NYC MV Collisions

NYC OpenData

BigQuery

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\snyc 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

Copyright © 2022 Athena IT Solutions All rights reserved.

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

Copyright © 2022 Athena IT Solutions All rights reserved.

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

Copyright © 2022 Athena IT Solutions All rights reserved.

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

Copyright © 2022 Athena IT Solutions All rights reserved.

NYC MV Collisions: Deliverables

Assignment 1: Create Preliminary Dimensional Model

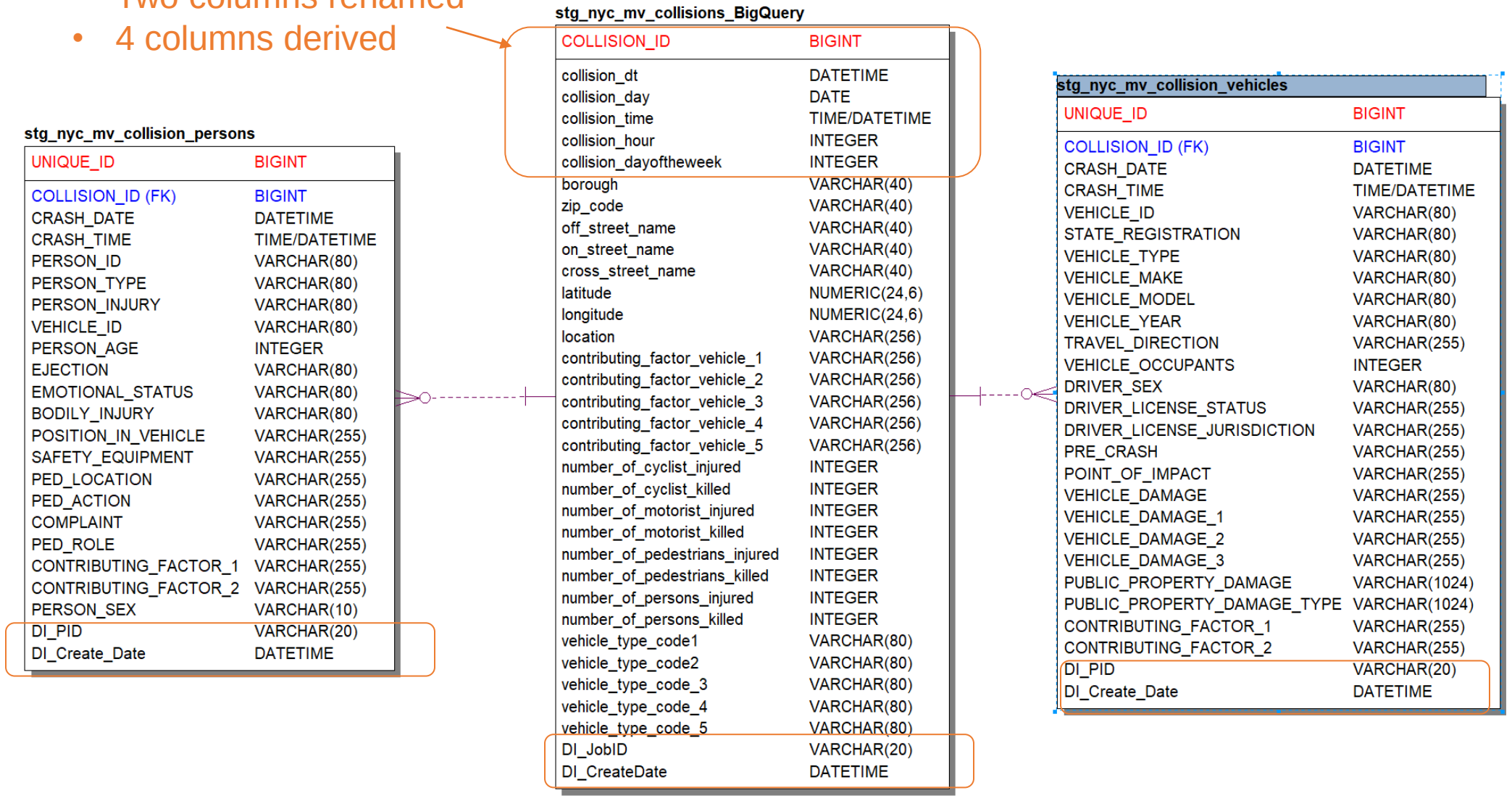
Deliverables: (upload the following)

- 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

Copyright © 2022 Athena IT Solutions All rights reserved.

Staging Tables: Crashes, Vehicles, Persons

- Two columns renamed
- 4 columns derived



Copyright © 2022 Athena IT Solutions All rights reserved.

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

Copyright © 2022 Athena IT Solutions All rights reserved.