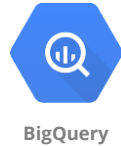


NYC MV Collisions

NYC OpenData



Assignment 1

- Perform Data Profiling (with Alteryx)
- Load Data into Staging Tables with Talend
 - SQL Scripts for staging tables:
OneDrive\...\damg7370_2022_Fall\Data - NYC Motor Vehicle Collision\snycc mv collisions Stage Tables.sql
- Create Preliminary Dimensional Model
 - List facts & dimensions
 - Create a list of Stage Tables' columns and map to your proposed dimensional model
 - A data model diagram or script are for bonus



NYC MV Collisions: Deliverables

Assignment 1:

- Perform Data Profiling (with Alteryx)
- Load Data into Staging Tables with Talend
 - SQL Scripts for staging tables: OneDrive\...\damg7370_2022_Fall\Data - NYC Motor Vehicle Collisions\nyc mv collisions Stage Tables.sql
 - Use SQL Server, Azure SQL, MySQL or Azure SQL
- Create Preliminary Dimensional Model
 - List facts & dimensions
 - Create a list of Stage Tables' columns and map to your proposed dimensional model
 - A data model diagram or script are for bonus

NYC MV Collisions: Deliverables

Assignment 1: Data Profiling using Alteryx

- Deliverables: (upload the following)
 - Screen shots of completed profiling
 - Time each profiling job takes
 - Completed Alteryx jobs

NYC MV Collisions: Deliverables

Assignment 1: Load Data into Staging Tables with Talend

- Deliverables: (upload the following)
 - Screen shots of completed jobs
 - Time each job takes and total time with all three loads in single job
 - Completed Talend jobs
 - List of tables with rows counts

NYC MV Collisions: Deliverables

Assignment 1: Create Preliminary Dimensional Model

Deliverables: (upload the following)

- List facts & dimensions
- Create a list of Stage Tables' columns and map to your proposed dimensional model
- A data model diagram or script not required but is a bonus

Staging Tables: Crashes, Vehicles, Persons

- Two columns renamed
- 4 columns derived

stg_nyc_mv_collision_persons

UNIQUE_ID	BIGINT
COLLISION_ID (FK)	BIGINT
CRASH_DATE	DATETIME
CRASH_TIME	TIME/DATETIME
PERSON_ID	VARCHAR(80)
PERSON_TYPE	VARCHAR(80)
PERSON_INJURY	VARCHAR(80)
VEHICLE_ID	VARCHAR(80)
PERSON_AGE	INTEGER
EJECTION	VARCHAR(80)
EMOTIONAL_STATUS	VARCHAR(80)
BODILY_INJURY	VARCHAR(80)
POSITION_IN_VEHICLE	VARCHAR(255)
SAFETY_EQUIPMENT	VARCHAR(255)
PED_LOCATION	VARCHAR(255)
PED_ACTION	VARCHAR(255)
COMPLAINT	VARCHAR(255)
PED_ROLE	VARCHAR(255)
CONTRIBUTING_FACTOR_1	VARCHAR(255)
CONTRIBUTING_FACTOR_2	VARCHAR(255)
PERSON_SEX	VARCHAR(10)
DI_PID	VARCHAR(20)
DI_CreateDate	DATETIME

stg_nyc_mv_collisions_BigQuery

COLLISION_ID	BIGINT
collision_dt	DATETIME
collision_day	DATE
collision_time	TIME/DATETIME
collision_hour	INTEGER
collision_dayoftheweek	INTEGER
borough	VARCHAR(40)
zip_code	VARCHAR(40)
off_street_name	VARCHAR(40)
on_street_name	VARCHAR(40)
cross_street_name	VARCHAR(40)
latitude	NUMERIC(24,6)
longitude	NUMERIC(24,6)
location	VARCHAR(256)
contributing_factor_vehicle_1	VARCHAR(256)
contributing_factor_vehicle_2	VARCHAR(256)
contributing_factor_vehicle_3	VARCHAR(256)
contributing_factor_vehicle_4	VARCHAR(256)
contributing_factor_vehicle_5	VARCHAR(256)
number_of_cyclist_injured	INTEGER
number_of_cyclist_killed	INTEGER
number_of_motorist_injured	INTEGER
number_of_motorist_killed	INTEGER
number_of_pedestrians_injured	INTEGER
number_of_pedestrians_killed	INTEGER
number_of_persons_injured	INTEGER
number_of_persons_killed	INTEGER
vehicle_type_code1	VARCHAR(80)
vehicle_type_code2	VARCHAR(80)
vehicle_type_code3	VARCHAR(80)
vehicle_type_code4	VARCHAR(80)
vehicle_type_code5	VARCHAR(80)
DI_JobID	VARCHAR(20)
DI_CreateDate	DATETIME

stg_nyc_mv_collision_vehicles

UNIQUE_ID	BIGINT
COLLISION_ID (FK)	BIGINT
CRASH_DATE	DATETIME
CRASH_TIME	TIME/DATETIME
VEHICLE_ID	VARCHAR(80)
STATE_REGISTRATION	VARCHAR(80)
VEHICLE_TYPE	VARCHAR(80)
VEHICLE_MAKE	VARCHAR(80)
VEHICLE_MODEL	VARCHAR(80)
VEHICLE_YEAR	VARCHAR(80)
TRAVEL_DIRECTION	VARCHAR(255)
VEHICLE_OCCUPANTS	INTEGER
DRIVER_SEX	VARCHAR(80)
DRIVER_LICENSE_STATUS	VARCHAR(255)
DRIVER_LICENSE_JURISDICTION	VARCHAR(255)
PRE_CRASH	VARCHAR(255)
POINT_OF_IMPACT	VARCHAR(255)
VEHICLE_DAMAGE	VARCHAR(255)
VEHICLE_DAMAGE_1	VARCHAR(255)
VEHICLE_DAMAGE_2	VARCHAR(255)
VEHICLE_DAMAGE_3	VARCHAR(255)
PUBLIC_PROPERTY_DAMAGE	VARCHAR(1024)
PUBLIC_PROPERTY_DAMAGE_TYPE	VARCHAR(1024)
CONTRIBUTING_FACTOR_1	VARCHAR(255)
CONTRIBUTING_FACTOR_2	VARCHAR(255)
DI_PID	VARCHAR(20)
DI_CreateDate	DATETIME

How develop Data Model

- Source systems analysis (data sources)
 - Any documentation
 - ~~Talk to the SME (subject matter expert)~~
 - Data profiling (if possible)
 - Ingest data into staging table for further data analysis
- Examine data for:
 - Data consistency
 - Columns use differently between data sources or at different times even within a single data source
 - Data quality
 - Invalid data values
 - Invalid data types
 - Data structures that are too normalized or too denormalized
 - Redundant data
 - Pre-summarized data
 - Repeating groups
- Map Source Tables, Columns to Integration Tables, columns