NYC MV Collisions

NYC OpenData

BigQuery

Assignment 2

- Load Data from Staging to Integration Schema (dimensional model) with Talend
- Use SQL Server, Azure SQL, MySQL or Azure SQL

ATHENA IT SOLUTIONS

Assignment 2: Tasks

- Load data from STG tables to dimensional model & perform data cleansing using Talend
  - Same database as STG tables reside

- Document any data cleansing tasks and results and explain

- Query dimensional data model with the listed business questions as a minimum

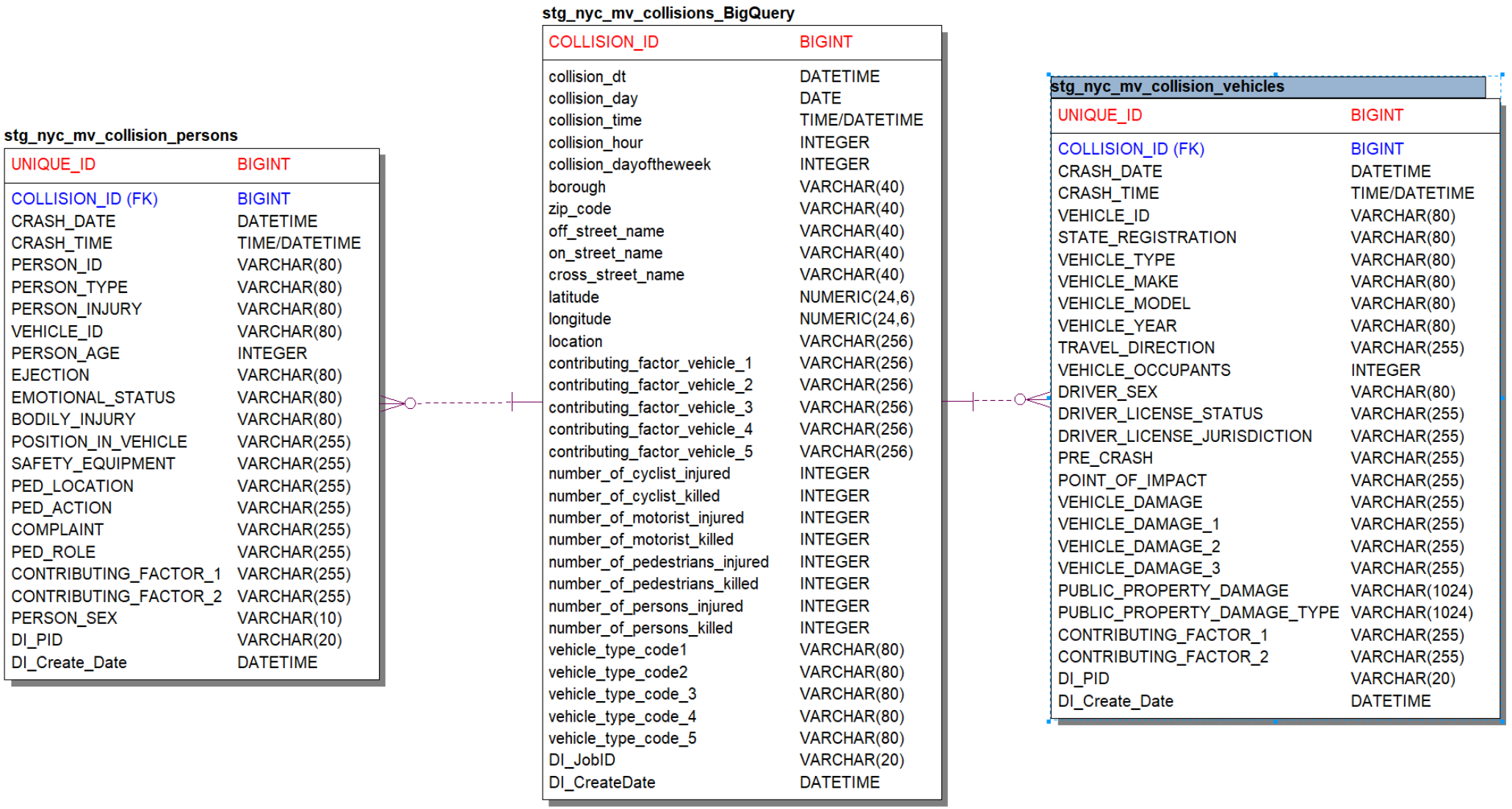
Assignment 2

- Upload
  - Screen shots of completed loads
  - Time each load took (should be on screen shots above)
    - Overall Orchestrator or Master job
  - Completed Talend jobs
  - List of tables with rows counts
  - Results of each query

NYC MV Collisions

Data Model

Staging Tables: Crashes, Vehicles, Persons



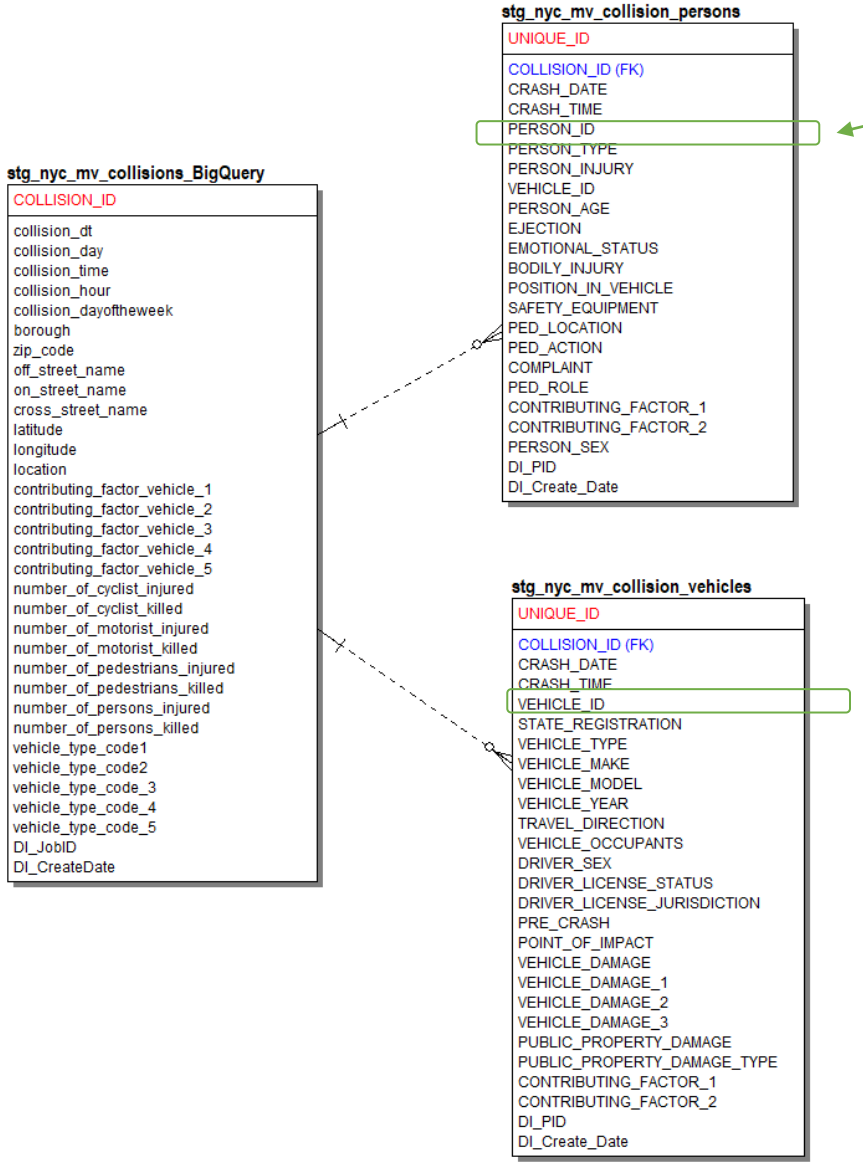
Dimensional Model: Examining Stage Tables

Primark Keys (PKs):

- Collision_ID & two Unique_IDs
- The Unique_IDs are surrogate keys & are NOT related to each other

Crash_date & Crash_Time (or collision_day & collision_time) repeated in each table

- Not necessary in tables when tables are used together

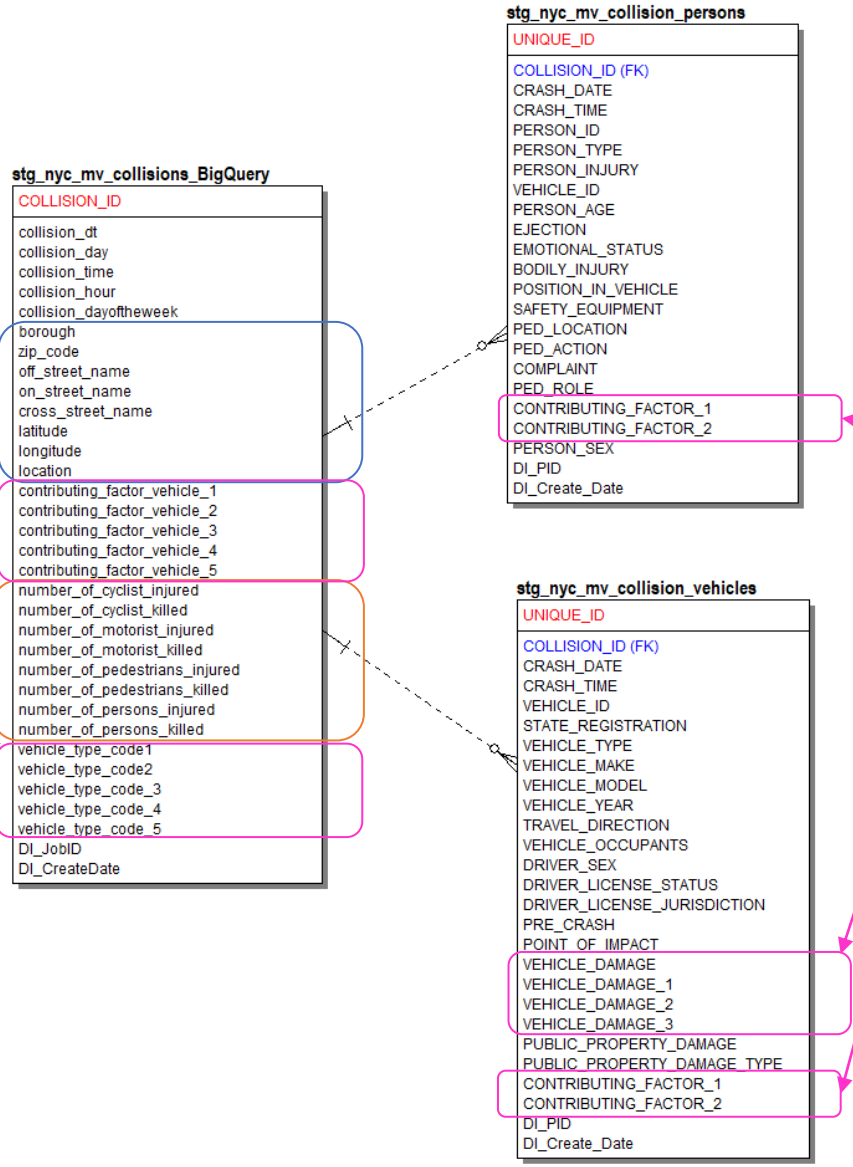


- 1-n number of rows per collision
- Person_ID
  - how it is populated changes over time (based on data profiling NOT documentation)
  - NOT a unique identifier

- 1-n number of rows per collision
- Vehicle_ID
  - how it is populated changes over time (based on data profiling NOT documentation)
  - NOT a unique identifier

Dimensional Model: Examining Stage Tables

- Collision Location
- Summary statistics
- Repeating Groups
  - Assumes 5 or less vehicles but there are sometimes more than 5
  - _n is associated with the n number vehicle
  - This data is redundant with what is in Vehicles tables but should we ignore?



- Repeating Groups
  - Used to provide multiple values in a single row
  - Should we normalize????

Dimensional Data Model (INT Schema)

- nyc mv collisions dimensional model.sql

Dimensional Model

Surrogate Keys (SKs) & Foreign Keys (FK)

Data Quality

Dimensions: NULL

Each dimension should have a “No Value Provided” or appropriate phrase with -99 (or other negative number) as the value for the surrogate key (SK)

```
INSERT INTO Dim_COMPLAINT
(COMPLAINT_SK,COMPLAINT, DI_PID, DI_Create_Date)
VALUES(-99, ‘No Value Provided', 'Manually input', getdate());
```

Note: In this dataset there are some values in the dimensions such as “Unknown”, “Not Appliable”, etc. Those should be used and NOT be replaced by the -99 SK above

Dimensions

- Cleanup
  - Dim_TRAVEL_DIRECTION – there should not be redundant directions descriptions so do some cleaning
  - VEHICLE_YEAR and PERSON_AGE have many invalid values. Next slide discusses cleaning

- Bonus
  - What other dimensions or dimensional attributes would you suggest to cleanup. You would have to have a reasonable ability to cleanup the data.

Two Error Tables

Determine what criteria you would use to invalidate model year or person age?
Based on that criteria populate the tables below and replace NULL with invalid value in FCT table

```
CREATE TABLE ERR_Model_Year (
UNIQUE_ID int NOT NULL,
COLLISION_ID int NULL,
VEHICLE_ID varchar(80),
CRASH_DATE date NULL,
VEHICLE_YEAR int NULL,
DI_PID varchar(20) NULL,
DI_Create_Date datetime DEFAULT getdate()
NULL,
PRIMARY KEY (UNIQUE_ID)
);
```

```
CREATE TABLE ERR_Person_Age (
UNIQUE_ID int NOT NULL,
COLLISION_ID int NULL,
PERSON_ID varchar(80),
CRASH_DATE date NULL,
PERSON_AGE int NULL,
DI_PID varchar(20) NULL,
DI_Create_Date datetime DEFAULT getdate()
NULL,
PRIMARY KEY (UNIQUE_ID)
);
```