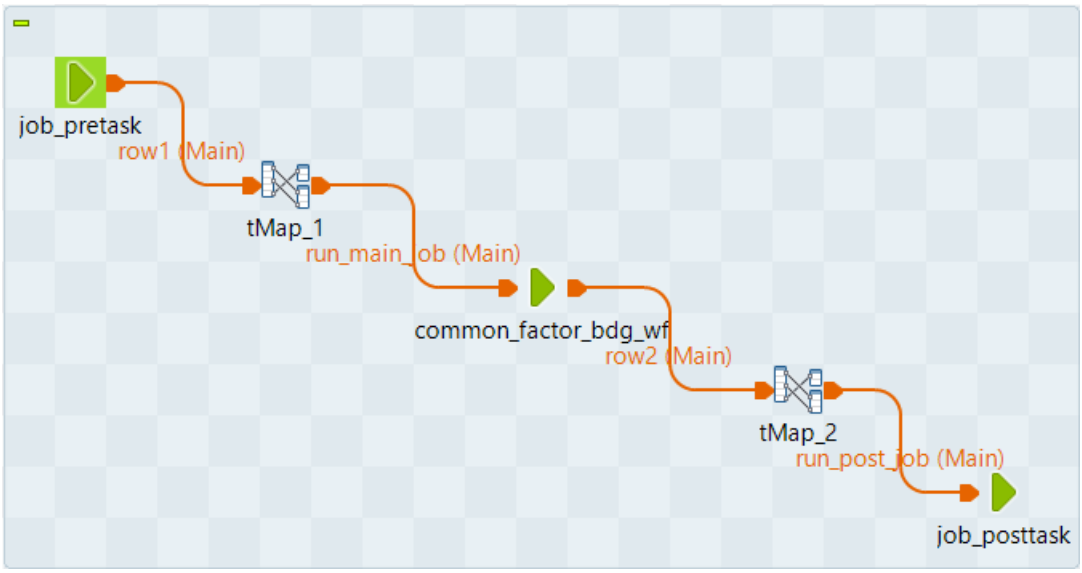
- **audit_vehicle_involved_bdg_wf**



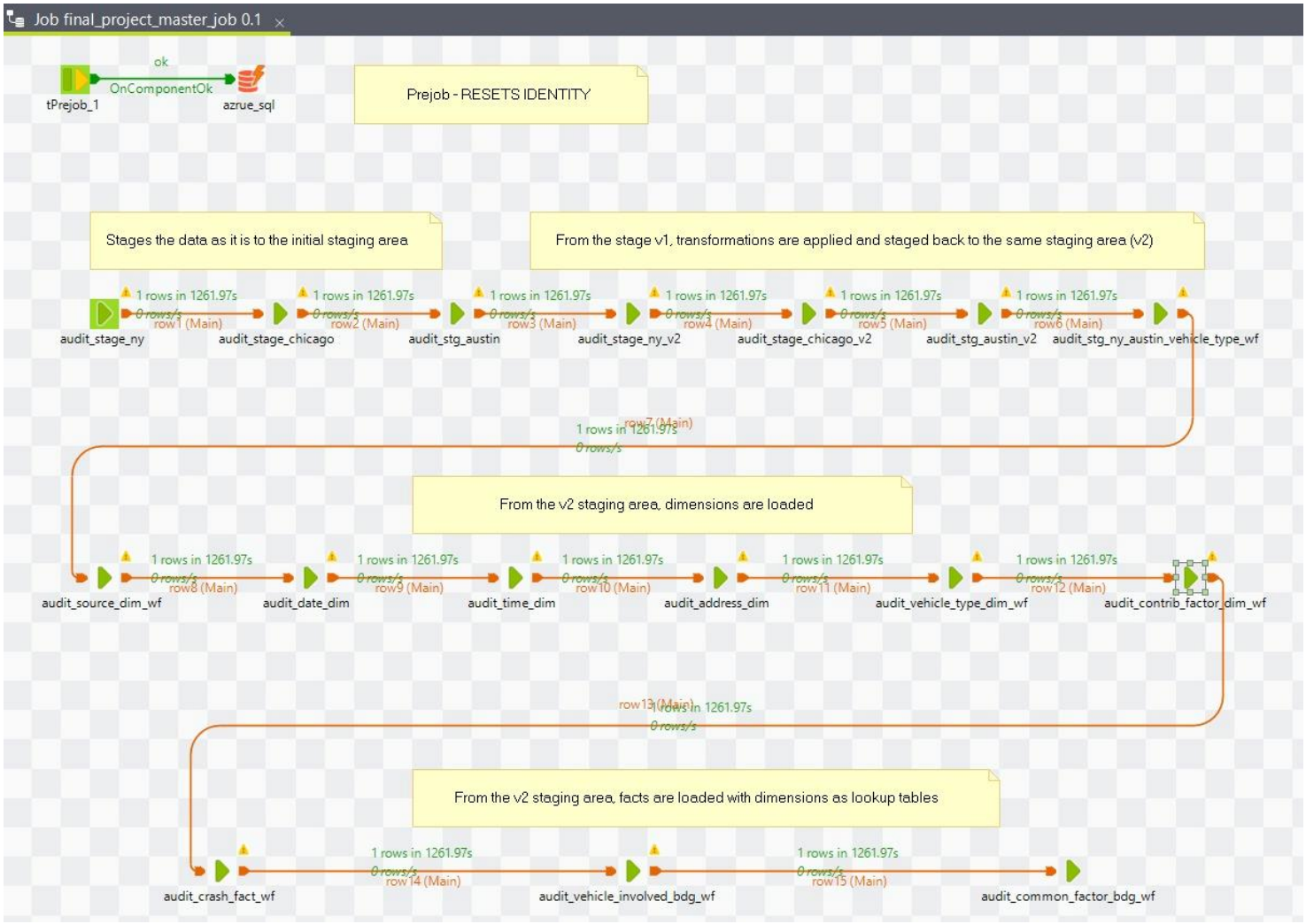
- **common_factor_bdg_wf**



- **audit_common_factor_bdg_wf**



- **final_project_master_job**



Row count validation

```sql
-- row count verification
select  isnull(a.tbl,'Total Rows') as tbl, sum(a.rowcnt) Row_Count from (
select 'stg_autin' as tbl ,count(crash_id) rowcnt from stg_austin
union all
select 'stg_chicago' , count(CRASH_RECORD_ID) from stg_chicago
union all
select 'stg_ny', count(COLLISION_ID) from stg_ny
union all
select 'stg_austin_v2', count(crash_id) from stg_austin_v2
union all
select 'stg_chicago_v2',count(CRASH_RECORD_ID) from stg_chicago_v2
union all
select 'stg_ny_v2', count(COLLISION_ID) from stg_ny_v2
```

3 %

Results | Messages

tbl	Row_Count
address_dim	814475
contrib_factor_dim	85
crash_fct	3040900
date_dim	7500
most_common_factor_bdg	5681670
soruce_dim	3
stg_austin_v2	147750
stg_autin	147750
stg_chicago	817723
stg_chicago_v2	817723
stg_ny	2075427
stg_ny_austin_vehicle_type	4229558
stg_ny_v2	2075427
time_dim	1440
vehicle_involved_bdg	4229558
vehicle_type_dim	1848
Total Rows	24088837

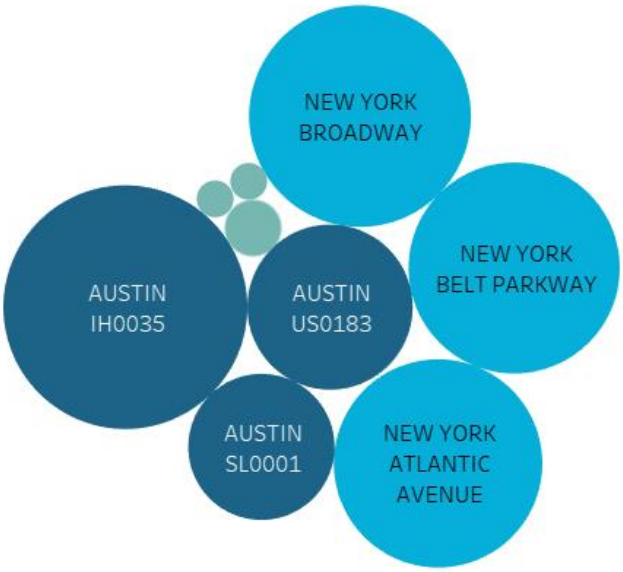
6. Data Visualization

- **Tableau**

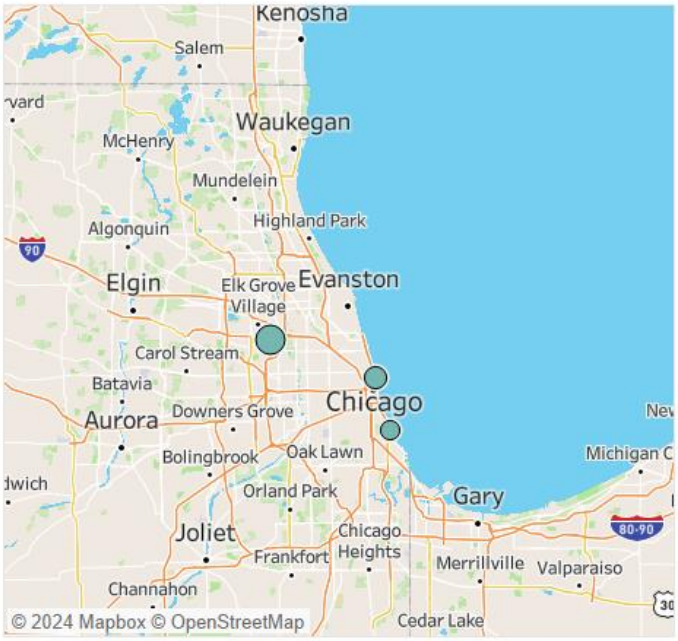
Geo-based Analysis

AUSTIN	CHICAGO	NEW YORK	OVERALL
1,47,750	8,17,723	20,75,427	30,40,900

Top 3 Areas with Greatest Number of Accidents by Street Name



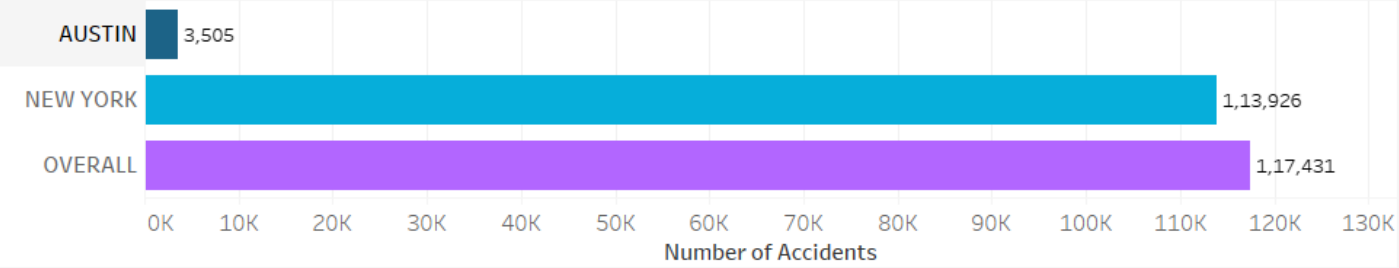
Top 3 Areas with Greatest Number of Accidents by Latitude & Longitude



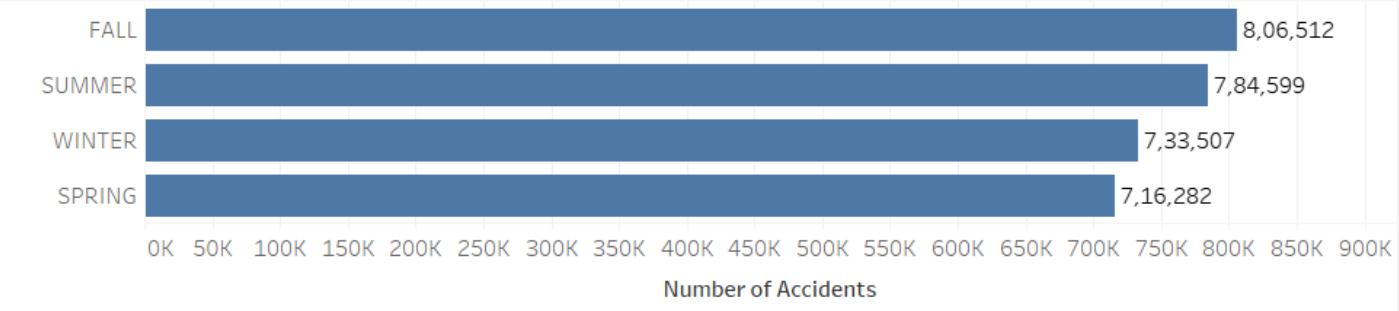
Number of Accidents Resulting in Just Injuries

AUSTIN	CHICAGO	NEW YORK	OVERALL
65,031	1,12,512	4,74,390	6,51,933

Number of Accidents Pedestrians Involved by City



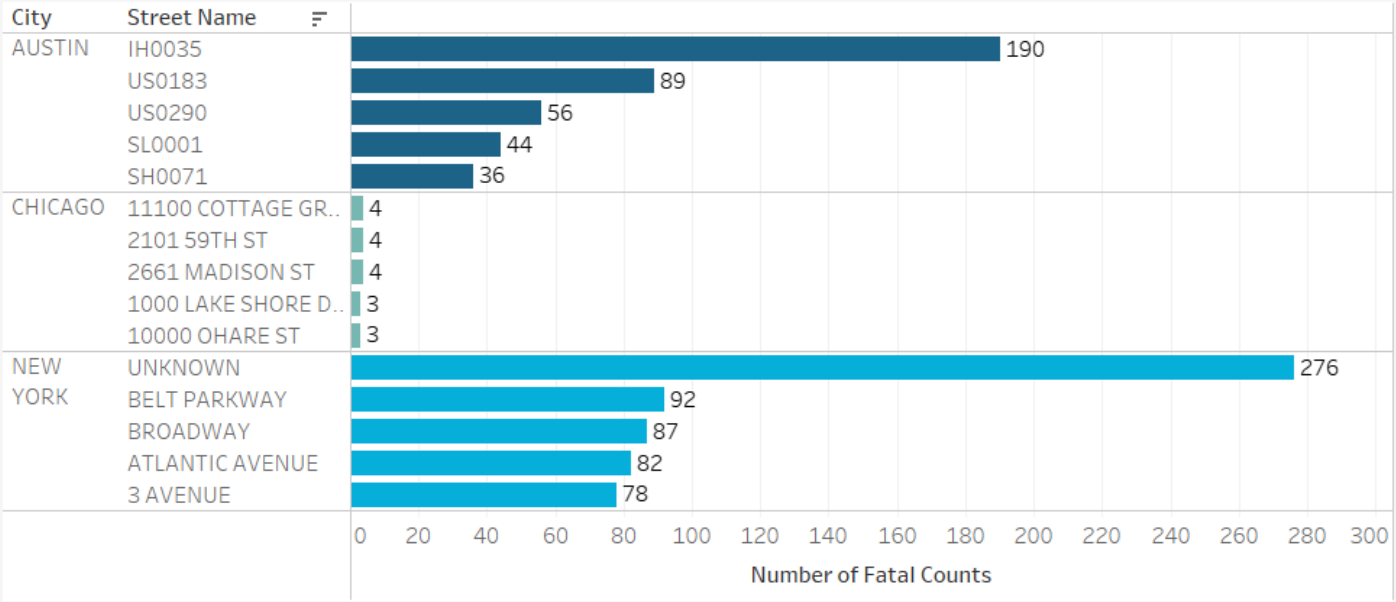
Number of Accidents by Season



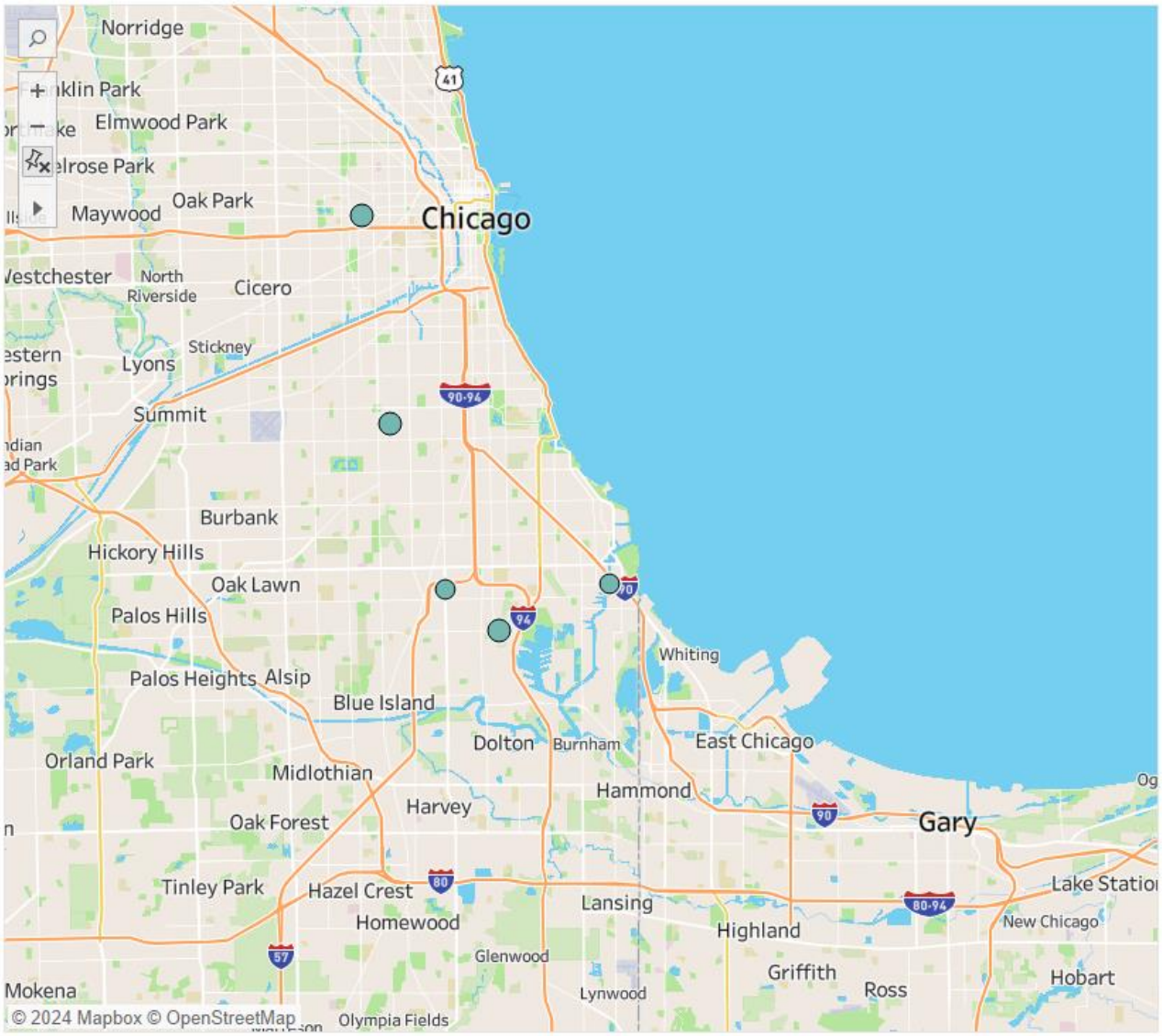
Number of Motorists Injured or Killed in Accidents

AUSTIN	NEW YORK	OVERALL
4,590	4,63,727	4,68,317

Top 5 areas with Most Fatal Count by Street Name

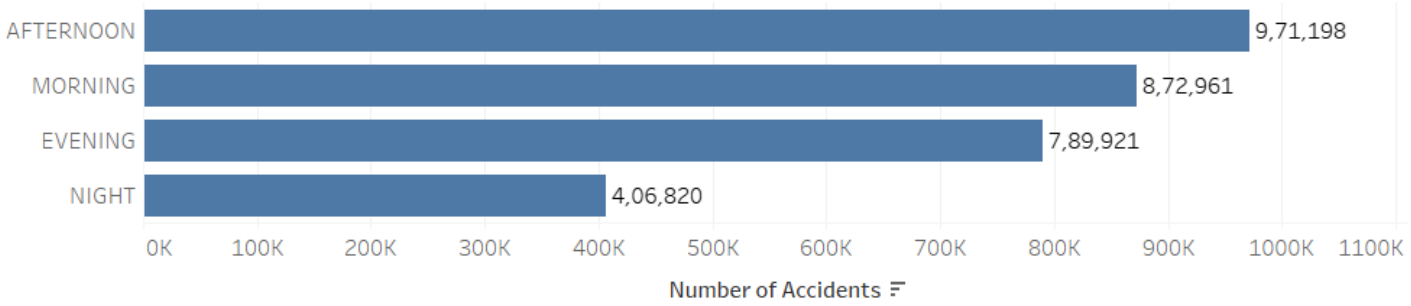


Top 5 Areas with Most Fatal Count by Latitude & Longitude

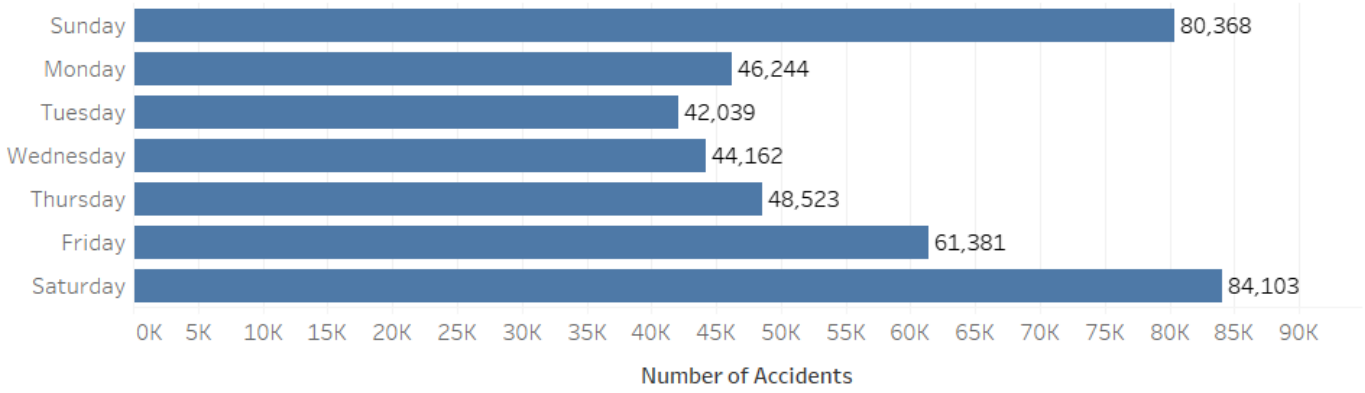


Date and Time-based Analysis

Number of Accidents Based on Time of Day



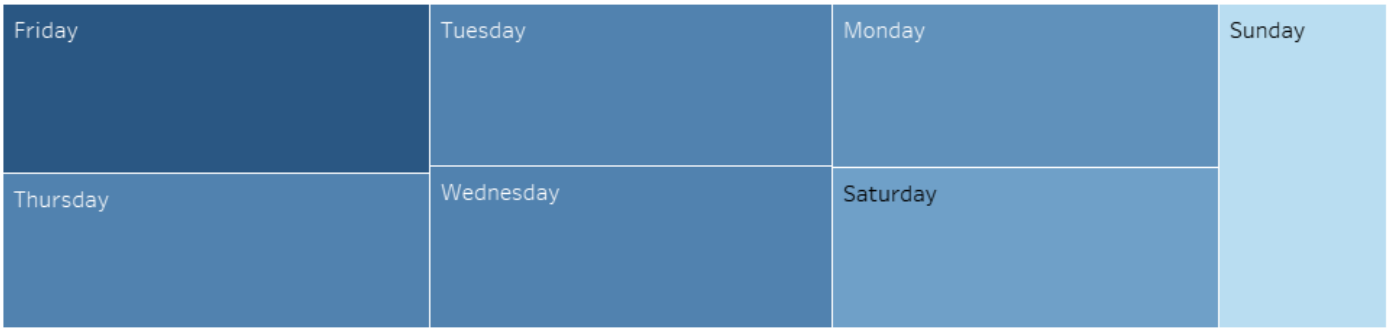
Number of Accidents by Day of the Week & Night



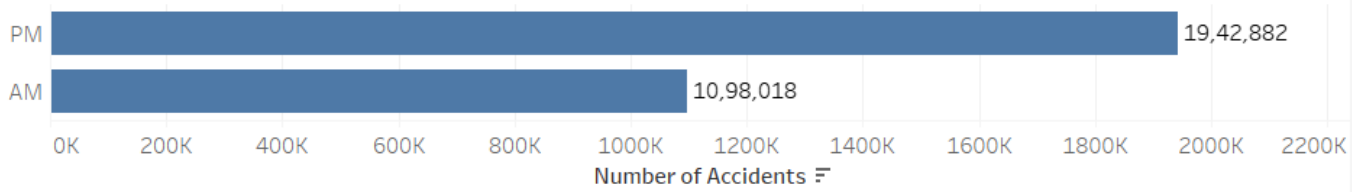
Date and Time-based Analysis

WEEKDAY	WEEKEND	WEEKEND NIGHT	WEEKDAY NIGHT
22,50,223	7,90,677	1,64,471	2,42,349

Number of Accidents Based on Day of the Week



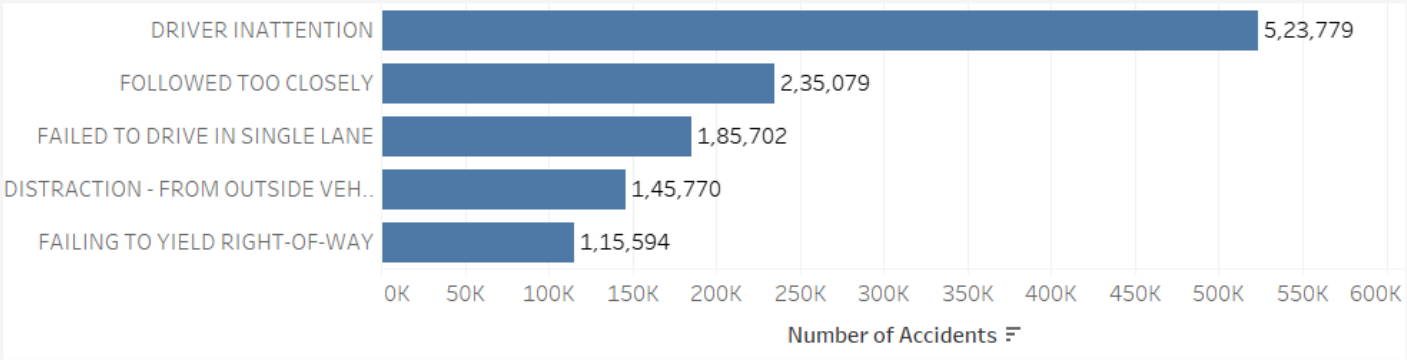
Number of Accidents Based on AM/PM



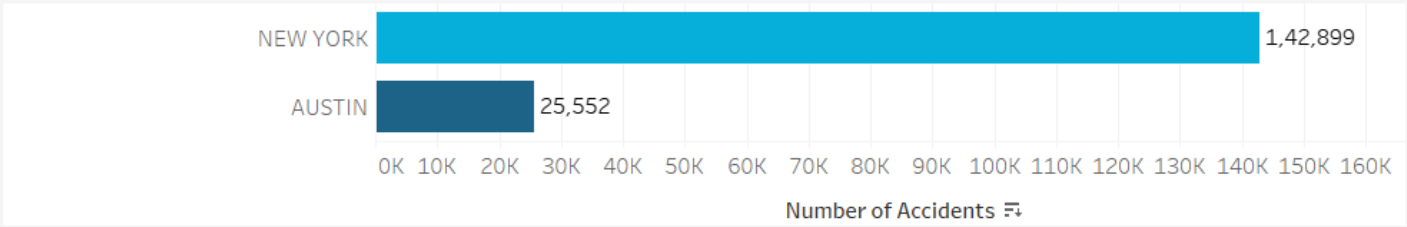
Fatality Analysis, Factors and Vehicle Involved Analysis

Motorist Fatal Count	Pedestrian Falal Count
1,820	1,860

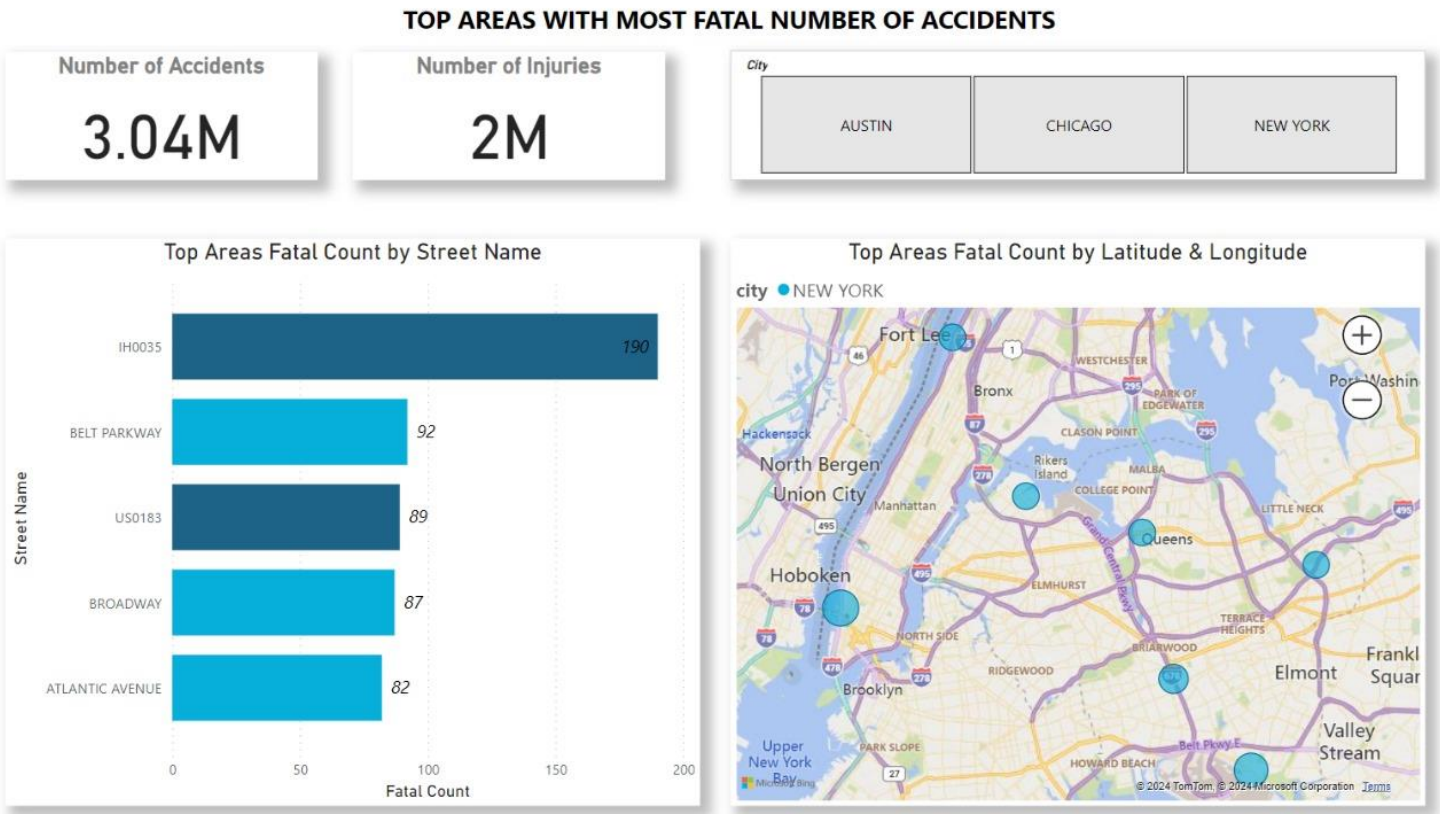
Most Common Factors Involved in Accidents

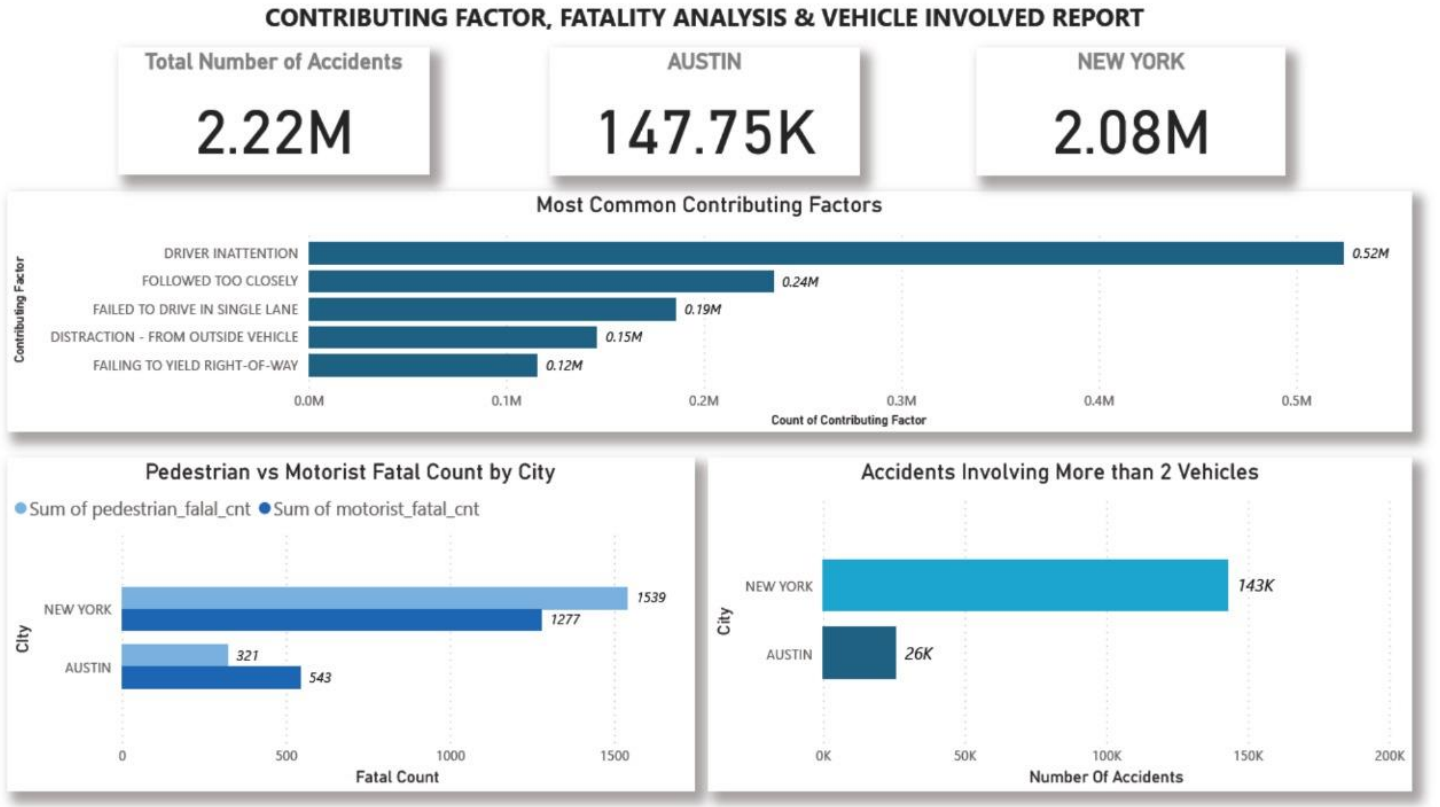


Number of Accidents Involved More than Two Vehicles

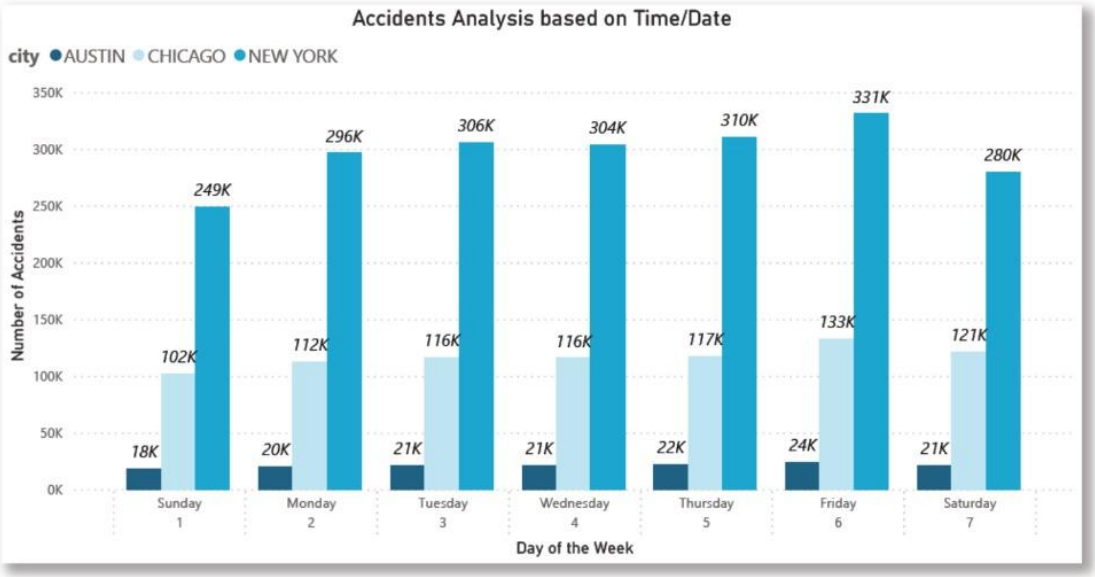


- **PowerBI**

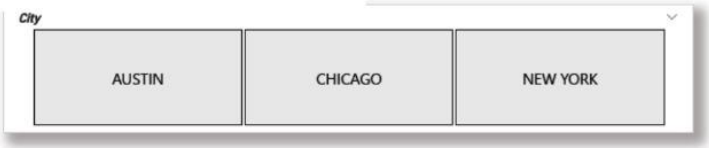

TOP AREAS WITH MOST FATAL NUMBER OF ACCIDENTS
Number of Accidents
3.04M
Number of Injuries
2M
City
AUSTIN
CHICAGO
NEW YORK
Top Areas Fatal Count by Street Name
IH0035
190
BELT PARKWAY
92
US0183
89
BROADWAY
87
ATLANTIC AVENUE
82
Street Name
0
50
100
150
200
Fatal Count
Top Areas Fatal Count by Latitude & Longitude
city NEW YORK
Fort Lee
Bronx
Hackensack
North Bergen
Union City
Manhattan
Rikers Island
COLLEGE POINT
Queens
Hoboken
ELMHURST
NORTH SIDE
RIDGEWOOD
Brooklyn
Elmont
Valley Stream
Upper New York Bay
PARK SLOPE
HOWARD BEACH
© 2024 TomTom, © 2024 Microsoft Corporation Terms

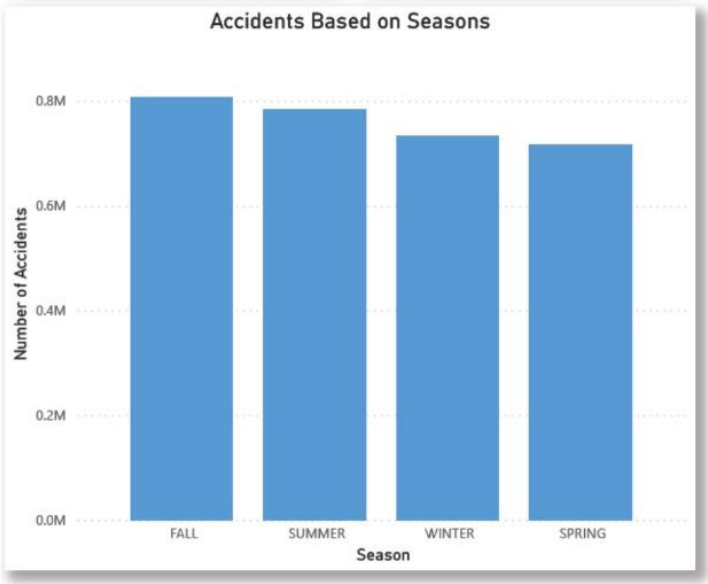

CONTRIBUTING FACTOR, FATALITY ANALYSIS & VEHICLE INVOLVED REPORT
Total Number of Accidents
2.22M
AUSTIN
147.75K
NEW YORK
2.08M
Most Common Contributing Factors
DRIVER INATTENTION
0.52M
FOLLOWED TOO CLOSELY
0.24M
FAILED TO DRIVE IN SINGLE LANE
0.19M
DISTRACTION - FROM OUTSIDE VEHICLE
0.15M
FAILING TO YIELD RIGHT-OF-WAY
0.12M
Contributing Factor
0.0M
0.1M
0.2M
0.3M
0.4M
0.5M
Count of Contributing Factor
Pedestrian vs Motorist Fatal Count by City
Sum of pedestrian_falal_cnt
Sum of motorist_fatal_cnt
NEW YORK
1539
1277
AUSTIN
321
543
City
0
500
1000
1500
Fatal Count
Accidents Involving More than 2 Vehicles
NEW YORK
143K
AUSTIN
26K
City
0K
50K
100K
150K
200K
Number Of Accidents

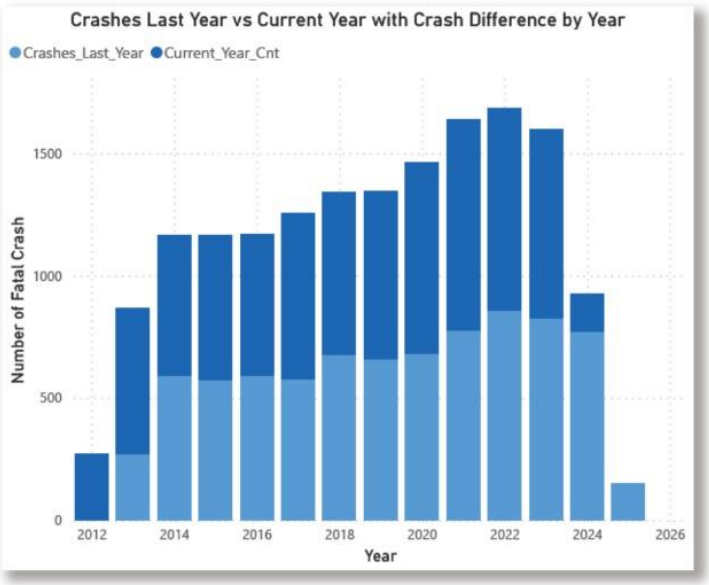
TIME BASED ANALYSIS REPORT



SEASONALITY REPORT OF ACCIDENTS

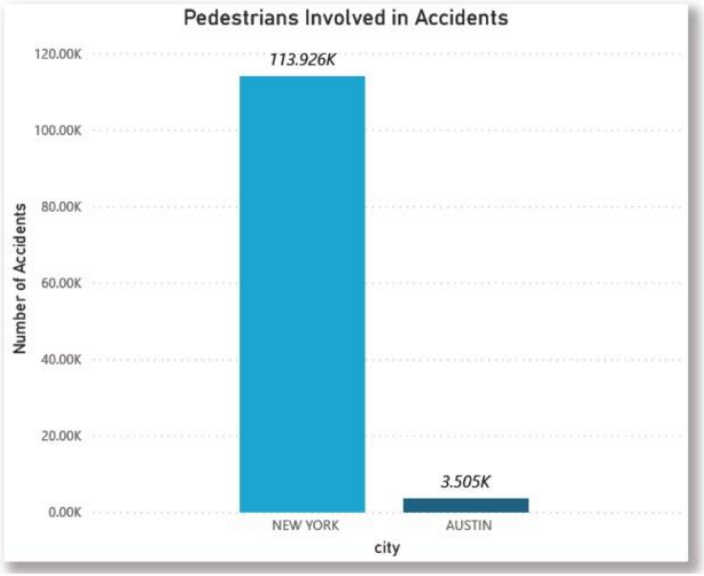


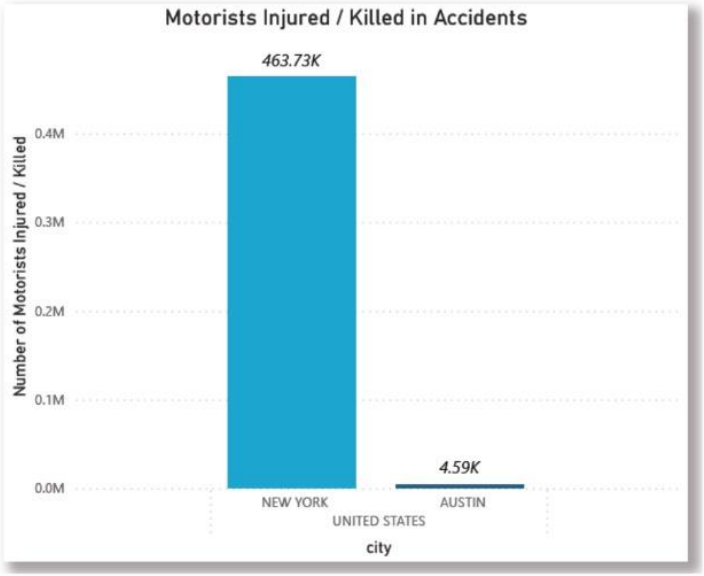




PEDESTRIANS INVOLVED & MOTORISTS INJURED/KILLED IN ACCIDENTS

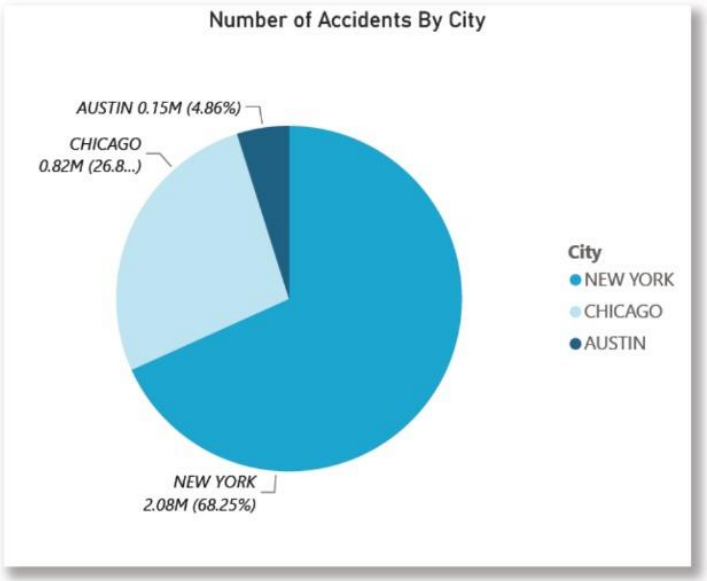
Total Number of Accidents	AUSTIN	NEW YORK
2.22M	147.75K	2.08M

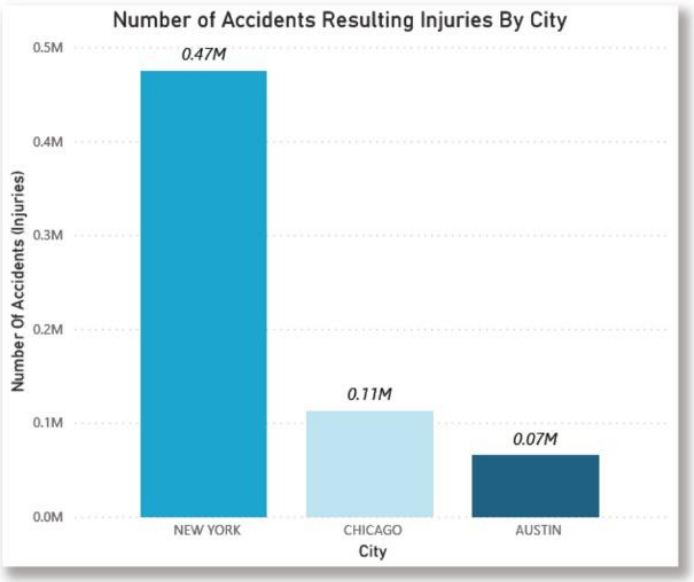




NUMBER OF ACCIDENTS IN EACH CITY

Overall Accidents	AUSTIN	CHICAGO	NEW YORK
3.04M	147.75K	817.72K	2.08M

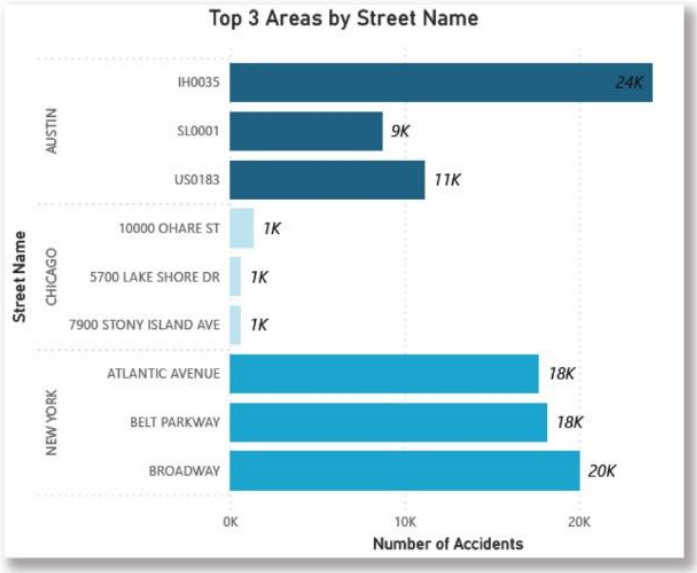




TOP 3 AREAS WITH GREATEST NUMBER OF ACCIDENTS

Number of Accidents	Number of Injuries
3.04M	2M

City: AUSTIN | CHICAGO | NEW YORK



8.Contributions

Subject Area	Contributing Member
Profiling	Manish and Praveen
Staging	Manish and Praveen
Dimension Model	Team
STTM	Mithali with little team contribution
ETL	Mithali and Sathya
Visualization	Tableau (Sathya and Praveen) PowerBI (Manish and Mithali)
SQL Validation	Sathya