**Data Profiling:** 

We started by listing down all the column, it’s data types and description for all the 3 staging tables (Persons, Crash and Vehicles) to look for missing values, nulls and any data type inconsistencies etc.

**Staging tables:**   
**Persons Table:**   
**String/Character Fields**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Datatype** | **% Missing** | **Unique Values** | **Shortest Value** | **Longest Value** | **Min Value Count** | **Max Value Count** | **Remarks** |
| CRASH\_TIME | varchar(20)    (change to Time in Dim) | 0.0% | 1,440 | 9:43 | 15:15    (24 hr time format) | 215 | 72,760 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| CONTRIBUTING\_FACTOR\_2 | Varchar(255)    Null: No Value Provided | 98.5% | 51 | Glare | Pedestrian/Bicyclist/Other Pedestrian Error/Confusion | 1 | 4,920,623 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| SAFETY\_EQUIPMENT | Varchar(255)    Null: No Value Provided | 48.8% | 19 | - | Stoppers Only (In-Line Skater/Bicyclist) | 2 | 2,437,109 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| PERSON\_ID | Varchar(80) | 0.0% | 4,800,319 | 1 | 31aa2bc0-f545-444f-8cdb-f1cb5cf00b89 | 1 | 142,787 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| PED\_ROLE | Varchar(255)    Null: No Value Provided | 3.9% | 11 | Owner | Notified Person | 305 | 2,070,186 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| CRASH\_DATE | Date | 0.0% | 3,934 | 10/26/2019 | 10/26/2019 | 21 | 4,220 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| PERSON\_TYPE | Varchar(80) | 0.0% | 4 | Occupant | Other Motorized | 6,215 | 4,811,588 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| POSITION\_IN\_VEHICLE | Varchar(255)    Null: No Value Provided | 48.8% | 12 | Driver | Any person in the rear of a station wagon, pick-up truck, all passengers on a bus, etc | 817 | 2,437,038 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| PED\_LOCATION | Varchar(255)    Null: No Value Provided | 98.5% | 5 | Unknown | Pedestrian/Bicyclist/Other Pedestrian Not at Intersection | 2,186 | 4,919,237 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| PED\_ACTION | Varchar(255)    Null: No Value Provided | 98.5% | 17 | Unknown | Emerging from in Front of/Behind Parked Vehicle | 63 | 4,919,338 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| PERSON\_INJURY | Varchar(80) | 0.0% | 3 | Killed | Unspecified | 2,816 | 4,402,459 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| EJECTION | Varchar(80) | 48.8% | 7 | Ejected | Partially Ejected | 541 | 2,510,988 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| PERSON\_SEX | Varchar(20)    Null: No Value Provided | 11.4% | 4 | U | U | 410,739 | 2,669,953 | This field has over 10% missing values. Consider imputing these values. |
| EMOTIONAL\_STATUS | Varchar(80)    Null: No Value Provided | 47.4% | 9 | Shock | Does Not Apply | 1,592 | 2,365,908 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| BODILY\_INJURY | Varchar(80)    Null: No Value Provided | 47.4% | 15 | Eye | Shoulder - Upper Arm | 765 | 2,365,865 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| COMPLAINT | Varchar(255)    Null: No Value Provided | 47.4% | 22 | Unknown | Fracture - Distorted - Dislocation | 26 | 2,365,858 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| CONTRIBUTING\_FACTOR\_1 | Varchar(255)    Null: No Value Provided | 98.5% | 54 | Glare | Pedestrian/Bicyclist/Other Pedestrian Error/Confusion | 1 | 4,920,523 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| DI\_PID | Integer | 0.0% |  |  |  |  |  | Auto generated using Formula in Talend (hence integer) |
| DI\_Create\_Date | Date |  |  |  |  |  |  |  |

**Numeric Fields** 

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** |  | **% Missing** | **Unique Values** | **Min** | **Mean** | **Median** | **Max** | **Std Dev** | **Remarks** |
| PERSON\_AGE | Integer    Null: -99 | 10.2% | 869 | -999.000 | 37.058 | 68.000 | 9,999.000 | 117.242 | This field has over 10% missing values. Consider imputing these values. |
| VEHICLE\_ID | Varchar(80)    Null: No value Provided | 4.0% | 2,304,434 | 123,423.000 | 18,378,373.033 | 18,603,597.500 | 20,419,746.000 | 1,516,560.745 |  |
| UNIQUE\_ID | BigInt | 0.0% | 4,995,141 | 10,922.000 | 8,766,994.564 | 9,149,930.000 | 12,571,292.000 | 2,504,216.834 |  |
| COLLISION\_ID | BigInt | 0.0% | 1,353,929 | 37.000 | 3,897,161.821 | 3,958,437.000 | 4,619,723.000 | 643,597.133 |  |

**Vehicle Table:**

**String/Character Fields**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Datatype** | | **% Missing** | | **Unique Values** | | **Shortest Value** | | **Longest Value** | | **Min Value Count** | | **Max Value Count** | | **Remarks** | | |
| VEHICLE\_ID | varchar(80) | | 0.0% | | 2,175,578 | | 1 | | 0553ab4d-9500-4cba-8d98-f4d7f89d5856 | | 1 | | 769,063 | | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| CRASH\_TIME | varchar(20)    (change to Time in Dim) | | 0.0% | | 1,440 | | 9:03 | | 17:18 | | 140 | | 53,106 | | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| CONTRIBUTING\_FACTOR\_2 | Varchar(255)    Null: No Value Provided | | 44.6% | | 57 | | 1 | | Pedestrian/Bicyclist/Other Pedestrian Error/Confusion | | 21 | | 1,650,688 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| PUBLIC\_PROPERTY\_DAMAGE | Varchar(1024)    Null: No Value Provided | | 41.3% | | 4 | | N | | Unspecified | | 11,168 | | 1,923,166 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| VEHICLE\_MODEL | Varchar(80)  Null: No Value Provided | | 98.6% | | 2,433 | | 6 | | school bus (medium length | | 1 | | 3,652,982 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| CRASH\_DATE | Date(80) | | 0.0% | | 3,444 | | 09/07/2012 | | 09/07/2012 | | 181 | | 2,375 | | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| VEHICLE\_TYPE | Varchar(80)  Null: No Value Provided | | 4.8% | | 2,146 | | A | | Enclosed Body - Nonremovable Enclosure | | 1 | | 859,697 | | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| VEHICLE\_DAMAGE\_3 | Varchar(80)  Null: No Value Provided | | 80.3% | | 20 | | Roof | | Right Front Quarter Panel | | 107 | | 2,976,150 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| PRE\_CRASH | Varchar(255)  Null: No Value Provided | | 23.8% | | 20 | | Parked | | Avoiding Object in Roadway | | 744 | | 1,369,802 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| VEHICLE\_DAMAGE\_1 | Varchar(255)  Null: No Value Provided | | 65.2% | | 20 | | Roof | | Right Front Quarter Panel | | 291 | | 2,416,919 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| PUBLIC\_PROPERTY\_DAMAGE\_TYPE | Varchar(1024)    Null: No Value Provided | | 99.5% | | 14,597 | | . | | AT T/P/O MV#1 STATES SHE WAS DRIVING EAST BOUND ON 115 AVENUE, SHE FURTHER STATES THAT MV#2 DID NOT STOP AT THE STOP SIGN AT 115 AVENUE AND 125 STREET CAUSING MV#1 TO IMPACT MV#2 CAUSING MV#2 TO STRIKE A CHAIN LINK FENCE. MV#2 STATES THAT HE WAS DRIVING SOUTH BOUND ON 125 STREET STOPPED ATTHE STOP SIGN THE PROCEEDED TO MAKE A LEFT HAND TURN WHEN MV#1 IMPACTED HIS VEHICLE CAUSING MV#2 TO STRIKE A CHAIN LINK FENCE AT 125-04 115 AVENUE. FRONT PASSENGER FROM MV#1 REMOVED TO JAMACIA HOSPITAL. FRONT P | | 1 | | 3,685,058 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| POINT\_OF\_IMPACT | Varchar(255)    Null: No Value Provided | | 44.9% | | 20 | | Roof | | Right Front Quarter Panel | | 1,091 | | 1,662,017 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| TRAVEL\_DIRECTION | Varchar(255)    Null: No Value Provided | | 44.1% | | 16 | | - | | Southwest | | 28 | | 1,634,411 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| VEHICLE\_YEAR | Varchar(80)  Null: No Value Provided | | 48.5% | | 300 | | 2002 | | 20063 | | 1 | | 1,796,971 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| VEHICLE\_DAMAGE | Varchar(255)    Null: No Value Provided | | 45.3% | | 20 | | Roof | | Right Front Quarter Panel | | 1,817 | | 1,679,039 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| DRIVER\_LICENSE\_STATUS | Varchar(255)    Null: No Value Provided | | 57.2% | | 4 | | Permit | | Unlicensed | | 11,830 | | 2,118,009 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| VEHICLE\_DAMAGE\_2 | Varchar(255)    Null: No Value Provided | | 74.1% | | 20 | | Roof | | Right Front Quarter Panel | | 119 | | 2,746,520 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| VEHICLE\_MAKE | Varchar(80)  Null: No Value Provided | | 48.2% | | 10,272 | | M | | GRUMMAN LLV OF THE UNITED STATES POSTAL SERVICE | | 1 | | 1,785,296 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| DRIVER\_SEX | Varchar(80)  Null: No Value Provided | | 55.3% | | 4 | | M | | M | | 7,109 | | 2,050,286 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| CONTRIBUTING\_FACTOR\_1 | Varchar(255)    Null: No Value Provided | | 3.2% | | 62 | | 1 | | Pedestrian/Bicyclist/Other Pedestrian Error/Confusion | | 16 | | 2,131,906 | | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| DRIVER\_LICENSE\_JURISDICTION | Varchar(255)    Null: No Value Provided | | 57.0% | | 71 | | NY | | A,NEJADE | | 1 | | 2,110,256 | | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| STATE\_REGISTRATION | Varchar(80)  Null: No Value Provided | | 5.9% | | 83 | | T | | NY | | 1 | | 2,923,340 | | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | |
| DI\_PID | | Integer | | 0.0% | |  | |  | |  | |  | |  | | Auto generated using Formula in Talend (hence integer) |  |
| DI\_Create\_Date | | Date | |  | |  | |  | |  | |  | |  | |  |  |

**Numeric Fields**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Datatype** | **% Missing** | **Unique Values** | **Min** | **Mean** | **Median** | **Max** | **Std Dev** | **Remarks** |
| VEHICLE\_OCCUPANTS | Integer    Null: -99 | 46.4% | 121 | 0.000 | 1,010.392 | 2.000 | 999,999,999.000 | 994,574.152 | This field has over 10% missing values. Consider imputing these values. |
| UNIQUE\_ID | BigInt | 0.0% | 3,704,406 | 111,711.000 | 16,060,871.002 | 17,306,057.500 | 20,121,717.000 | 3,244,903.357 |  |
| COLLISION\_ID | BigInt | 0.0% | 1,845,435 | 22.000 | 2,996,659.054 | 3,567,067.500 | 4,484,197.000 | 1,495,340.474 |  |

**BigQuery Table (Crash dataset):**

**String/Character Fields** 

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | | **Datatype** | | **% Missing** | **Unique Values** | | **Shortest Value** | **Longest Value** | | **Min Value Count** | | **Max Value Count** | **Remarks** | |
| contributing\_factor\_vehicle\_4 | |  | | 98.4% | 41 | | Glare | Traffic Control Device Improper/Non-Working | | 1 | | 1,951,965 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| vehicle\_type\_code\_5 | |  | | 99.6% | 65 | | PK | Station Wagon/Sport Utility Vehicle | | 1 | | 1,975,056 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| borough | |  | | 31.1% | 6 | | BRONX | STATEN ISLAND | | 57,317 | | 616,697 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| contributing\_factor\_vehicle\_3 | |  | | 93.0% | 52 | | 1 | Pedestrian/Bicyclist/Other Pedestrian Error/Confusion | | 1 | | 1,843,420 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| vehicle\_type\_code1 | |  | | 0.6% | 1,500 | | 1 | Enclosed Body - Nonremovable Enclosure | | 1 | | 534,747 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| contributing\_factor\_vehicle\_1 | |  | | 0.3% | 62 | | 1 | Pedestrian/Bicyclist/Other Pedestrian Error/Confusion | | 10 | | 684,508 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| vehicle\_type\_code\_3 | |  | | 93.2% | 236 | | PK | Station Wagon/Sport Utility Vehicle | | 1 | | 1,848,167 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| contributing\_factor\_vehicle\_5 | |  | | 99.6% | 30 | | Glare | Traffic Control Device Improper/Non-Working | | 1 | | 1,974,808 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| vehicle\_type\_code2 | |  | | 18.4% | 1,669 | | D | Enclosed Body - Nonremovable Enclosure | | 1 | | 379,493 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| contributing\_factor\_vehicle\_2 | |  | | 15.1% | 62 | | 1 | Pedestrian/Bicyclist/Other Pedestrian Error/Confusion | | 3 | | 1,416,930 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| cross\_street\_name | |  | | 37.2% | 19,891 | | 0 | EAGLE STREET | | 1 | | 737,221 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| location | |  | | 11.4% | 268,287 | | (0.0, 0.0) | (40.5258882, -74.2298047) | | 1 | | 226,896 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| on\_street\_name | |  | | 20.9% | 17,718 | | 72 | BROOKLYN QUEENS EXPRESSWAY RAMP | | 1 | | 414,857 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| off\_street\_name | |  | | 83.7% | 208,514 | | AVENUE Y | ATLANTIC AVE | | 1 | | 1,660,745 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
| vehicle\_type\_code\_4 | |  | | 98.5% | 94 | | PK | Station Wagon/Sport Utility Vehicle | | 1 | | 1,952,992 | This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | |
|  | |  | |  |  | |  |  | |  | |  |  | |
| **Name** | **Datatype** | | **% Missing** | | | **Unique Values** | | | **Latest Date** | | **Earliest Date** | | **Interval** | **Remarks** |
| timestamp | time | | 0.0% | | | 1,070,284 | | | 04/09/2023 23:00 | | 07/01/2012 00:00 | | Unknown |  |

**Numeric Fields** 

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Datatype** | **% Missing** | **Unique Values** | **Min** | **Mean** | **Median** | **Max** | **Std Dev** | **Remarks** |
| number\_of\_persons\_killed |  | 0.0% | 7 | 0.000 | 0.001 | 0.000 | 8.000 | 0.040 | This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string". |
| latitude |  | 11.4% | 125,147 | 0.000 | 40.628 | 40.722 | 43.344 | 1.978 | This field has over 10% missing values. Consider imputing these values. |
| number\_of\_cyclist\_injured |  | 0.0% | 5 | 0.000 | 0.026 | 0.000 | 4.000 | 0.160 | This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string". |
| longitude |  | 11.4% | 97,487 | -201.360 | -73.753 | -73.926 | 0.000 | 3.726 | This field has over 10% missing values. Consider imputing these values. |
| zip\_code |  | 31.1% | 233 | 10,000.000 | 10,839.379 | 11,207.000 | 11,697.000 | 554.774 | This field has over 10% missing values. Consider imputing these values. |
| number\_of\_motorist\_killed |  | 0.0% | 6 | 0.000 | 0.001 | 0.000 | 5.000 | 0.026 | This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string". |
| number\_of\_pedestrians\_killed |  | 0.0% | 4 | 0.000 | 0.001 | 0.000 | 6.000 | 0.027 | This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string". |
| number\_of\_cyclist\_killed |  | 0.0% | 3 | 0.000 | 0.000 | 0.000 | 2.000 | 0.010 | This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string". |
| number\_of\_pedestrians\_injured |  | 0.0% | 14 | 0.000 | 0.055 | 0.000 | 27.000 | 0.240 |  |
| number\_of\_persons\_injured |  | 0.0% | 31 | 0.000 | 0.298 | 1.000 | 43.000 | 0.689 |  |
| number\_of\_motorist\_injured |  | 0.0% | 30 | 0.000 | 0.215 | 0.000 | 43.000 | 0.650 |  |
| unique\_key |  | 0.0% | 1,983,231 | 22.000 | 3,089,604.018 | 3,627,826.000 | 4,619,988.000 | 1,503,457.266 |  |

**Data Profiling for Cleaning Data**

1. **Dim\_arrest\_borough**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| borough\_sk | Integer | Auto generated |  |
| Boro\_code | (char(1)) --> Integer | Nyc.gov--> boro\_code is an integer value  **A picture containing text, screenshot, font, algebra  Description automatically generated** |  |
| borough | Varchar(80) | Null Replace: “No Value Provided” | In All caps  Unique: |

1. **Dim\_BODILY\_INJURY**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| BODILY\_INJURY\_SK | Integer | Auto generated |  |
| BODILY\_INJURY | Varchar(80) | Blanks Replace: “No Value Provided”  *Contains Unknown and Does not Apply rows* | In Title Case  Unique: 15 |

1. **Dim\_COMPLAINT**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| COMPLAINT\_SK | Integer | Auto generated |  |
| COMPLAINT | Varchar(255) | Blanks Replace: “No Value Provided”  *Contains Unknown and Does not Apply rows* | In Title Case  Unique: 22 |

1. **Dim\_CONTRIBUTING\_FACTOR**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| CONTRIBUTING\_FACTOR\_SK | Integer | Auto generated |  |
| CONTRIBUTING\_FACTOR | Varchar(255) | **Persons:**  Blanks Replace: “No Value Provided”  *Does not Contain Unknown and Does not Apply rows*  **Vehicle:**  Blanks Replace: “No Value Provided”  Invalid Values: 1, 80  **Crash:**  Null Replace: “No Value Provided”  Invalid Values: 1, 80    *Contains “Unspecified” rows* | In Title Case  Unique:  61    (59 + 1 invalid + no value provided) |
|  |  |  |  |

1. **Dim\_DRIVER\_LICENSE\_JURISDICTION**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| DRIVER\_LICENSE\_JURISDICTION\_SK | Integer | Auto generated |  |
| DRIVER\_LICENSE\_JURISDICTION | Varchar(255) | Blanks Replace: “No Value Provided”    *Does not Contain Unknown and Does not Apply rows*    *There are only 50 states so rest we are replacing them with invalid:*    *Invalid values:*   *A,NEJADE*  *AB*  *BC*  *DC*  *ETA*  *MB*  *MX*  *NB*  *NB1*  *NF*  *NS*  *NT*  *ON*  *PA'*  *PE*  *PQ*  *PR*  *SK*  *YT*  *ZZZ* | In Title Case  Unique: 50 + No value Provided |

1. **Dim\_DRIVER\_LICENSE\_STATUS:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| DRIVER\_LICENSE\_STATUS\_SK | Integer | Auto generated |  |
| DRIVER\_LICENSE\_STATUS | Varchar(255) | Blanks Replace: “No Value Provided”    *Does not Contain Unknown and Does not Apply rows* | In Title Case  Unique: 4 |

1. **Dim\_EJECTION:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| EJECTION\_SK | Integer | Auto generated |  |
| EJECTION | Varchar(80) | Blanks Replace: “No Value Provided”    *Contains Unknown and Does not Apply rows* | In Title Case  Unique: 7 |

1. **Dim\_EMOTIONAL\_STATUS:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| EMOTIONAL\_STATUSSK | Integer | Auto generated |  |
| EMOTIONAL\_STATUS | Varchar(80) | Blanks Replace: “No Value Provided”    *Contains Unknown and Does not Apply rows* | In Title Case  Unique: 9 |

1. **Dim\_PED\_ACTION:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| PED\_ACTION\_SK | Integer | Auto generated |  |
| PED\_ACTION | Varchar(255) | Blanks Replace: “No Value Provided”    *Contains Unknown and Does not Apply rows* | In Title Case  Unique: 17 |

1. **Dim\_PED\_LOCATION:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| PED\_LOCATION\_SK | Integer | Auto generated |  |
| PED\_LOCATION | Varchar(255) | Blanks Replace: “No Value Provided”    *Contains Unknown and Does not Apply rows* | In Title Case  Unique: 5 |

1. **Dim\_PED\_ROLE:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| PED\_ROLE\_SK | Integer | Auto generated |  |
| PED\_ROLE | Varchar(255) | Blanks Replace: “No Value Provided”    *Does not Contain Unknown and Does not Apply rows* | In Title Case  Unique: 12 |

1. **Dim\_PERSON\_INJURY:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| PERSON\_INJURY\_SK | Integer | Auto generated |  |
| PERSON\_INJURY | Varchar(80) |  | In Title Case  Unique: 3 |

1. **Dim\_PERSON\_SEX:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| **PERSON\_SEX** \_SK | Integer | Auto generated |  |
| PERSON\_SEX | Varchar(20)    changed from varchar(10) to varchar(20) | Blanks Replace: “No Value Provided”      *Contains U* | In Title Case  Unique: 4 |

1. **Dim\_PERSON\_TYPE:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| PERSON\_ TYPE\_SK | Integer | Auto generated |  |
| PERSON\_TYPE | Varchar(80) |  | In Title Case  Unique: 4 |

1. **Dim\_POINT\_OF\_IMPACT:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| POINT\_OF\_IMPACT\_SK | Integer | Auto generated |  |
| POINT\_OF\_IMPACT | Varchar(255) | Blanks Replace: “No Value Provided” | In Title Case  Unique: 20 |

1. **Dim\_POSITION\_IN\_VEHICLE:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| POSITION\_IN\_VEHICLE\_SK | Integer | Auto generated |  |
| POSITION\_IN\_VEHICLE | Varchar(255) | Blanks Replace: “No Value Provided”    *Does not Contain Unknown and Does not Apply rows* | In Title Case  Unique: 12 |

1. **Dim\_PRE\_CRASH:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| PRE\_CRASH\_SK | Integer | Auto generated |  |
| PRE\_CRASH | Varchar(255) | Blanks Replace: “No Value Provided” | In Title Case  Unique: 20 |

1. **Dim\_SAFETY\_EQUIPMENT:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| SAFETY\_EQUIPMENT\_SK | Integer | Auto generated |  |
| SAFETY\_EQUIPMENT | Varchar(255) | Blanks Replace: “No Value Provided”    *Contains Unknown and*   *‘-’ replace Invalid* | In Title Case  Unique: 19 |

1. **Dim\_STATE\_REGISTRATION:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| STATE\_REGISTRATION\_SK | Integer | Auto generated |  |
| STATE\_REGISTRATION | Varchar(80) | Blanks Replace: “No Value Provided”    *There are only 50 states so rest we are replacing them with invalid:*    Invalid values:  ZZ  QB  ON  MH  NB  MB  FO  GU  FM  DP  NS  LR  BC  DI  T  AB  NF  AS  DC  MX  PR  YT  UK  GV  PQ  SK  NT  OT  UA  PE  VI  MP | In Title Case  Unique: 51 |



|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| TRAVEL\_DIRECTION\_SK | Integer | Auto generated |  |
| TRAVEL\_DIRECTION | Varchar(255) | Blanks Replace: “No Value Provided”    *Contains Unknown and*   *‘-’and ‘U’  replace Invalid*  *‘S’ replace South*  *‘W’ replace West*  *‘E’ replace East*  *‘N’ replace North* | In Title Case  Unique: 11 |

1. **Dim\_VEHICLE\_MAKE:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| VEHICLE\_MAKE\_SK | Integer | Auto generated |  |
| VEHICLE\_MAKE | Varchar(80) | Blanks Replace: “No Value Provided”    Replace this all with Unknown:    unk  UNK BLACK VEH  Unk Bus  UNK CITY BUS  unk garbage truck  UNK L/S/A  unk left scene  unk make  UNK MAKE BICYCLE  UNK MAKE OF SCOOTER  unk red vehicle  UNK SCOOTER  UNK TRUCK  UNK VAN  UNK YEAR  UNK.  UNK. YW TAXI  UNK/L/SA  UNK/LEFT SCENE  UNK/LSA  UNK/MAKE/YEAR  UNKINOWN  UNKL  UNKN  UNKN GOVT VEH (USPS)  UNKN NOT REGISTERED  UNKN SELF INSURED  UNKN UNREGISTERED  UNKNIWN  unknkwn  UNKNNOWN  unknow  UNKNOW MAKE & YEAR  unknown  unknown (electric scooter)  unknown - left scene  UNKNOWN - LEFT THE SCENE  UNKNOWN 18 WHEEL TRUCK  UNKNOWN BICYCLE  UNKNOWN BUS  unknown chinese  UNKNOWN ELECTRIC  Unknown fled scene  UNKNOWN FORKLIFT  UNKNOWN GAS SCOOTER  unknown left scene  UNKNOWN LEFT SCENE OF AN ACCIDENT  unknown left the scene  UNKNOWN LSA  unknown LSOA  UNKNOWN MAKE  UNKNOWN MAKE SUBURBAN  unknown make/ model  UNKNOWN MAKE/YEAR  unknown pickup truck  unknown postal truck  unknown trailer  UNKNOWN TRUCK  UNKNOWN VAN  UNKNOWN VEHICLE  UNKNOWN VEHICLE LEFT SCENE  UNKNOWN YEAR  unknown, left scene  UNKNOWN-LEFT SCENE  unknown.  UNKNOWN/LEFT SCENE  UNKNOWN/NONE  UNKNOWNS  unkonwn  unkown  UNKOWN BLACK SEDAN 4DR  UNKOWN/BUS  unkwon | Mixed Case  Unique: 7879 |

1. **Dim\_VEHICLE\_MODEL:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| VEHICLE\_MODEL\_SK | Integer | Auto generated |  |
| VEHICLE\_MODEL | Varchar(80) | Blanks Replace: “No Value Provided”    Replace this all with Unknown:UNK  unk left scene  unk make or model  unk model  UNK/LSA  UNKN  unkn lsoa  unknkwn  UNKNOWN  UNKNOWN LEFT SCENE  Unknown Vehicle  unknown white van  UNKNOWN-LEFT SCENE  unkown  AMBULANCE, | Mixed Case  Unique: 2764 |

1. **Dim\_VEHICLE\_TYPE:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| VEHICLE\_TYPE\_SK | Integer | Auto generated |  |
| VEHICLE\_TYPE | Varchar(80) | Blanks Replace: “No Value Provided”    Replace this all with Unknown:   unk  UNK FEMALE  UNK L  UNK T  Unk,  UNK.  UNK/L  unkl  UNKN  UNKNO  UNKNOW  UNKNOWN  UNKNOWN VE  UNKNW  UNKOW  UNKOWN  UNKWN | Mixed Case  Unique: 1478 |

1. **Dim\_ vehicle\_type\_code \_:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| vehicle\_type\_code\_SK | Integer | Auto generated |  |
| vehicle\_type\_code | Varchar(255) | Blanks Replace: “No Value Provided”    *Contains Unknown and*    5 different columns, we took union and unique of those values | In Title Case  Unique: 11 |

**25. Dim\_VEHICLE\_DAMAGE**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| VEHICLE\_DAMAGE\_SK | Integer | Auto generated |  |
| VEHICLE\_DAMAGE | Varchar(255) | Blanks Replace: “No Value Provided” | In Title Case  Unique: 20 |

**26. Dim\_PUBLIC\_PROPERTY\_DAMAGE**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Datatype (changes if any)** | **Cleansing** | **Details** |
| PUBLIC\_PROPERTY\_DAMAGE\_SK | Integer | Auto generated |  |
| PUBLIC\_PROPERTY\_DAMAGE | Varchar(255) | Blanks Replace: “No Value Provided”    *Contains “Unspecified”* | In Title Case  Unique: 4 |

1. **PUBLIC\_PROPERTY\_DAMAGE\_TYPE:**

**Unique: 14597**

**Blanks Replace: “No Value Provided”**

1. CROSS\_STREET\_NAME:

Blanks & Nulls replaced with “No Value Provided”

1. **VEHICLE\_YEAR:**

**Unique: 300**

**Blanks Replace: “No Value Provided”**

1. **On\_street\_name & off\_street\_name**

Blanks & Nulls replaced with “No Value Provided”

1. Zip\_code:

Nulls replaced with “No Value Provided”

1. ERR\_MODEL\_YEAR:

select \* from [NYC\_MV\_stg\_tables].[dbo].[stg\_nyc\_mv\_collision\_vehicles]

where [VEHICLE\_YEAR] > YEAR([CRASH\_DATE]) or YEAR([CRASH\_DATE]) >= 2022

ORDER BY [VEHICLE\_YEAR] ASC

1. For ERR\_PERSON\_AGE:   
   We considered all the person age between 0 and 100. So to populate this table we took person age <0 and person age >100. And in the fact we replace these with invalid values.

**Dimension: Null, Blank, Unknown, Does Not Apply**

|  |  |  |
| --- | --- | --- |
| **SK** | **Type of Missing Data** |  |
| -99 | No Value Provided | Nulls and Blanks  (In ERR\_Person\_Age we’ve replaced the NULLS with –99 directly as Person\_Age column is of *Integer datatype*) |
| -1 | Unknown |  |
| -2 | Does Not Apply |  |
| -3 | Invalid |  |
| -4 | Unspecified |  |

**We need to make sure to ignore distinct and consider case sensitivity while replacing invalid column values coming from stage while joining it with dimension.**

**DIMENSIONAL MODEL NOTES & CHANGES: -**

**A screenshot of a computer

Description automatically generated with medium confidence**

ER Model and Changes: -

**A picture containing diagram, plan, line, schematic

Description automatically generated**

1. *UNIQUE\_ID in* both the Fct\_Collision\_Person and Fct\_Collision\_Vehicle were transformed to  *Unique\_ID\_NK* since it is coming from the source table. And we added surrogate keys for each of theses tables
2. *FCT\_Collision\_Crash* will be loaded first (we will join on Collision\_ID) since there is a ONE-to-MANY relation between FCT\_Collision\_Person and FCT\_Collision\_Vehicle respectively.
3. Vehicle\_Contributing\_Number (FCT\_Collision\_Vehicles\_Contributing\_Factors):

* This is basically the count of Contributing factors involved in the Crash, we take **unique combination** of *Unique ID & Collision ID* (count will be row wise)

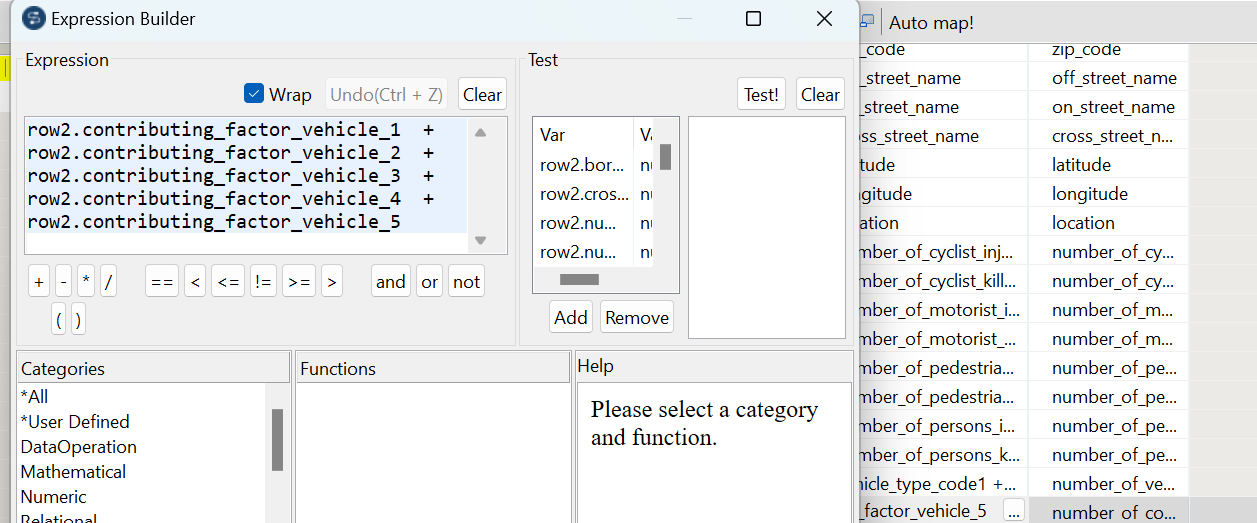
**A screenshot of a computer

Description automatically generated with medium confidence**

* Same for no\_of contributing\_factors (in Fct\_Collision\_Crashes)
* Same for **Factor\_Number(**FCT\_Collision\_Persons\_Contributing\_Factors)
* Same for no\_of\_vehicles\_involved (Fct\_Collision\_Crashes)   
  **A picture containing text, font, line, diagram

  Description automatically generated**

What we did was, using tAggregaterow, we first calculated the count of cfv1, cfv2 etc.. Group by collision\_id, then we mapped the count to another tMap and performed summation to get the column (no\_of contributing\_factors)   
   
   
**A computer screen shot of a calculator

Description automatically generated with low confidence**   
   
****

1. We connected FCT\_Collision\_Crashes\_Vehicles & FCT\_Collisions\_Vehicles\_Contributing\_Factors

Because FCT\_Collision\_Crashes gives us an overview of the contributing factors, if we want a detailed information, we can get it through this connection along with vehicle type code detail since it’s dim is also connected through that.   
   
**A screenshot of a computer

Description automatically generated with medium confidence**

1. **Distinct Contributing\_Factors\_Vehicle\_1 (from Crash stg table) encompassed all the distinct values from all other contributing\_factor (Vehicles, Persons staging table) columns from all staging tables**

****

1. While joining the FCT\_collision\_crashes table with fct\_collision\_persons table and fct\_collision\_ vehicles tables we used only collision id because:    
      
   a) Reason for not choosing crash\_date along with collision id is that, there were many discrepancies between all the 3 tables with regards to crash\_date   
      
   b) Reason for not choosing Unique\_id along with collision\_id is because unique\_id was different for persons and vehicles tables and they were not related to each other and this column was also not present in the fct\_crash

7. Discrepancies in vehicle type:   
   
**A picture containing text, screenshot, document

Description automatically generated**   **A picture containing text, screenshot, font, handwriting

Description automatically generated**   
   
   
   
**A screenshot of a computer

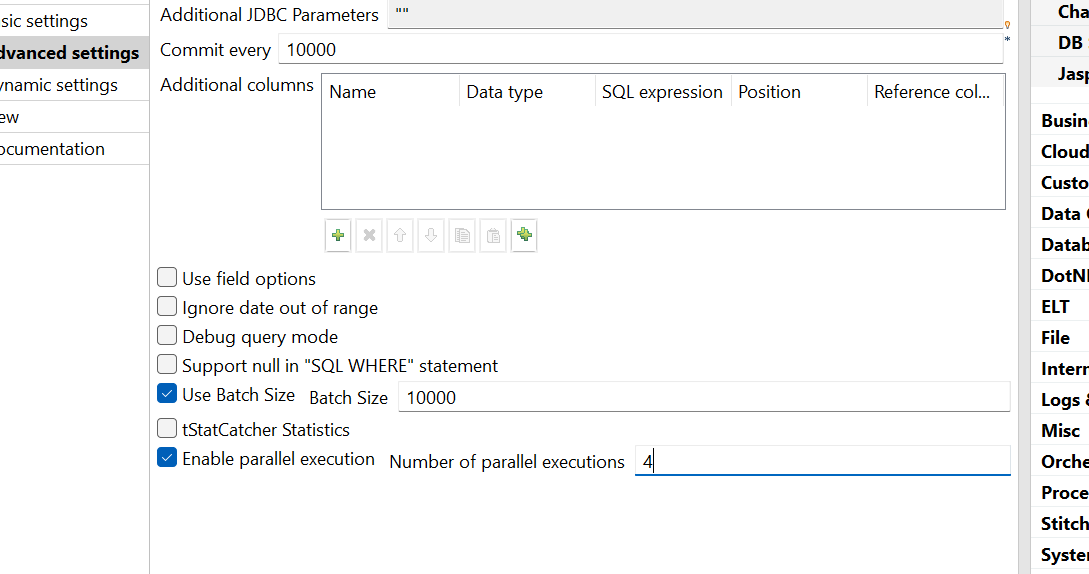
Description automatically generated with medium confidence** 

\

**Optimizing techniques:**   
   
1) We made sure to select only the required columns in the tinput component.

2) Watching Professor Rick Sherman’s, we made changes to the JVM arguments according to the laptop specification to increase the load time whenever we got heap size or memory issues.    
   
So according to the video since we were using 16 GB ram so we allocated maximum memory 8G and the minimum starts with 512MB, these changes will be implemented in talend backend file   
   
**A screenshot of a computer

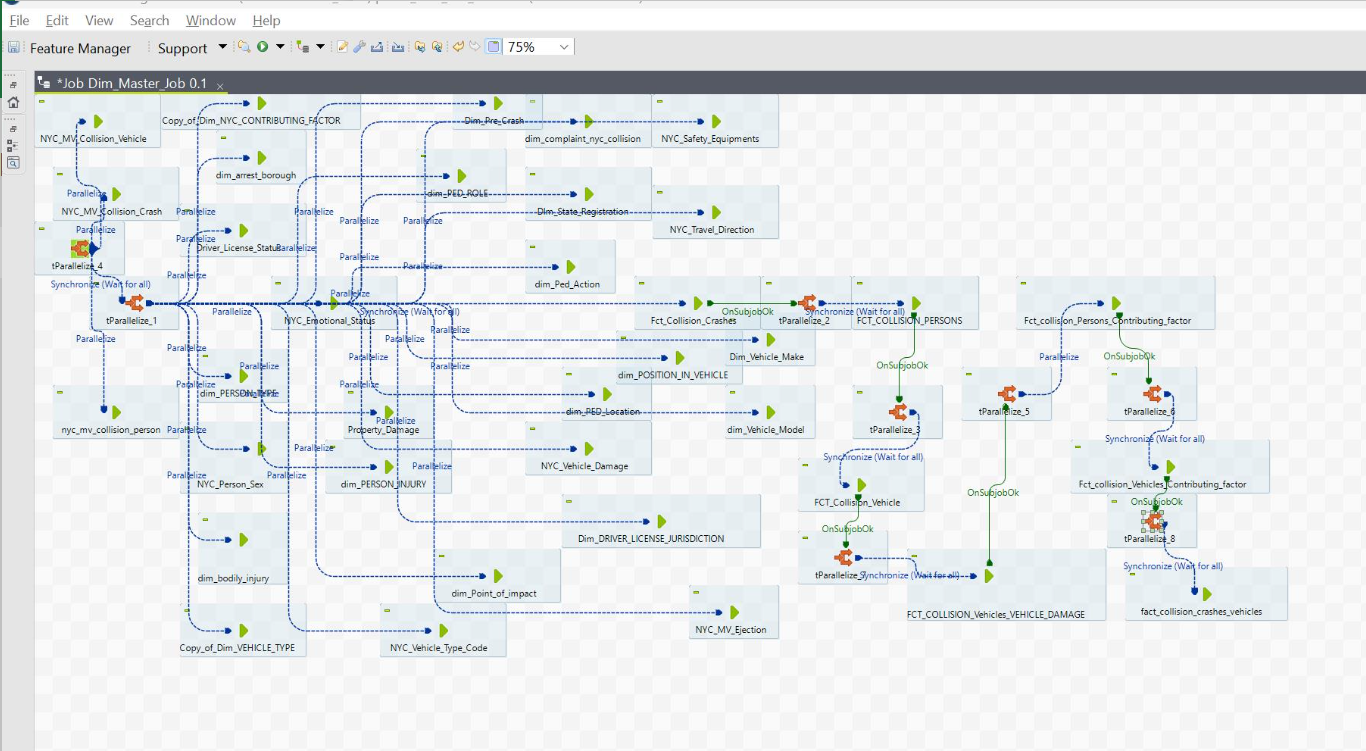
Description automatically generated with medium confidence**

3) In our master job we followed the same sequence of loading the tables is that:    
a) Using tparrallelize, we first loaded all 3 staging tables in parallelize method. (Here it was taking a very long time to get data directly from big query so we enabled parallel execution and made it to 4)   
   
****

B) Then we used Synchronize (Wait for all) feature. After the staging tables were loaded, we added all the 26 dimensions in the parallelize mode.    
c) Then the final step was loading 7 facts one by one using Synchronize (Wait for all) and on subjob ok trigger due to fk pk constraints.   
The sequence was: Fct\_Collision\_Crash   
FCT\_Collision\_Person   
FCT\_Vehicles   
FCT\_Collision\_Vehicle\_Vehicles\_Damage   
FCT\_Collision\_Person\_Contributing\_Factors   
FCT\_Collision\_Vehicles\_Contributing\_Factors   
Fct\_Collision\_crashes\_vehicles 

**A screenshot of a computer

Description automatically generated**   
 

****