Group 14 - Final Report

Prasad Gavas
Shantanu Mahakal
Prarthana Shetty
Soham Shah

Project Overview:

Austin Dataset:

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 Austin Police Department (APD), and maintained by TXDOT.

This dataset contains crash-level records for crashes which have occurred in the last ten years. Crash data takes 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 observed that ensures more accurate and complete results.

Chicago Dataset:

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 Austin Police Department (APD), and maintained by TXDOT.

This dataset contains crash-level records for crashes which 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 observed that ensures more accurate and complete results.

NYC Dataset:

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. It should be noted that the data is preliminary and subject to change when the MV-104AN forms are amended based on revised crash details.

Police officers complete form MV-104AN for all vehicle collisions. The MV-104AN is a New York State form that has all of the details of a traffic collision.

Step 1: Understanding the Dataset

- Analyzed and discussed the Schema and Content of the provided dataset
- Identified key tables, fields, and relationships to establish a foundational understanding
- Noted and handled all unique characteristics or challenges within the dataset's organization
- Identified potential data quality issues, outliers, or anomalies
- Discussed a schema structure for both datasets

Step 2: Data Profiling using ydata_profiling

- Imported the datasets
- Data Import and Initial Assessment:
- Data Quality Review
- Column Profiling
- Data Transformation Insight

Aim of the Assignment:

The aim of this project is to construct a comprehensive and advanced data architecture that will serve as the backbone for business intelligence and analytical solutions in the domain of traffic safety and vehicular incidents. Leveraging detailed crash data from three major cities—New York, Chicago, and Austin—this initiative strives to harmonize disparate datasets into a singular, unified schema that will facilitate in-depth analysis and reporting on various aspects of motor vehicle collisions.

The goal is to create a reliable and scalable data warehouse that will enable stakeholders to uncover critical insights such as accident hotspots, temporal trends in accident occurrences, the incidence of injuries and fatalities, and contributory factors to accidents.

Data Source Observations:

NYC Dataset

Name	Datatype	Unique values	Null values	Metadata	Min Value	Max Value
CRASH DATE	Date	0.2%	0.00%	Occurrence date of collision	20212-07-01	2024-03-22
CRASH TIME	Date	0.1%	0.00%	Occurrence time of collision	2024-03-29 00:00:00	2024-03-29 23:59:00
BOROUGH	Categorical	<0.1%	31.1%	Borough where collision occurred	5	13
ZIP CODE	Text	<0.1%	31.1%	Postal code of incident occurrence	5	5
LATITUDE	Real Number	6.9%	11.3%	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)	0	43.344
LONGITUDE	Real Number	5.3%	11.3%	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326).	-201.359	0
LOCATION	Text	15.4%	11.3%	Latitude , Longitude pair	25	10
ON STREET NAME	Text	1.1%	21.2%	Street on which the collision occurred	2	32
CROSS STREET NAME	Text	1.6%	37.8%	Nearest cross street to the collision	1	32
OFF STREET NAME	Text	64.9%	83.2%	Street address if known	8	40
NUMBER OF PERSONS INJURED	Real Number	<0.1%	18	Number of persons injured	0	43
NUMBER OF PERSONS KILLED	Real Number	<0.1%	<0.1%	Number of persons killed	0	8
NUMBER OF PEDESTRIANS INJURED	Real Number	<0.1%	0.0%	Number of pedestrians injured	0	27
NUMBER OF PEDESTRIANS KILLED	Categorical-Wrong	<0.1%	0%	Number of pedestrians killed	0	6
NUMBER OF CYCLIST INJURED	Categorical-Wrong	<0.1%	0%	Number of cyclists injured	0	4

NUMBER OF CYCLIST KILLED	Categorical	<0.1%	0%	Number of cyclists killed	0	2
NUMBER OF MOTORIST INJURED	Real Number	<0.1%	0%	Number of vehicle occupants injured	0	43
NUMBER OF MOTORIST KILLED	Real Number	<0.1%	0%	Number of vehicle occupants killed	0	5
CONTRIBUTING FACTOR VEHICLE 1	Text	<0.1%	0.3%	Factors contributing to the collision for designated vehicle	1	53
CONTRIBUTING FACTOR VEHICLE 2	Text	<0.1%	15.5%	Factors contributing to the collision for designated vehicle	1	53
CONTRIBUTING FACTOR VEHICLE 3	Categorical	<0.1%	92.9%	Factors contributing to the collision for designated vehicle	1	53
CONTRIBUTING FACTOR VEHICLE 4	Categorical	<0.1%	98.4%	Factors contributing to the collision for designated vehicle	5	43
CONTRIBUTING FACTOR VEHICLE 5	Categorical	0.3%	99.6%	Factors contributing to the collision for designated vehicle	5	43
COLLISION_ID	Real Number	100%	0%	Unique record code generated by system. Primary Key for Crash table.	22	4712252
VEHICLE TYPE CODE 1	Text	0.1%	0.7%	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)	1	38
VEHICLE TYPE CODE 2	Text	0.1%	19.1%	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, e-scooter, truck/bus, motorcycle, other)	1	38
VEHICLE TYPE CODE 3	Text	0.2%	93.1%	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, e-scooter, truck/bus, motorcycle, other)	2	35
VEHICLE TYPE CODE 4	Text	0.3%	98.4%	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, e-scooter, truck/bus, motorcycle, other)	2	35
VEHICLE TYPE CODE 5	Text	0.08%	99.6%	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, e-scooter, truck/bus, motorcycle, other)	2	35

Austin Dataset

Name	Datatype	Unique values	Null values	Metadata	Min Value	Max Value
crash_id	Real_number	100%	0.00%	TxDOT C.R.I.S. system-generated unique identifying number for a crash	1001	1.802 *10^8
crash_fatal_fl	Boolean	<0.1%	0%	Fatal Crash Identifier - Indicates that the crash involved one or more fatalities	False	True
crash_date	Date	97.9%	0.0%	Crash Date	2014-03-26 06:41:00	2024-03-11 22:05:00
crash_time	Date	1%	0.0%	Crash Time - Time crash occurred	2024-03-29 00:00:00	2024-03-29 23:59:00
case_id	Text	99.9%	1.3%	Case ID	1	20
rpt_latitude	Real Number	77.5%	93%	Reported Latitude	25.83746	36.50048
rpt_longitude	Real Number	70.6%	93%	Reported Longitude	-106.645	-93.50795
rpt_block_num	Text	3.7%	13.3%	Reported Block Number (road on which crash occurred)	1	9
rpt_street_pfx	Categorical	<0.1%	45.9%	Reported Street Prefix (road on which crash occurred: N, S, E, W, SW..)	1	2
rpt_street_name	Text	6.6%	<0.1%	Reported Street Name (road on which crash occurred)	1	49
rpt_street_sfx	Categorical	<0.1%	34.1%	Reported Street Suffix (road on which crash occurred)	2	4
crash_speed_limit	Real number	<0.1%	<0.1%	Speed Limit	-1	85
road_constr_zone_fl	Boolean	<0.1%	<0.1%	Construction Zone - Indicates whether the crash occurred in or was related to a construction, maintenance, or utility work zone, regardless of whether or not workers were actually present at the time of the crash	False	True

Latitude	Real Number	66.2%	1.5%	Derived Latitude map coordinate of the crash	30.0987	30.5116
Longitude	Real Number	66.1%	1.5%	Derived Longitude map coordinate of the crash	-97.926	-97.570
street_name	Text	3.1%	<0.1%	Derived Street Name - Name of the road crash occurred on, as determined by the Locator application.	3	41
street_nbr	Real Number	16.2%	58.9%	Derived Street Number - Block number of primary street where crash occurred as determined by the Locator application	0	21146
street_name_2	Text	5.1%	55.1%	Derived Street Name 2 - The road name for the secondary road related to the crash location (If applicable)	3	38
street_nbr_2		<0.1%	0.3%	Derived Street Number 2 - Block number of secondary street related to the crash location as determined by the Locator application (If applicable)	1	53
crash_sev_id	Real Number	<0.1%	0%	Crash Severity - Most severe injury suffered by any one person involved in the crash (0=UNKNOWN, 1=INCAPACITATING INJURY, 2=NON-INCAPACITATING INJURY, 3=POSSIBLE INJURY, 4=KILLED, 5=NOT INJURED)	0	99
sus_serious_injry_cnt	Real Number	<0.1%	0%	Total Suspected Serious Injury Count	0	10
nonincap_injry_cnt	Real Number	<0.1%	<0.1%	Total Non-incapacitating Injury Count	0	14
poss_injry_cnt	Real Number	<0.1%	<0.1%	Total Possible Injury Count	0	20
non_injry_cnt	Real Number	<0.1%	<0.1%	Total Not Injured Count.	0	56
unkn_injry_cnt	Real Number	<0.1%	<0.1%	Total Unknown Injury Count	0	41
tot_injry_cnt	Real Number	<0.1%	<0.1%	Total Injury Count	0	21
death_cnt	Categorical- Wrong	<0.1%	0%	Total Death Count	0	4

contrib_factr_p1_id	Real Number	0.2%	80.6%	The first factor for a given vehicle which the officer felt possibly contributed to the crash	1	80
contrib_factr_p2_id	Real Number	1.4%	96.9%	The second factor for a given vehicle which the officer felt possibly contributed to the crash	1	79
units_involved	Text	0.8%	<0.1%	Mode of units involved in crash	10	321
atd_mode_category_metadata	Text	100%	<0.1%	Description of units involved in crash	214	3936
pedestrian_fl	Boolean	<0.1%	97.6%	Pedestrian involved crash flag	Missing	True
motor_vehicle_fl	Boolean	<0.1%	0.8%	Motor vehicle involved crash flag	Missing	True
motorcycle_fl	Boolean	<0.1%	97.6%	Motorcycle involved crash flag	Missing	True
bicycle_fl	Boolean	<0.1%	98.3%	Bicyclist involved crash flag	Missing	True
other_fl	Boolean	<0.1%	96.7%	Other involved crash flag	Missing	True
point	Text	67.2%	1.5%	Point data type created with crash latitude and longitude to enable request of GeoJSON.	22	45
apd_confirmed_fatality	Boolean	<0.1%	0%	APD Fatality flag	False	True
apd_confirmed_death_count	Categorical-Wrong	<0.1%	0%	APD Fatality Count	0	4
motor_vehicle_death_count	Categorical-Wrong	<0.1%	0%		0	4
motor_vehicle_serious_injury_count	Real Number	<0.1%	0%		0	5
bicycle_death_count	Categorical-Wrong	<0.1%	0%		0	1
bicycle_serious_injury_count	Categorical-Wrong	<0.1%	0%		0	3
pedestrian_death_count	Categorical-Wrong	<0.1%	0%		0	2
pedestrian_serious_injury_count	Categorical-Wrong	<0.1%	0%		0	9
motorcycle_death_count	Categorical-Wrong	<0.1%	0%		0	2
motorcycle_serious_injury_count	Categorical-Wrong	<0.1%	0%		0	2
other_death_count	Categorical-Wrong	<0.1%	0%		0	0
other_serious_injury_count	Categorical-Wrong	<0.1%	0%		0	3

onsys_fl	Boolean	<0.1%	0%	Flag indicates whether primary road of crash was on the TxDOT highway system.	False	True
private_dr_fl	Boolean	<0.1%	0%	Flag indicating whether crash occurred on a private drive or road/private property/parking lot.	False	False
micromobility_serious_injury_count	Categorical-Wrong	<0.1%	0%		0	2
micromobility_death_count	Categorical-Wrong	<0.1%	0%		0	1
micromobility_fl	Boolean	0.3%	99.8%		Missing	True

Chicago Dataset:

Name	Datatype	Unique values	Null values	Metadata	Min Value	Max Value
CRASH_RECORD_ID	Text	100%	0.00%	This number can be used to link to the same crash in the Vehicles and People datasets. This number also serves as a unique ID in this dataset	128	128
CRASH_DATE_EST_I	Boolean	<0.1%	92.5%	Crash date estimated by desk officer or reporting party (only used in cases where crash is reported at police station days after the crash)	False	True
CRASH_DATE	Date	65.7%	0.0%	Date and time of crash as entered by the reporting officer	2013-03-03 16:48:00	2024-03-26 01:40:00
POSTED_SPEED_LIMIT	Real Number	<0.1%	0.0%	Posted speed limit, as determined by reporting officer	0	99
TRAFFIC_CONTROL_DEVICE	Categorical	<0.1%	0%	Traffic control device present at crash location, as determined by reporting officer	5	24
DEVICE_CONDITION	Categorical	<0.1%	0%	Condition of traffic control device, as determined by reporting officer	5	24
WEATHER_CONDITION	Categorical	<0.1%	0%	Weather condition at time of crash, as determined by reporting officer	4	24
LIGHTING_CONDITION	Categorical	<0.1%	0%	Light condition at time of crash, as determined by reporting officer	4	22
FIRST_CRASH_TYPE	Categorical	<0.1%	0%	Type of first collision in crash	5	28
TRAFFICWAY_TYPE	Categorical	<0.1%	0%	Trafficway type, as determined by reporting officer	4	31
LANE_CNT	Real Number	<0.1%	75%	Total number of through lanes in either direction, excluding turn lanes, as determined by reporting officer (0 = intersection)	0	1191625
ALIGNMENT	Categorical	<0.1%	0%	Street alignment at crash location, as determined by reporting officer	12	21
ROADWAY_SURFACE_COND	Categorical	<0.1%	0%	Road surface condition, as determined by reporting officer	3	15
ROAD_DEFECT	Categorical	<0.1%	0%	Road defects, as determined by reporting officer	5	17

REPORT_TYPE	Categorical	<0.1%	3.0%	Administrative report type (at scene, at desk, amended)	7	26
CRASH_TYPE	Categorical	<0.1%	0%	A general severity classification for the crash. Can be either Injury and/or Tow Due to Crash or No Injury / Drive Away	22	32
INTERSECTION_RELATED_I	Boolean	<0.1%	77.1%	A field observation by the police officer whether an intersection played a role in the crash. Does not represent whether or not the crash occurred within the intersection.	FALSE	TRUE
NOT_RIGHT_OF_WAY_I	Boolean	<0.1%	95.4%	Whether the crash begun or first contact was made outside of the public right-of-way.	FALSE	TRUE
HIT_AND_RUN_I	Boolean	<0.1%	68.7%	Crash did/did not involve a driver who caused the crash and fled the scene without exchanging information and/or rendering aid	FALSE	TRUE
DAMAGE	Categorical	<0.1%	0%	A field observation of estimated damage.	11	13
DATE_POLICE_NOTIFIED	Date	75.9%	0%	Calendar date on which police were notified of the crash	2013-06-01 20:31:00	2024-03-26 01:42:00
PRIM_CONTRIBUTORY_CAUSE	Categorical	<0.1%	0%	The factor which was most significant in causing the crash, as determined by officer judgment	6	80
SEC_CONTRIBUTORY_CAUSE	Categorical	<0.1%	0%	The factor which was second most significant in causing the crash, as determined by officer judgment	6	80
STREET_NO	Real Number	1.4%	0%	Street address number of crash location, as determined by reporting officer	0	451100
STREET_DIRECTION	Categorical	<0.1%	<0.1%	Street address direction (N,E,S,W) of crash location, as determined by reporting officer	1	1
STREET_NAME	Text	<0.2%	<0.1%	Street address name of crash location, as determined by reporting officer	4	31
BEAT_OF_OCCURRENCE	Real Number	<0.1%	<0.1%	Chicago Police Department Beat ID. Boundaries available at https://data.cityofchicago.org/d/aerh-rz74	111	6100

PHOTOS_TAKEN_I	Boolean	2%	98.7%	Whether the Chicago Police Department took photos at the location of the crash	FALSE	TRUE
STATEMENTS_TAKEN_I	Boolean	<0.1%	97.8%	Whether statements were taken from unit(s) involved in crash	FALSE	TRUE
DOORING_I	Boolean	0.1%	99.7%	Whether crash involved a motor vehicle occupant opening a door into the travel path of a bicyclist, causing a crash	FALSE	TRUE
WORK_ZONE_I	Boolean	<0.1%	99.4%	Whether the crash occurred in an active work zone	FALSE	TRUE
WORK_ZONE_TYPE	Categorical	0.1%	99.6%	The type of work zone, if any	7	12
WORKERS_PRESENT_I	Boolean	0.2%	99.9%	Whether construction workers were present in an active work zone at crash location	FALSE	TRUE
NUM_UNITS	Real Number	<0.1%	0%	Number of units involved in the crash. A unit can be a motor vehicle, a pedestrian, a bicyclist, or another non-passenger roadway user. Each unit represents a mode of traffic with an independent trajectory.	1	18
MOST_SEVERE_INJURY	Categorical	<0.1%	0.2%	Most severe injury sustained by any person involved in the crash	5	24
INJURIES_TOTAL	Real Number	<0.1%	0.2%	Total persons sustaining fatal, incapacitating, non-incapacitating, and possible injuries as determined by the reporting officer	0	21
INJURIES_FATAL	Categorical-Wrong	<0.1%	0.2%	Total persons sustaining fatal injuries in the crash.	0	4
INJURIES_INCAPACITATING	Real Number	<0.1%	0.2%	Total persons sustaining incapacitating/serious injuries in the crash as determined by the reporting officer. Any injury other than fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities they were capable of performing before the injury occurred. Includes severe lacerations, broken limbs, skull or chest injuries, and abdominal injuries.	0	10

INJURIES_NON_INCAPACITATING	Real Number	<0.1%	0.2%	Total persons sustaining non-incapacitating injuries in the crash as determined by the reporting officer. Any injury, other than fatal or incapacitating injury, which is evident to observers at the scene of the crash. Includes lump on head, abrasions, bruises, and minor lacerations.	0	21
INJURIES_REPORTED_NOT_EVIDENT	Real Number	<0.1%	0.2%	Total persons sustaining possible injuries in the crash as determined by the reporting officer. Includes momentary unconsciousness, claims of injuries not evident, limping, complaint of pain, nausea, and hysteria.	0	15
INJURIES_NO_INDICATION	Real Number	<0.1%	0.2%	Total persons sustaining no injuries in the crash as determined by the reporting officer	0	61
INJURIES_UNKNOWN	Categorical	<0.1%	0.2%	Total persons for whom injuries sustained, if any, are unknown	3	3
CRASH_HOUR	Real Number	<0.1%	0%	The hour of the day component of CRASH_DATE.	0	23
CRASH_DAY_OF_WEEK	Real Number	<0.1%	0%	The day of the week component of CRASH_DATE. Sunday=1	1	7
CRASH_MONTH	Real Number	<0.1%	0%	The month component of CRASH_DATE.	1	12
LATITUDE	Real Number	37%	0.7%	The latitude of the crash location, as determined by reporting officer, as derived from the reported address of crash	0	42.02278
LONGITUDE	Real Number	36.9%	0.7%	The longitude of the crash location, as determined by reporting officer, as derived from the reported address of crash	-87.936193	0
LOCATION	Text	37%	0.7%	The crash location, as determined by reporting officer, as derived from the reported address of crash, in a column type that allows for mapping and other geographic analysis in the data portal software	11	40

NYC Dataset:

CRASH DATE & CRASH TIME: These columns contain the date and time of the collision, respectively. They have been appropriately formatted and are crucial for analyzing trends over time.

BOROUGH & ZIP CODE: These categorical variables identify the location of the collision, with a significant percentage of null values. Boroughs where collisions occur are crucial for understanding geographic patterns.

LATITUDE & LONGITUDE: Real number coordinates of the collision location. They have a considerable percentage of null values and provide precise spatial information.

LOCATION: This column contains latitude and longitude pairs, which can be utilized for mapping and spatial analysis. However, it has a high percentage of null values.

NUMBER OF PERSONS INJURED & NUMBER OF PERSONS KILLED: These numerical variables provide information on the severity of the collisions, with a low percentage of null values.

CONTRIBUTING FACTOR VEHICLE 1-5: These categorical variables identify factors contributing to collisions, with varying percentages of null values.

COLLISION_ID: This serves as a unique identifier for each collision record and is crucial for data management.

VEHICLE TYPE CODE 1-5: These categorical variables describe the types of vehicles involved in the collisions, with varying percentages of null values.

Key Observations

Spatial Distribution:Spatial distribution analysis using latitude and longitude coordinates enables us to map and visualize the geographical spread of collisions across New York City. By plotting these coordinates on a map, we can identify collision hotspots, concentration areas, and spatial patterns of collision occurrence.

Temporal Trends: we observed interesting temporal patterns in the crash date and time columns, indicating potential daily, weekly, or seasonal variations in collision occurrence. Understanding these trends is crucial for identifying high-risk periods and implementing targeted interventions to enhance road safety.

Severity Analysis: By examining the number of persons injured and killed, we gained insight into the severity of collisions. Understanding the factors contributing to severe collisions is essential for developing effective strategies to reduce injuries and fatalities on NYC roads and to answer the business requirements.

Contributing Factors: Analysis of contributing factor variables revealed common causes of collisions, such as distracted driving, speeding, or poor road conditions. Addressing these factors through targeted interventions and public awareness campaigns is crucial for improving overall road safety.

Data Quality Concerns: The presence of null values and inconsistencies in data completeness across columns raises concerns about data quality. It's imperative to conduct thorough data cleansing and validation to ensure the reliability of analytical results and insights derived from the dataset.

Austin Dataset:

crash_id: This column serves as a unique identifier.

crash_date & crash_time: These columns represent the date and time of the collision occurrence, crucial for temporal analysis.

Latitude & Longitude: Derived coordinates of the collision location, providing spatial information for mapping and analysis.

crash_sev_id: This numerical variable indicates the severity of the crash, ranging from unknown to fatal injuries.

contrib_factr_p1_id & contrib_factr_p2_id: These numerical variables identify primary and secondary contributing factors to the collision, respectively.

Key Observations:

Temporal Patterns: Analysis of crash date and time revealed distinct temporal trends in collision occurrence, including daily, weekly, or hourly patterns. This information is essential for implementing timely interventions and resource allocation during high-risk periods.
Spatial Distribution: With latitude and longitude coordinates available, we conducted spatial analysis to identify collision hotspots and distribution patterns across Austin. Understanding these patterns can inform urban planning and infrastructure improvements to enhance road safety.
Severity Assessment: By examining the crash_sev_id column, We gained insight into the severity of collisions. Understanding factors contributing to severe collisions is crucial for developing strategies to reduce injury rates and improve road safety.
Contributing Factors: Analysis of contributing factor variables highlighted common causes of collisions in Austin, such as impaired driving or adverse weather conditions. Addressing these factors through targeted interventions can mitigate collision risks and enhance road safety.
Data Quality Considerations: Similar to the NYC dataset, concerns regarding null values and data completeness necessitate thorough data cleansing and validation processes. Ensuring data accuracy and reliability is essential for deriving actionable insights from the dataset.

Chicago Dataset:

CRASH_RECORD_ID: These columns serve as identifiers

CRASH_DATE & CRASH_TIME: These columns represent the date and time of the collision occurrence, essential for temporal analysis.

POSTED_SPEED_LIMIT & TRAFFIC_CONTROL_DEVICE: These categorical variables provide information on speed limits and traffic control devices present at the crash location.

WEATHER_CONDITION & LIGHTING_CONDITION: These categorical variables describe weather and lighting conditions at the time of the collision, impacting visibility and road conditions.

LANE_CNT & ALIGNMENT: These numerical and categorical variables offer insights into the number of lanes and street alignment at the crash location.

PRIM_CONTRIBUTORY_CAUSE & SEC_CONTRIBUTORY_CAUSE: These categorical variables identify primary and secondary contributing factors to the collision, respectively.

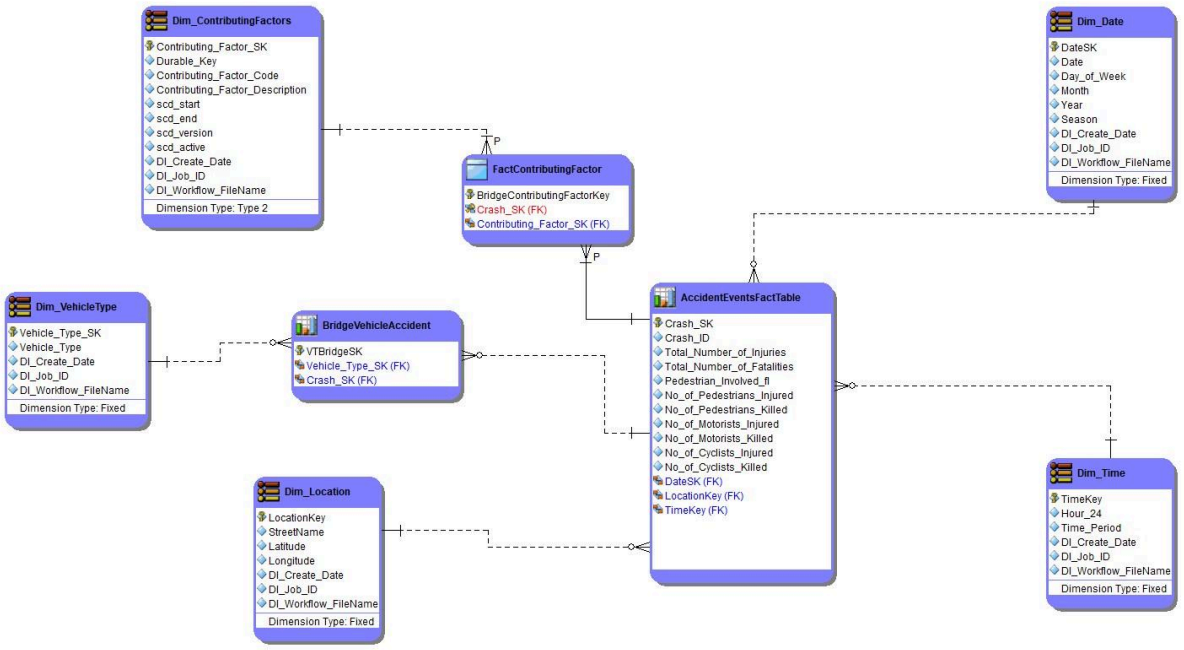
NUM_UNITS & MOST_SEVERE_INJURY: These numerical and categorical variables provide information on the number of units involved and the most severe injury sustained in the collision.

Key Observations:

Temporal Analysis: Analysis of crash date and time revealed distinct temporal patterns in collision occurrence, including daily, weekly, or monthly trends. Understanding these patterns is essential for implementing timely interventions and allocating resources effectively.
Environmental Conditions: Variables such as weather_condition and lighting_condition provided insights into environmental factors influencing collision occurrence. Analyzing the impact of these factors on collision rates can inform strategies to improve road safety under adverse conditions.
Roadway Characteristics: By examining lane_cnt and alignment variables, we gained insights into roadway features at crash locations. Assessing the relationship between these features and collision occurrence is crucial for guiding infrastructure improvements and traffic management strategies.
Contributing Factors: Analysis of contributing cause variables identified primary and secondary factors contributing to collisions in Chicago. Understanding these factors is essential for developing targeted interventions to address root causes of collisions.
Data Completeness and Accuracy: Concerns regarding null values and data completeness across columns underscore the importance of thorough data validation and cleansing. Addressing these issues is essential to ensure the reliability and validity of analytical insights derived from the dataset.

Note: In the process of profiling all 3 accident dataset, a critical observation has been made regarding the data types assigned to certain variables. A few variables have been inaccurately classified as categorical due to the presence of a limited number of distinct values within the dataset. However, these variables are inherently continuous and should not be confined to observed categories. These variables have been duly highlighted in yellow to signify their misclassification.
For instance, variables such as NUMBER OF PERSONS INJURED and NUMBER OF PERSONS KILLED might exhibit a small range of distinct numbers due to the dataset's constraints (e.g., most accidents involve 0-5 injuries). However, their theoretical range is not bounded by these observed values, and they can take on any non-negative integer value.

DIMENSIONAL MODELING:



Dimension Tables:

Dim_ContributingFactors:

- Keys: Contributing_Factor_SK (surrogate key), Durable_Key
- Attributes: Contributing_Factor_Code, Contributing_Factor_Description, SCD (Slowly Changing Dimension) attributes (start, end, version, active), and metadata fields (DI_Create_Date, DI_Job_ID, DI_Workflow_FileName)
- Type: Type 2 Dimension, which means it tracks historical changes to data.

Dim_VehicleType:

- Keys: Vehicle_Type_SK
- Attributes: Vehicle_Type, and metadata fields (DI_Create_Date, DI_Job_ID, DI_Workflow_FileName)
- Type: Fixed Dimension, indicating that the content does not change over time.

Dim_Location:

- Keys: LocationKey
- Attributes: StreetName, Latitude, Longitude, and metadata fields (DI_Create_Date, DI_Job_ID, DI_Workflow_FileName)
- Type: Fixed Dimension, implying stable location data.

Dim_Date:

- Keys: DateSK
- Attributes: Date, Day_of_Week, Month, Year, Season, and metadata fields (DI_Create_Date, DI_Job_ID, DI_Workflow_FileName)
- Type: Fixed Dimension, for time-invariant date data.

Dim_Time:

- Keys: TimeKey
- Attributes: Hour_24, Time_Period, and metadata fields (DI_Create_Date, DI_Job_ID, DI_Workflow_FileName)
- Type: Fixed Dimension, for consistent time data.

Fact Tables:

FactContributingFactor:

- Keys: BridgeContributingFactorKey
- Foreign Keys: Contributing_Factor_SK
- A bridge linking contributing factors to vehicle accidents.

BridgeVehicleAccident:

- Keys: TVBridgeSK
- Foreign Keys: Vehicle_Type_SK, Crash_SK
- A bridge table connecting vehicle types and accidents.

AccidentEventsFactTable:

- Keys: Crash_SK
- Attributes: Total_Number_of_Injuries, Total_Number_of_Fatalities, counts of injured/killed pedestrians, motorists, and cyclists, and foreign keys linking to dimensions (DateSK, LocationKey, TimeKey).
- Central fact table for accident events, tracking key accident metrics.

Chosen Data Columns and Justification Austin:

The data columns chosen for the analysis of traffic accidents in Austin, as outlined previously, serve as a template for the selection of analogous columns from the New York dataset. The selection is guided by the necessity to answer pivotal business questions related to accident occurrence, location, severity, and contributing factors. Hereafter is the justification for the inclusion of each column, which is believed to hold relevance for NYC and Chicago datasets as well:

Core Identifiers (crash_id, crash_date, crash_time)

- These columns serve as the primary identifiers for each incident, enabling us to track individual accidents, pinpoint their occurrence in time, and provide foundational data for frequency analysis.

Geographic Coordinates (latitude, longitude, street_name)

- Geographic data offers precise accident location information, crucial for identifying high-risk areas and informing resource allocation for emergency response services and urban planning.

Injuries & Fatalities Indicators (motor_vehicle_death_count, serious_injury_count, etc.)

- The impact of each accident on human life is a critical aspect of road safety analysis. These columns help differentiate between the severity of incidents, a vital factor in policy formulation and priority setting.

Contributing Factors (contrib_factr_p1_id, contrib_factr_p2_id)

- The identification of contributing factors is essential for preventive measures. This data provides insight into potential areas for intervention, such as road conditions, vehicle performance issues, or driver behavior patterns.

Application of Data for Business Questions

The chosen columns provide the necessary context to address several business-oriented questions:

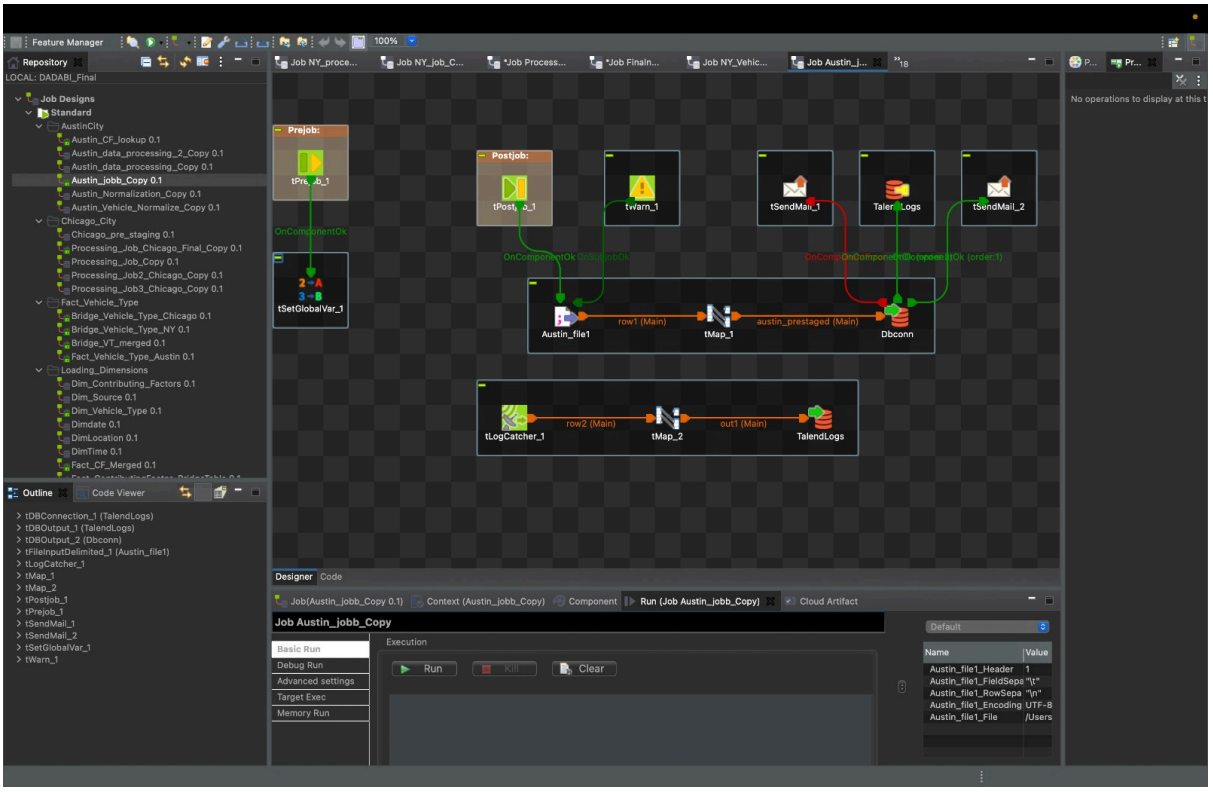
- Accident Frequency: By counting entries based on core identifiers (Crash_ID), we can determine the number of accidents in each city.

- High-Risk Locations: Geographic data enables us to create heat maps highlighting the top three areas with the greatest number of accidents. (LATITUDE, LONGITUDE)
- Injury Analysis: Severity indicators allow for a dual-level report on accidents resulting in injuries, both overall and by city. (No_of_injuries)
- Pedestrian Involvement: Columns indicating pedestrian involvement offer specific data for overall and city-level analysis of pedestrian safety.(CRASH_ID,PEDESTRIANS_FL)
- Accident Timing: Date and time information assist in generating seasonality reports to understand when most accidents happen.(CRASH_ID, HOUR_24)
- Motorist Safety: Severity indicators related to motor vehicles provide data for reports on motorist injuries or fatalities.
- Fatal Accident Zones: Geographic and severity data combined reveal the top five fatal areas across the cities.
- Temporal Trends: Core identifiers and time columns enable time-based analysis to determine patterns in accident occurrences.

Austin Jobs:

Dataset	Job Number	Job Name	Source Table	Target Table	Columns Transformed	Transformation
Austin	1	Processing_Job3_NY	delimited Austin_file1	austin_pre_staging1	-	
	2	Austin_data_processing_Copy	austin_pre_staging1	Austin_p_part1	crash_date, apd_confirmed_death_count,motor_vehicle_death_count,motor_vehicle_serious_injury_count,bicycle_death_count, bicycle_serious_injury_count,pedestrian_death_count, pedestrian_serious_injury_count, motorcycle,death_count,motorcycle_serious_injury_count,other_death_count,other_serious_injury_count,micromobility_serious_injury_count, micormobility_death_count, city added	Formatted according to required date, handling nulls for all integer values
	3	Austin_data_processing_2_Copy	Austin_p_part1	Austin_p_part2	latitude, longitude, contrib_factr_p1_id, contrib_factr_p2_id	Handled nulls for all
	4	Austin_Normalization_Copy	Austin_p_part2	Austin_p_part3	contrib_factr_p1_id, contrib_factr_p2_id both concatenated and normalized	Both contributing factors are concatenated, normalized and filtered for duplicates, normalized contributing factors , filtered contributing factors for duplicates
	5	Austin_Vehicle_Normalize_Copy	Austin_p_part3	Austin_p_part4	Units_Invlolved	Normalized units_involved, trimmed units_involved and handled null values with Other/Unknown
	6	Austin_CF_lookup	Austin_p_part4, Austin_Lookup_CF	Austin_p_final	Contibuting_Factor_Description	1: Added column Contibuting_Factor_Description via lookup on Contributing_Factor, 2: Contibuting_Factor_Description (assigned others to all null values)

Austin Job 1:



Initialization (Pre-Job)

Component: tPrejob_1
Function: Executes initial setup operations necessary before the main ETL process begins. This includes establishing global variable settings for inputting appropriate tFileInputDelimited_1 according to current date and ensuring all system prerequisites are met.

Warning Generator

Component: tWarn_1
Function: Configured to generate warnings that might arise during the pre-job phase. These warnings help in monitoring the job's execution without aborting the process for non-critical issues. They are caught by the tLogCatcher_1

Data Extraction

Component: Austin_file1 (tFileInputDelimited_1)
Function: Reads daily accident data files from Austin with a dynamic file path that incorporates the date variable. This ensures that the correct file is processed according to the scheduled job run.

File Name/Stream

Configuration: The file path includes a global variable configured to have the date component to identify the correct input file, suggesting the handling of time-sequenced data files.

Data Transformation

Components: tMap_1 and tMap_2

Function: These components perform pre-staging data. They map input fields to their respective output fields necessary for subsequent processing steps.

Logging

Component: tLogCatcher_1
Function: Captures log messages, including errors and warnings generated by other components. This component centralized logging, thereby simplifying error handling and debugging.

Notification

Components: tSendMail_1 and tSendMail_2
Function: These components are configured to send out email notifications post-job execution, which include completion confirmations, error reports, or data quality summaries to designated stakeholders.

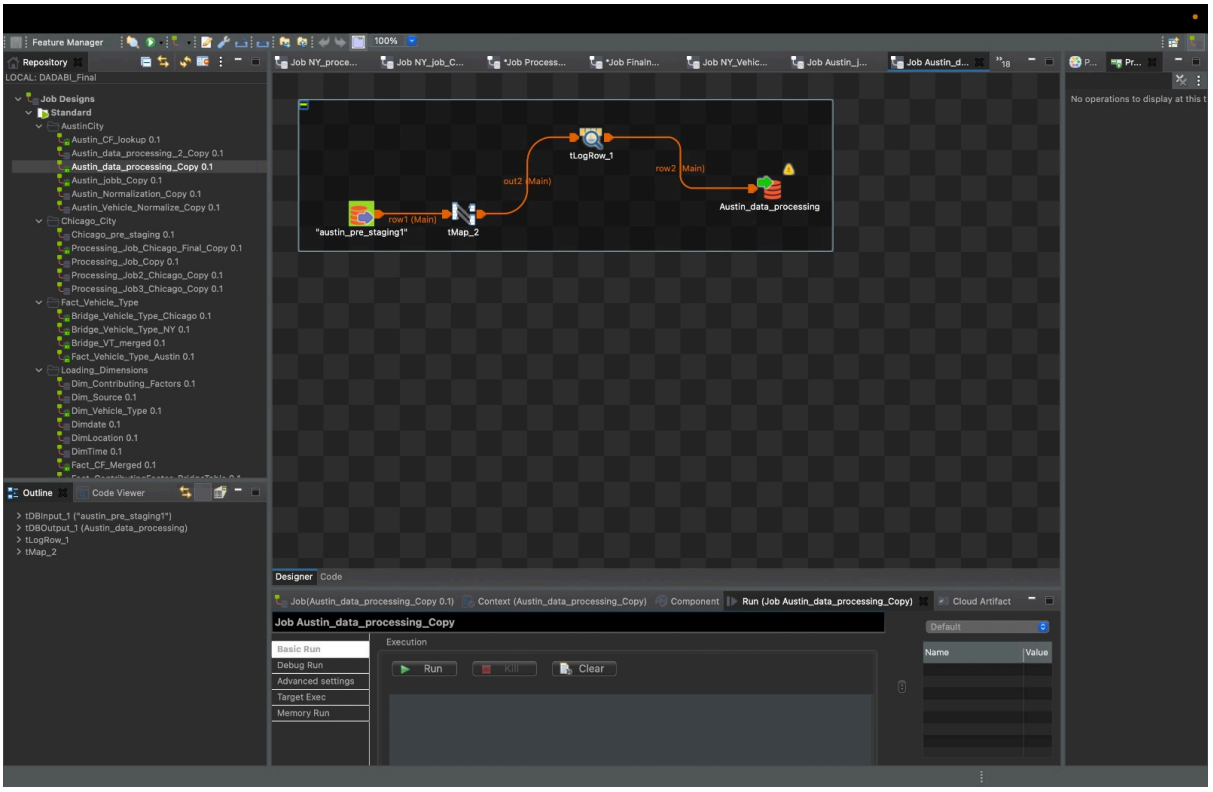
Variable Setting

Component: tSetGlobalVar_1
Function: Sets a global variable used across the job, to pass parameters or conditions between components during execution.

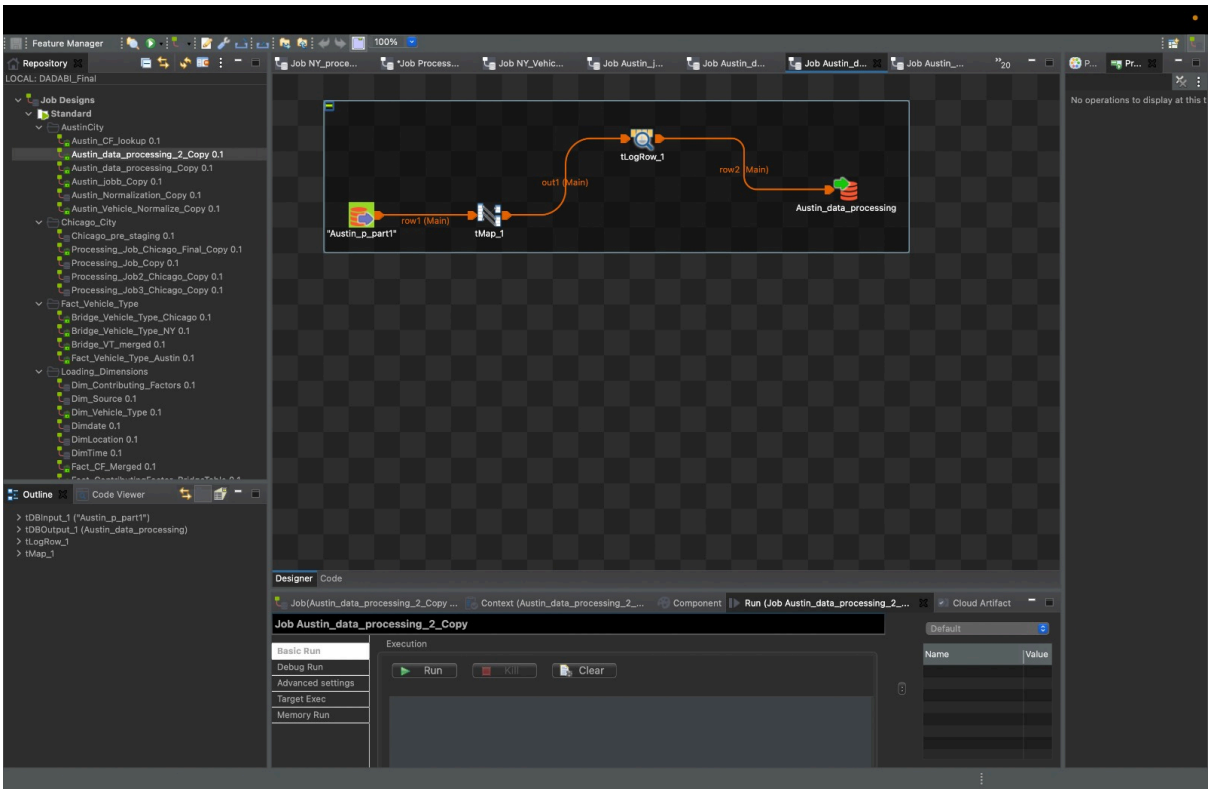
Clean-Up (Post-Job)

Component: tPostjob_1
Function: Executes after the pre-job to perform tasks, such as mapping connections, sending final notifications, etc

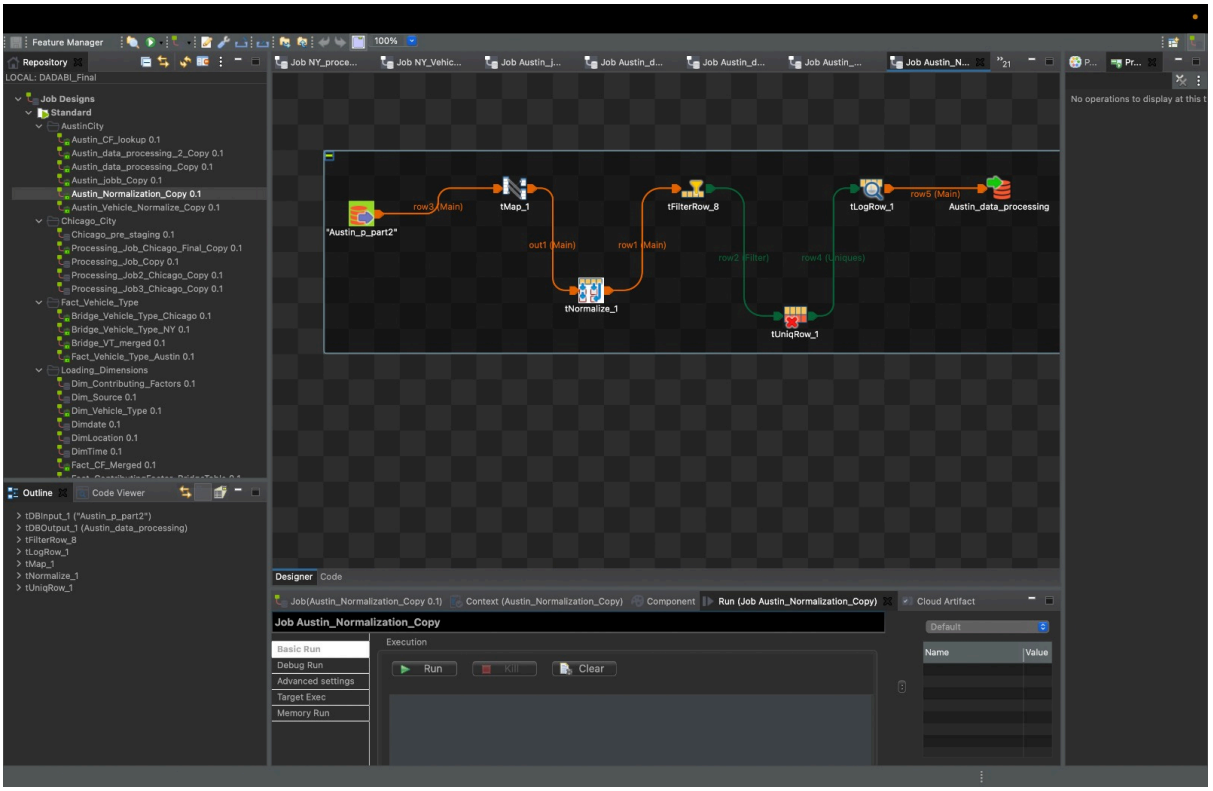
Austin Job 2 :



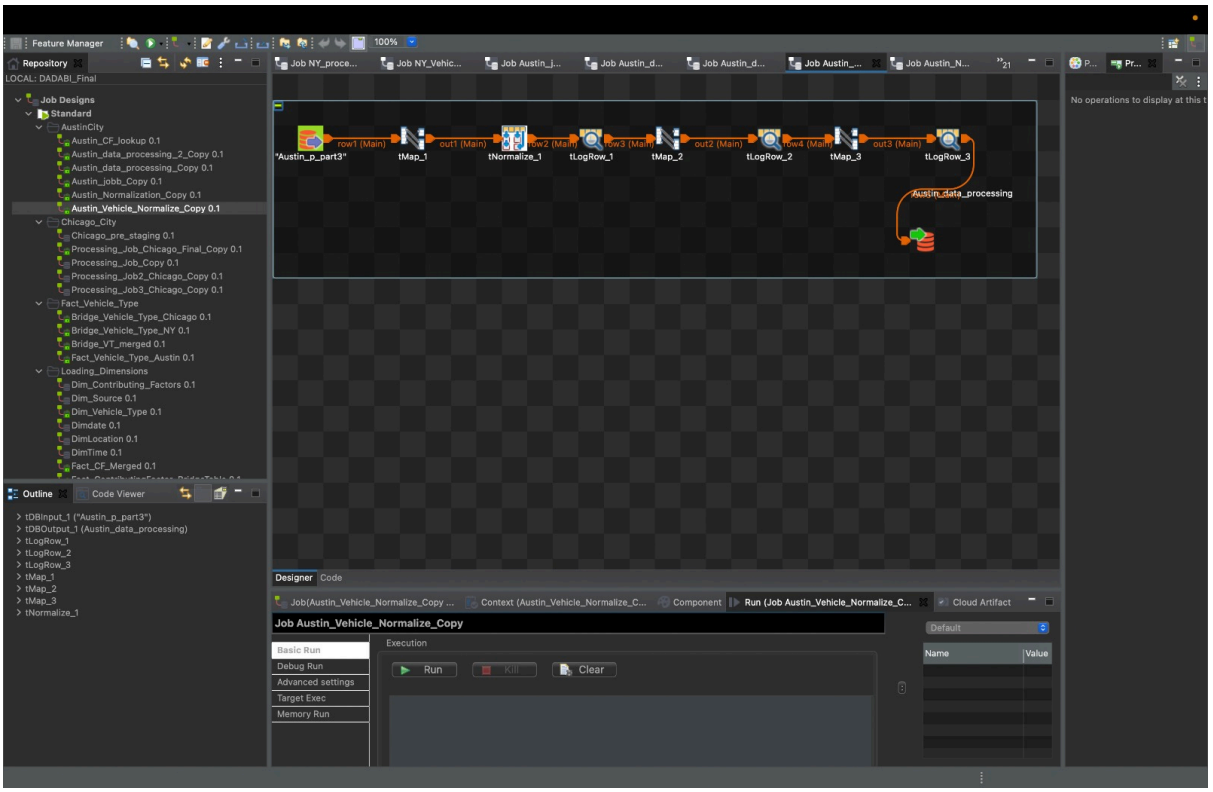
Austin Job 3 :



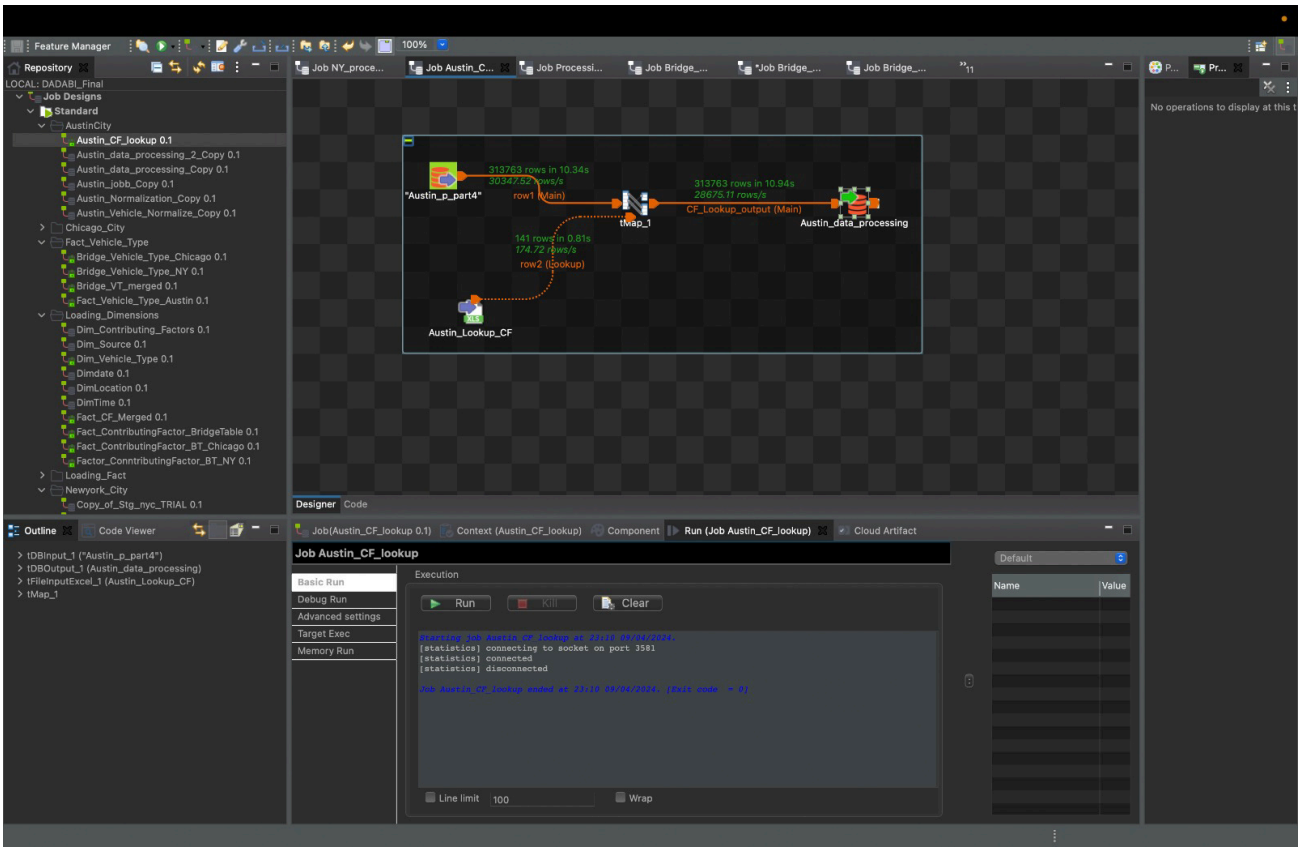
Austin Job 4:



Austin Job 5:



Austin Job 6:


Feature Manager
Repository
LOCAL: DADABI_Final
Job Designs
Standard
AustinCity
Austin_CF_lookup 0.1
Austin_data_processing_2_Copy 0.1
Austin_data_processing_Copy 0.1
Austin_jobb_Copy 0.1
Austin_Normalization_Copy 0.1
Austin_Vehicle_Normalize_Copy 0.1
Chicago_City
Fact_Vehicle_Type
Bridge_Vehicle_Type_Chicago 0.1
Bridge_Vehicle_Type_NY 0.1
Bridge_VT_merged 0.1
Fact_Vehicle_Type_Austin 0.1
Loading_Dimensions
Dim_Contributing_Factors 0.1
Dim_Source 0.1
Dim_Vehicle_Type 0.1
Dimdate 0.1
DimLocation 0.1
DimTime 0.1
Fact_CF_Merged 0.1
Fact_ContributingFactor_BridgeTable 0.1
Fact_ContributingFactor_BT_Chicago 0.1
Factor_ConntributingFactor_BT_NY 0.1
Loading_Fact
Newyork_City
Copy_of_Stg_nyc_TRIAL 0.1
Job NY_proce...
Job Austin_C...
Job Processi...
Job Bridge_...
*Job Bridge_...
Job Bridge_...
313763 rows in 10.34s
30347.52 rows/s
"Austin_p_part4"
row1 (Main)
tMap_1
313763 rows in 10.94s
28675.11 rows/s
CF_Lookup_output (Main)
Austin_data_processing
141 rows in 0.81s
174.72 rows/s
row2 (Lookup)
Austin_Lookup_CF
Designer
Code
Outline
Code Viewer
tDBInput_1 ("Austin_p_part4")
tDBOutput_1 (Austin_data_processing)
tFileInputExcel_1 (Austin_Lookup_CF)
tMap_1
Job(Austin_CF_lookup 0.1)
Context (Austin_CF_lookup)
Component
Run (Job Austin_CF_lookup)
Cloud Artifact
Job Austin_CF_lookup
Basic Run
Debug Run
Advanced settings
Target Exec
Memory Run
Execution
Run
Kill
Clear
[statistics] connecting to socket on port 3581
[statistics] connected
[statistics] disconnected
Line limit
100
Wrap
Default
Name
Value

Chicago

Introduction

In response to the need for a comprehensive understanding of traffic accidents across major cities, we have conducted an extensive data selection process for the city of Chicago. This report justifies the choice of data columns from the Chicago dataset, aligning with the business objectives of quantifying, qualifying, and analyzing traffic accidents to inform policy and improve public safety.

Data Column Selection Justification

The columns chosen from the Chicago dataset are critical for creating an accurate and detailed portrayal of traffic accidents. Below is a justification for each column selected, directly corresponding to the project's business questions:

CRASH_RECORD_ID

- Serves as a unique identifier for each accident, which is fundamental for counting the total number of accidents and tracking individual cases over time.

CRASH_DATE

- Provides the date of the accident, allowing for temporal analysis and identifying trends over days, months, or years. The ability to derive time from the date also adds the dimension of time-based patterns.

LATITUDE & LONGITUDE

- These geographic coordinates are essential for pinpointing accident locations. They enable the mapping of accidents and identification of hotspots within the city.

STREET_NAME

- Complements the geographic coordinates and aids in identifying high-risk streets and intersections, facilitating targeted interventions in the top areas with the most accidents.

WEATHER_CONDITION

- Weather plays a significant role in driving conditions. Understanding its impact on accident frequency can guide infrastructure improvements and public advisories.

INJURIES_FATAL & INJURIES_TOTAL

- Key indicators of accident severity, these columns provide statistics for accidents resulting in injuries or fatalities, both overall and disaggregated by city.

PRIM_CONTRIBUTORY_CAUSE & SEC_CONTRIBUTORY_CAUSE

- Recognizing the primary and secondary causes of accidents is vital for preventive measures, informing both public safety campaigns and road users' education.

Business Questions Addressed

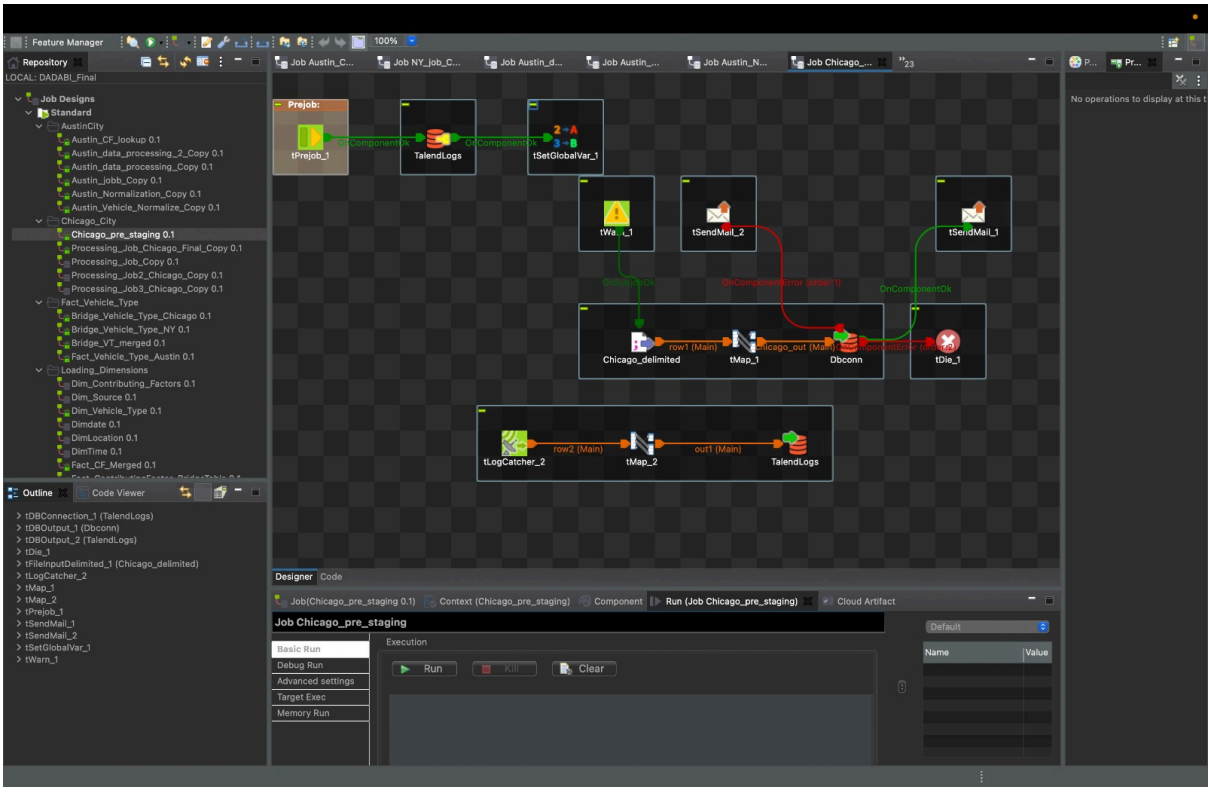
Accident Frequency: CRASH_RECORD_ID and CRASH_DATE provide the foundation for determining the total number of accidents across all cities.
Geographic Analysis: LATITUDE, LONGITUDE, and STREET_NAME identify which areas in the city have the highest number of accidents.
Injury and Fatality Reporting: INJURIES_FATAL and INJURIES_TOTAL contribute to the report that details the number of accidents resulting in injuries, both at an overall level and within Chicago.
Seasonality and Timing: CRASH_DATE allows us to analyze when accidents are most likely to occur, thus enabling the creation of a seasonality report.
Contributory Factors Analysis: PRIM_CONTRIBUTORY_CAUSE and SEC_CONTRIBUTORY_CAUSE aid in understanding the most common factors involved in accidents.

Chicago Jobs:

Dataset	Job Number	Job Name	Source Table	Target Table	Columns	Transformatio

					Transformed	n
Chicago	1	Chicago_pre_staging	Chicago_delimited	Chicago_pre_staging	-	
	2	Processing_Job_Copy	Chicago_pre_staging	Chicago_p_part1	CRASH_DATE, STREET_NAME,LATITUDE,LONGITUDE,CITY,WEATHER CONDITION	Appropriate format for crash_date assigned; handled nulls for street name, latitude, longitude, weather confition ;added city hardcoded for source info
	3	Processing_Job_Copy2_Chicago_Copy	Chicago_p_part1	Chicago_p_part2	Injuries_Non_Incapacitating,Injuries_Incapacitating,Injuries Total,Number_of_Pedestrians_Injured, Numbe_of_Pedestrians_killed, Numbe_of_Motorist_Injured, Numbe_of_Motorist_killed, Numbe_of_Cyclist_Injured, Numbe_of_Cyclist_Killed, Micromobility_death_count, VehicleType, Crash_Date, VehicleType added	Handled nulls for Vehicle Type, Injuries Total, Injuries_Non_Incapacitating,Injuries_Incapacitating, Injuries_Unknown, Injuries_Reported_Not_evident, Injuries_fatal;Hardcoded values for Numbe_of_Pedestrians_Injured, Numbe_of_Pedestrians_killed, Numbe_of_Motorist_Injured, Numbe_of_Motorist_killed, Numbe_of_Cyclist_Injured, Numbe_of_Cyclist_Killed, Micromobility_death_count as 0 , CRASH_DATE, added Vehicle_Type as "Others/Unknown"
	4	Processing_Job_Copy3_Chicago_Copy	Chicago_p_part2	Chicago_p_part3	PRIM_CONTRIBUTORY_CAUSE, SEC_CONTRIBUTORY_CAUSE	Handled null values for PRIM_CONTRIBUTORY_CAUSE, SEC_CONTRIBUTORY_CAUSE, latitude, longitude; Crash date

	5	Procesing_Job_Chicago_Final_Copy	Chicago_p_part3, Lookup_file_Chicago	Chicago_p_final	Code	1: Added column Codevia lookup on Contributing_Factor, 2: Contibuting_Factor_Description (assigned others to all null values)



Chicago Job 1:

Initialization (Pre-Job)

Component: tPrejob_1
Function: Executes initial setup operations necessary before the main ETL process begins. This includes establishing global variable settings for inputting appropriate tFileInputDelimited_1 according to current date and ensuring all system prerequisites are met.

Warning Generator

Component: tWarn_1
Function: Configured to generate warnings that might arise during the pre-job phase. These warnings help in monitoring the job's execution without aborting the process for non-critical issues. They are caught by the tLogCatcher_1

Data Extraction

Component: Chicago_delimited (tFileInputDelimited_1)
Function: Reads daily accident data files from Chicago with a dynamic file path that incorporates the date variable. This ensures that the correct file is processed according to the scheduled job run.

File Name/Stream

Configuration: The file path includes a global variable configured to have the date component to identify the correct input file, suggesting the handling of time-sequenced data files.

Data Transformation

Components: tMap_1 and tMap_2
Function: These components perform pre-staging data. They map input fields to their respective output fields necessary for subsequent processing steps.

Logging

Component: tLogCatcher_1
Function: Captures log messages, including errors and warnings generated by other components. This component centralized logging, thereby simplifying error handling and debugging.

Notification

Components: tSendMail_1 and tSendMail_2
Function: These components are configured to send out email notifications post-job execution, which include completion confirmations, error reports, or data quality summaries to designated stakeholders.

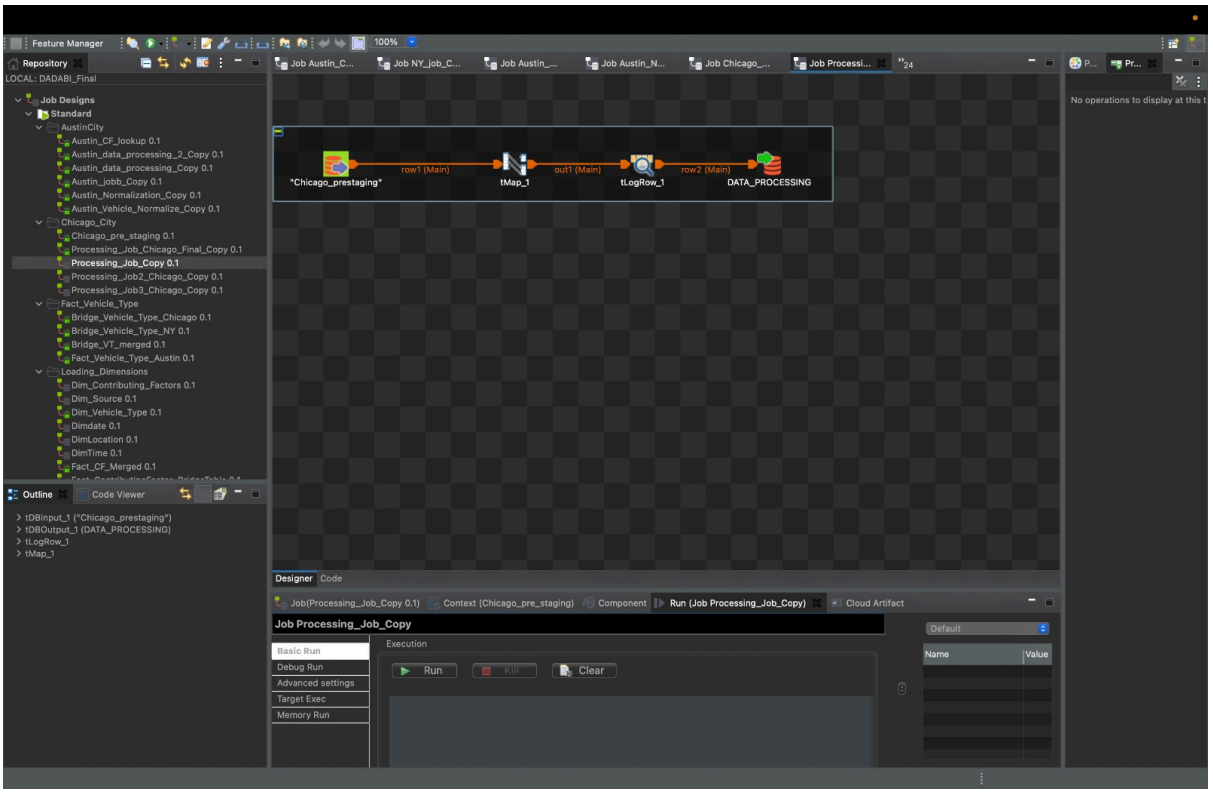
Variable Setting

Component: tSetGlobalVar_1
Function: Sets a global variable used across the job, to pass parameters or conditions between components during execution.

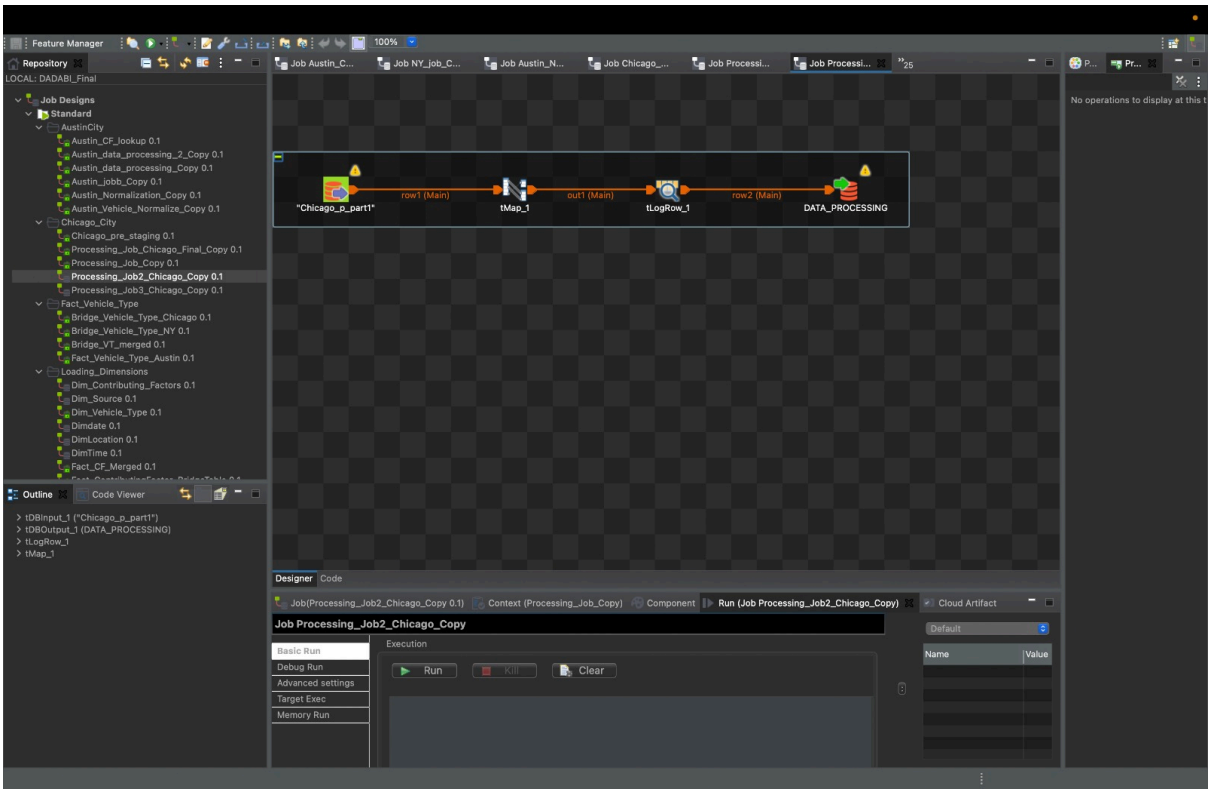
Clean-Up (Post-Job)

Component: tPostjob_1
Function: Executes after the pre-job to perform tasks, such as mapping connections, sending final notifications, etc

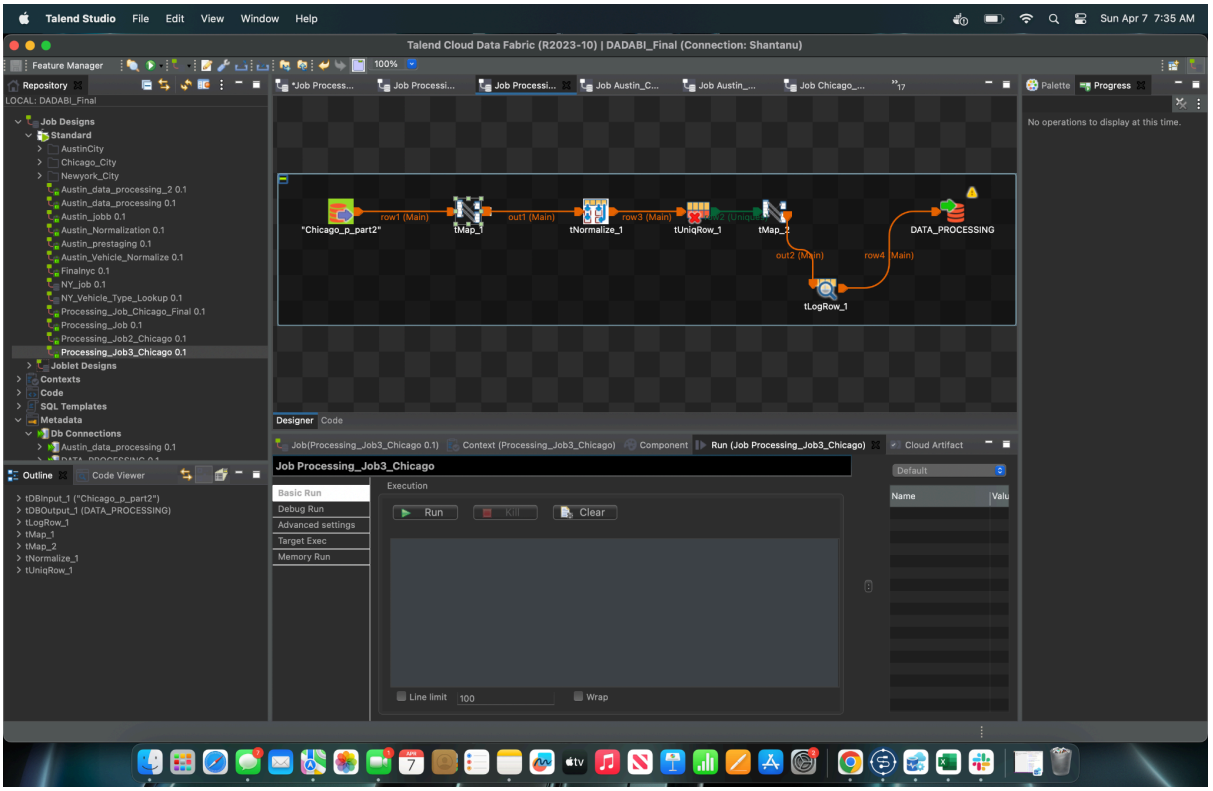
Chicago Job 2:



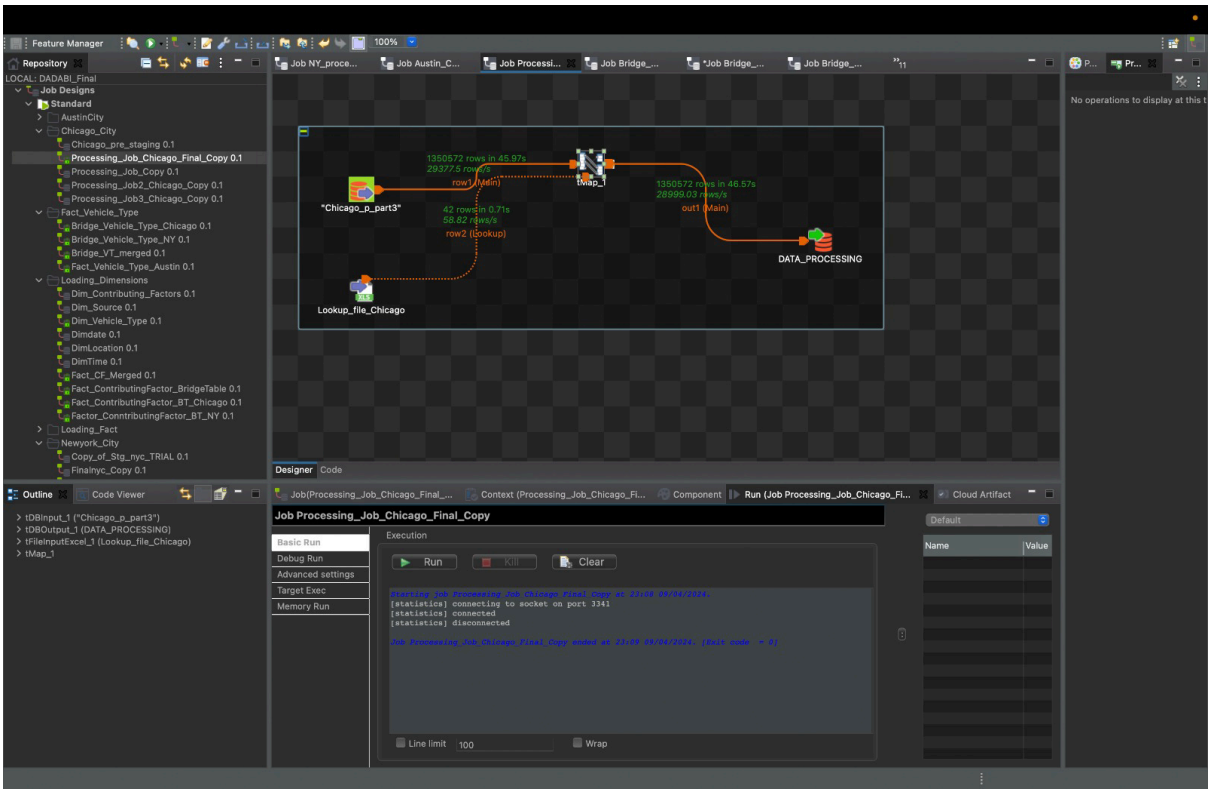
Chicago Job 3:



Chicago Job 4 :



Chicago Job 5:



NYC:
Introduction

This report justifies the selection of specific data columns from the New York City (NYC) traffic accident dataset. These columns are vital for generating insights into the traffic accident trends within the city, with the ultimate objective of improving road safety and informing urban traffic management strategies.

Data Column Selection Justification

The selection of NYC data columns is based on their relevance to the critical business questions related to the occurrence, location, severity, and causative factors of traffic accidents.

COLLISION_ID

- As the unique identifier for each accident event in the raw dataset, it allows for an accurate count of incidents and the ability to reference each accident distinctly from the raw dataset.

CRASH DATE & CRASH TIME

- These columns provide essential temporal data, enabling the analysis of accident frequency over time, identifying peak times, and understanding seasonal trends in accident occurrence.

LATITUDE & LONGITUDE

- Geospatial data is indispensable for pinpointing the exact location of accidents. It is crucial for mapping accident hotspots and conducting location-based trend analyses.

"NY" hardcoded for City

- Including the city's name as a constant value ensures clarity when aggregating data from multiple cities and allows for straightforward city-level analyses.

Derived STREET NAME from ON STREET NAME

- Street name data is critical for localizing accidents to specific streets, which is essential for identifying high-risk roads and intersections.

NUMBER OF PERSONS INJURED & KILLED

- These figures provide all human impact of traffic accidents aggregated in that crash, informing emergency response and healthcare provisioning.

NUMBER OF PEDESTRIANS, CYCLISTS, AND MOTORISTS INJURED & KILLED

- Disaggregating data by road user type is vital for targeted safety measures and understanding which groups are most at risk in traffic incidents.

CONTRIBUTING FACTOR VEHICLE 1-5

- Identifying the contributing factors to accidents is crucial for developing preventive measures and policy interventions. These columns can reveal patterns and common causes of accidents.

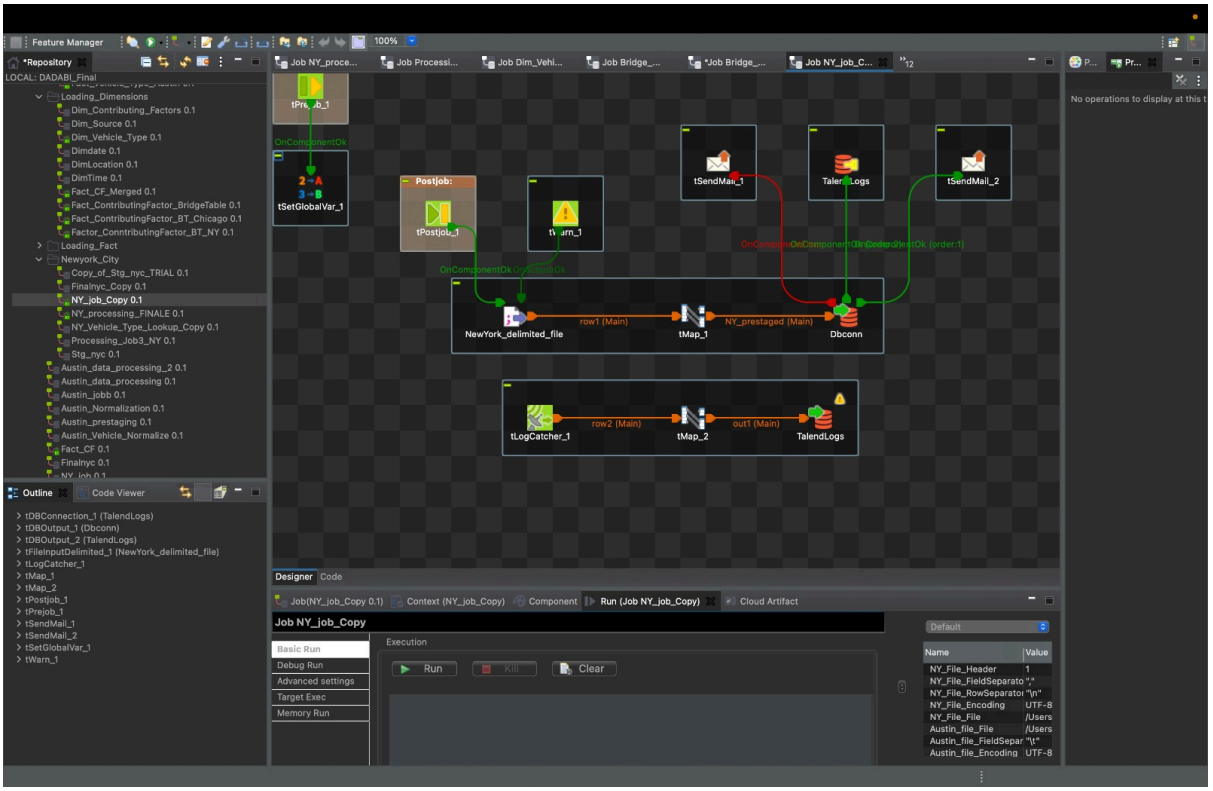
Business Questions Addressed

Accident Frequency: The COLLISION_ID along with CRASH DATE and CRASH TIME enable analysis of the total number of accidents over various time frames.
Accident Localization: LATITUDE, LONGITUDE, and the derived STREET NAME enable identification of the most accident-prone areas in NYC.
Severity and Demographics: Data on the number of injuries and fatalities, including the breakdown by road user type, help measure the impact of accidents on public health.
Causal Analysis: CONTRIBUTING FACTOR VEHICLE columns help identify prevalent contributing factors, which are essential for crafting preventative strategies.

NewYork City Jobs:

Dataset	Job Number	Job Name	Source Table	Target Table	Columns Transformed	Transformation
NY	1	NY_Job_Copy	NY_job_Copy (TSV file)	NY_pre_staging	-	
	2	Copy_of_Stg_nyc	NY_pre_staging	NY_p_part1	CRASH DATE, LATITUDE, LONGITUDE, ON_STREET_NAME,NUMBER_OF_PERSONS_INJURED, NUMBER_OF_PERSONS_KILLED, NUMBER_OF_PEDESTRIANS_INJURED, NUMBER_OF_PEDESTRIANS_KILLED, NUMBER_OF_CYCLISTS_INJURED, NUMBER_OF_CYCLISTS_KILLED, NUMBER_OF_CYCLISTS_INJURED, NUMBER_OF_MOTORISTS_KILLED, CONTRIBUTING_FACTORS(1-5), city added\	Check for digits, handled nulls added relevant columns to maintain source info
	3	Finalnyc_Copy	NY_p_part1	NY_p_part2	VEHICLE_TYPE_CODE(1-5)	Merged Vehicle_Type_Codes
	4	Processing_Job3_NY	NY_p_part2, Custom LookUp file for Contirbuting Cause	NY_p_part3	Code, Contibuting_Cause	1: Added column Code via lookup , 2: Contibuting_Cause (assigned others to all values 101)
	5	NY_Vehicle_Type_Lookup	NY_p_part3, Custom LookUp_Vehicle_Type_NY	NY_p_part4	Vehicle_Type	1: Generalized column Vehicle_Type via custom lookup file made wrt Austin
	6	NY_Processing_Finale	NY_p_part4	NY_p_final	Vehicle_Type_C Category, Crash_Hour	Added the 2 new columns referencing Crash_Time

New York Job 1:



Initialization (Pre-Job)

Component: tPrejob_1
Function: Executes initial setup operations necessary before the main ETL process begins. This includes establishing global variable settings for inputting appropriate tFileInputDelimited_1 according to current date and ensuring all system prerequisites are met.

Warning Generator

Component: tWarn_1
Function: Configured to generate warnings that might arise during the pre-job phase. These warnings help in monitoring the job's execution without aborting the process for non-critical issues. They are caught by the tLogCatcher_1

Data Extraction

Component: NewYork_delimited (tFileInputDelimited_1)
Function: Reads daily accident data files from New York with a dynamic file path that incorporates the date variable. This ensures that the correct file is processed according to the scheduled job run.

File Name/Stream

Configuration: The file path includes a global variable configured to have the date component to identify the correct input file, suggesting the handling of time-sequenced data files.

Data Transformation

Components: tMap_1 and tMap_2
Function: These components perform pre-staging data. They map input fields to their respective output fields necessary for subsequent processing steps.

Logging

Component: tLogCatcher_1
Function: Captures log messages, including errors and warnings generated by other components. This component centralized logging, thereby simplifying error handling and debugging.

Notification

Components: tSendMail_1 and tSendMail_2
Function: These components are configured to send out email notifications post-job execution, which include completion confirmations, error reports, or data quality summaries to designated stakeholders.

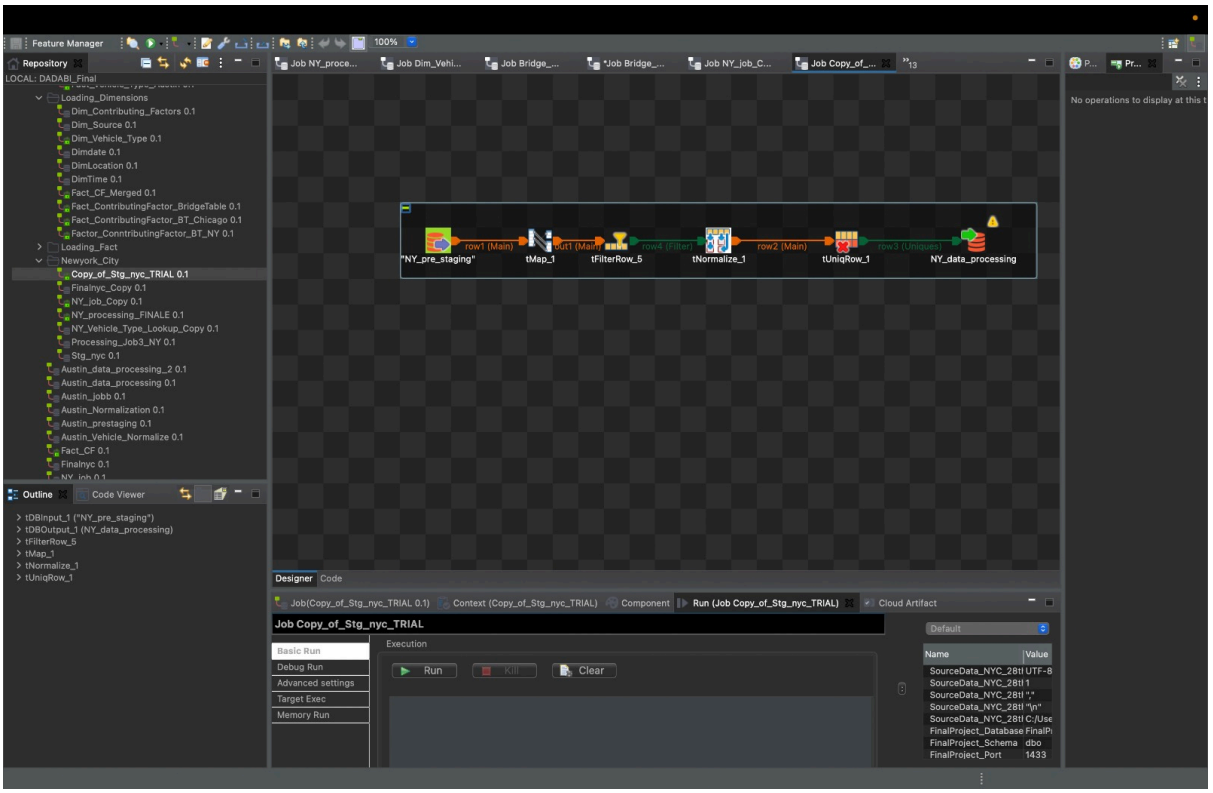
Variable Setting

Component: tSetGlobalVar_1
Function: Sets a global variable used across the job, to pass parameters or conditions between components during execution.

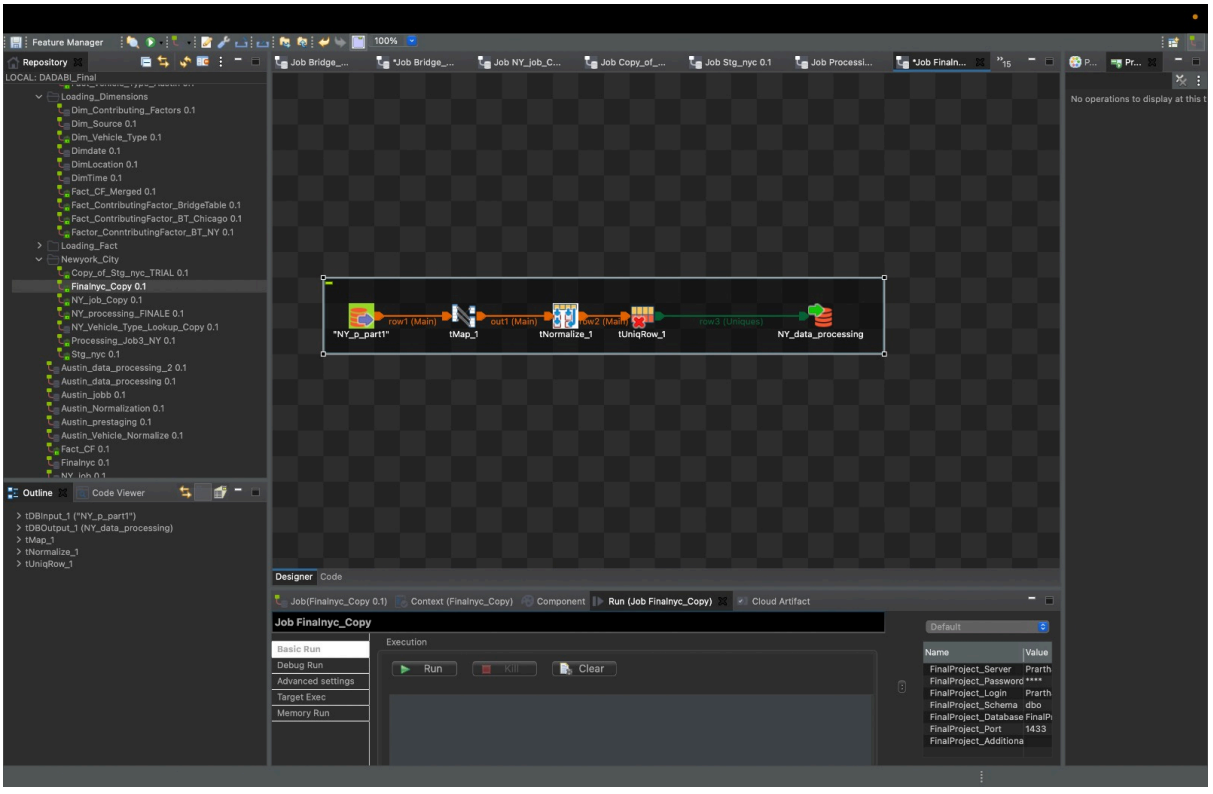
Clean-Up (Post-Job)

Component: tPostjob_1
Function: Executes after the pre-job to perform tasks, such as mapping connections, sending final notifications, etc

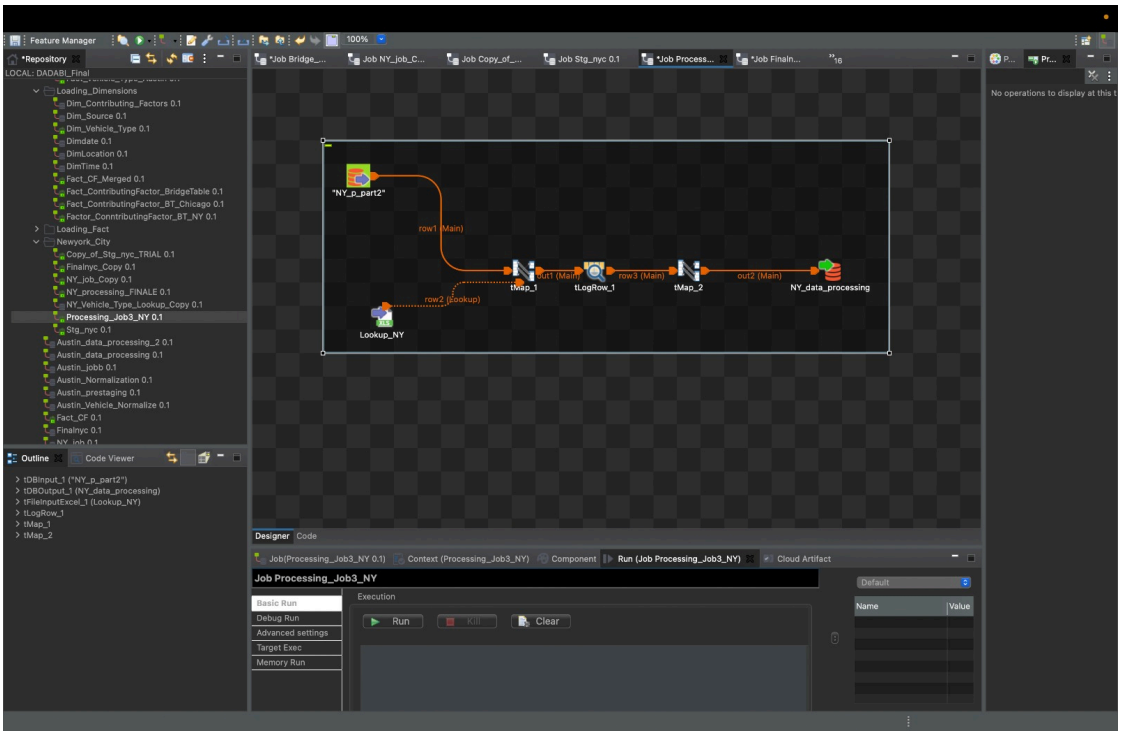
New York Job 2:



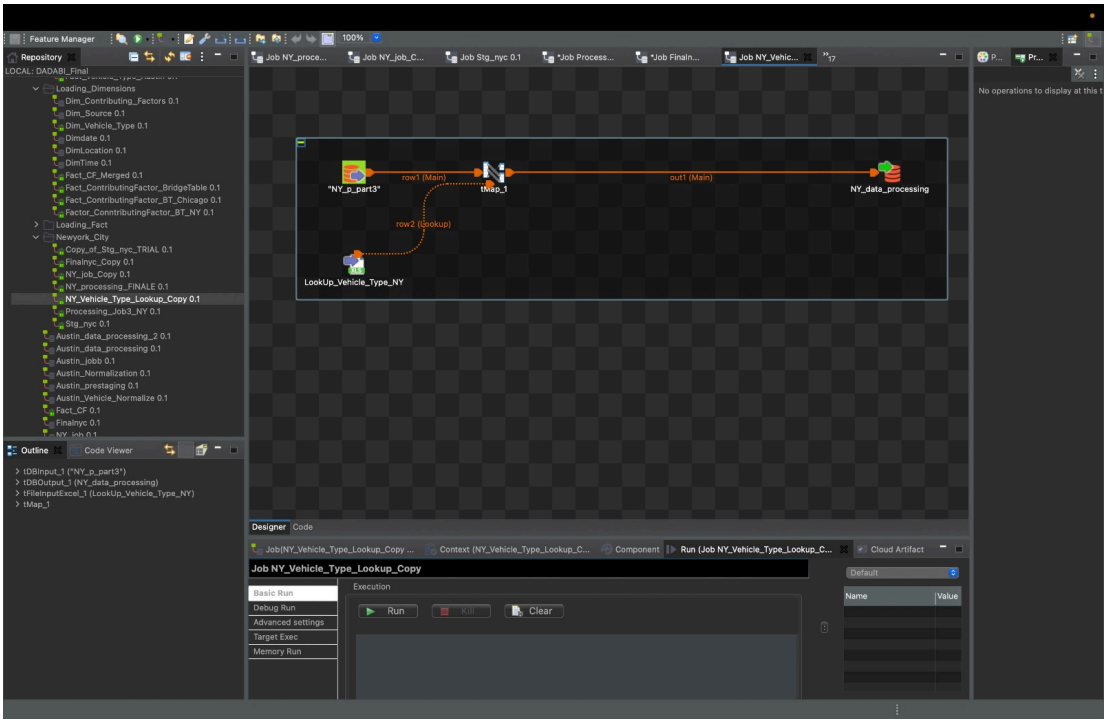
New York Job 3:



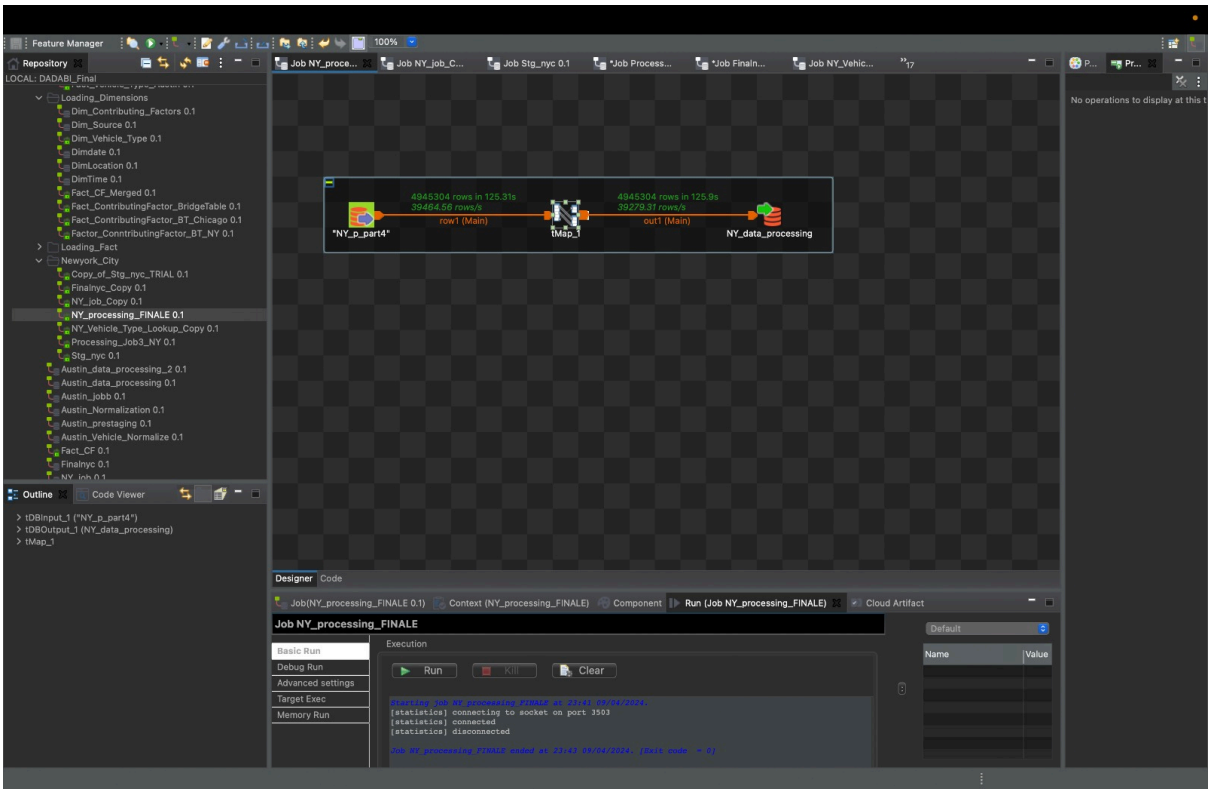
New York Job 4:



New York Job 5:



New York Job 6:



Dimensions:

- Dim_ContributingFactors
- Dim_VehicleType
- Dim_Location
- Dim_Date
- Dim_Time

Fact Tables:

- FactContributingFactor:

- BridgeVehicleAccident:

- AccidentEventsFactTable:

DDL Scripts:

-- DDL for Dim_ContributingFactors

```
CREATE TABLE Dim_ContributingFactors (
    Contributing_Factor_SK INT PRIMARY KEY,
    Durable_Key INT,
    Contributing_Factor_Code VARCHAR(255),
    Contributing_Factor_Description TEXT,
    scd_start DATE,
    scd_end DATE,
    scd_version INT,
    scd_active BIT,
    DI_Create_Date DATE,
    DI_Job_ID INT,
    DI_Workflow_FileName VARCHAR(255)
);
```

-- DDL for Dim_VehicleType

```
CREATE TABLE Dim_VehicleType (
    Vehicle_Type_SK INT PRIMARY KEY,
    Vehicle_Type VARCHAR(255),
    DI_Create_Date DATE,
    DI_Job_ID INT,
    DI_Workflow_FileName VARCHAR(255)
);
```

-- DDL for Dim_Location

```
CREATE TABLE Dim_Location (
    LocationKey INT PRIMARY KEY,
    StreetName VARCHAR(255),
    Latitude DECIMAL(9,6),
    Longitude DECIMAL(9,6),
    DI_Create_Date DATE,
    DI_Job_ID INT,
    DI_Workflow_FileName VARCHAR(255)
);
```

-- DDL for Dim_Date

```
CREATE TABLE Dim_Date (
    DateSK INT PRIMARY KEY,
    Date DATE,
    Day_of_Week VARCHAR(9),
    Month VARCHAR(9),
    Year INT,
    Season VARCHAR(255),
    DI_Create_Date DATE,
    DI_Job_ID INT,
    DI_Workflow_FileName VARCHAR(255)
);
```

-- DDL for Dim_Time

```
CREATE TABLE Dim_Time (
    TimeKey INT PRIMARY KEY,
```

```
    Hour_24 INT,
    Time_Period VARCHAR(255),
    DI_Create_Date DATE,
    DI_Job_ID INT,
    DI_Workflow_FileName VARCHAR(255)
);

-- DDL for FactContributingFactor
CREATE TABLE FactContributingFactor (
    BridgeContributingFactorKey INT PRIMARY KEY,
    Contributing_Factor_SK INT,
    FOREIGN KEY (Contributing_Factor_SK) REFERENCES
Dim_ContributingFactors(Contributing_Factor_SK)
);

-- DDL for BridgeVehicleAccident
CREATE TABLE BridgeVehicleAccident (
    TVBridgeSK INT PRIMARY KEY,
    Vehicle_Type_SK INT,
    Crash_SK INT,
    FOREIGN KEY (Vehicle_Type_SK) REFERENCES Dim_VehicleType(Vehicle_Type_SK),
    FOREIGN KEY (Crash_SK) REFERENCES AccidentEventsFactTable(Crash_SK)
);

-- DDL for AccidentEventsFactTable
```

```
CREATE TABLE AccidentEventsFactTable (
    Crash_SK INT PRIMARY KEY,
    Crash_ID VARCHAR(255),
    Total_Number_of_Injuries INT,
    Total_Number_of_Fatalities INT,
    Pedestrian_Involved_fl BIT,
    No_of_Pedestrians_Injured INT,
    No_of_Pedestrians_Killed INT,
    No_of_Motorists_Injured INT,
    No_of_Motorists_Killed INT,
    No_of_Cyclists_Injured INT,
    No_of_Cyclists_Killed INT,
    DateSK INT,
    LocationKey INT,
    TimeKey INT,
    FOREIGN KEY (DateSK) REFERENCES Dim_Date(DateSK),
    FOREIGN KEY (LocationKey) REFERENCES Dim_Location(LocationKey),
    FOREIGN KEY (TimeKey) REFERENCES Dim_Time(TimeKey));
```

CONCLUSION

As we conclude our project on Motor Vehicle Collisions and Crashes, we reflect on the comprehensive journey we've embarked upon. Our venture has spanned across three major cities: New York, Chicago, and Austin, each presenting unique challenges and insights in analyzing traffic crash data sourced from their respective Departments of Transportation.

Through meticulous data profiling, staging, and the integration of ETL jobs using Talend, we have adhered to the highest standards of data handling, ensuring accuracy and reliability in

our findings. Our dimensional modeling has enabled us to create meaningful relationships within the data, encapsulating facts and dimensions that provide a multi-faceted view of the underlying trends.

We addressed null values with care, ensuring that the integrity of our analysis remained uncompromised. By implementing slowly changing dimensions (SCD2) within at least one aspect of our model, we maintained a robust link to the historical data, capturing the evolution of the data points over time.

In conclusion, this project has not only equipped us with valuable insights into motor vehicle collisions but also reinforced our capabilities in managing big data, ensuring data quality, and deriving actionable intelligence from complex datasets. We believe that the methodologies and practices we have employed can serve as a blueprint for future analyses within this domain.