PSEUDOCODE FOR ETL TRANSFORMATION

Accident Analysis Data Warehouse

Step 1: Make empty tables to store our data

- Make a table Dim_Location with columns Location ID, Location Name
- Make a table Dim_Vehicle with columns Vehicle ID, Vehicle Type
- Make a table Dim_Road Condition with columns Road Condition ID, Surface, Visibility
- Make a table Dim_Date with columns Accident ID, Date ID, Location ID, Vehicle ID, Road Condition ID, Vehicles Involved, Severity Score

Note: These are our star schema tables!

Step 2: fill Dim_location with unique locations

Get all unique location names from accidents table
Start with location ID = 1
for each unique location
    Add a row to Dim_location with (location_ID, location)
    Increase location ID by 1

Note: Giving each location a number (like 1, 2, 3, ...)

Step 3: fill Dim_Vehicle with vehicle info

for each row in vehicles table:
    Add a row to Dim Vehicle with (row.Vehicle ID, row.Vehicle Type)

Note: Just copying Vehicle ID and Vehicle Type as-is

Step 4: fill Dim_Road Condition with road conditions

for each row in Road Conditions table:
    Add a row to Dim_Road Condition with (row.Condition ID, row.Surface, row.Visibility).

Note: Just copying Condition ID, Surface and Visibility as-is

Step 5: fill Dim_Date with dates and times.

Get all unique Reported At timestamps from Accidents table
Starts with Date ID = 1.
for each timestamp:
    Get the date part (like 2025-02-20)
    Get the month part (like 2 for February).
    Get the year part (like 2025)

Get the time part (like 15 15 34)

Add a row to Dim-Date with (Date ID, date, month, year, time)

Increase Date ID by 1

Note: Splitting Reported At into pieces for easier analysis

Step 6: fill fact_Accidents with accident data

for each row in Accidents table

* find the Date ID by matching the timestamp

find the Date ID from Dim_Date where the date and time match row. Reported At.

find the Location ID by matching the location

find the location ID from Dim_location where Location Name = row. location

Pick a random Vehicle ID (pretends it's the main vehicle).

Pick random number between 1 and 200 for Vehicle ID

find the Road Condition ID by matching location and closest timestamp.

find the condition ID from Road Conditions where:

location matches row. Location AND

Recorded At is closest to row. Reported At

If no match found:

Pick random number between 1 and 100 for Condition ID

Note: We're guessing the road condition if we can't find a close match

Turn Severity into a number (Severity score)

If row. Severity into a number (Severity score)

If row. Severity is "Minor", set Severity score to 1

If row. Severity is "Moderate", set Severity score to 2

If row Severity is "Severe", set Severity score to 3

If row. Severity is something else, set Severity score to 1

Note: Making Severity a number for easier math

Add the row to fact_Accidents.

Add a row to fact_Accidents with:

(row Accident ID, Date ID, location ID, Vehicle ID, Road Condition, row. Vehicle Involved, Severity Score)

Note: This table connects everything together!

END OF PSEUDOCODE

(Super easy steps for our ETL process!)