Step-by-Step Guide: Traffic Data Modeling (Star Schema)

STEP 0: Understand the Dataset (Very Important)

Rule #1 of data modeling:

First understand the grain of your data.

Grain of your Excel sheet

One row = One traffic crash

Each record contains:

Date & time

Location info

Weather & road conditions

Injury counts

Crash causes

-This clearly indicates:

We need ONE Fact table + multiple Dimension tables

STEP 1: Identify Fact vs Dimension Data

Ask these two questions for every column:

- Is it numeric and aggregatable (sum, avg, count)?
- Does it describe context (who, when, where, how)?

FACT DATA (Measures → go to Fact table)

From your Excel sheet:

num_units

injuries_total

injuries_fatal

injuries_incapacitating

injuries_non_incapacitating

injuries_reported_not_evident

injuries_no_indication

➞ These answer "how much / how many"

DIMENSION DATA (Descriptors → go to Dimension tables)

Category	Columns
Time	crash_date, crash_hour, crash_day_of_week, crash_month
Location	region, street, intersection_related_i
Weather	weather_condition, lighting_condition
Road	trafficway_type, alignment, road_defect
Severity/Cause	crash_type, first_crash_type, prim_contributory_cause

---

STEP 2: Design the Star Schema (Conceptual)

Why Star Schema?

As a senior analyst, I always prefer star schema for:

Clean relationships
Faster Power BI performance
Easy DAX writing
Interview & industry standard

Central Table: FACT_CRASH
crash_id (Primary Key – generated)
date_id (FK)
weather_id (FK)
road_id (FK)
severity_id (FK)
Measures
num_units
injuries_total
injuries_fatal
injuries_incapacitating
injuries_non_incapacitating
injuries_reported_not_evident
injuries_no_indication

STEP 3: Build Dimension Tables (One by One)

DIM_TIME
Purpose
Enables time-based analysis (trend, seasonality, peak hours)
Columns
date_id (PK)
crash_date
year
month
day
day_of_week
crash_hour
Time Hierarchy
Year
└── Month
└── Day
└── Hour
Used for:
Monthly trends
Daily analysis
Hourly congestion patterns

DIM_WEATHER

Purpose
Analyze environmental impact on crashes
Columns
weather_id (PK)
weather_condition
lighting_condition
roadway_surface_cond
- Example questions answered:
Do crashes increase during rain?
Is night-time more dangerous?

DIM_ROAD
Purpose
Stores road infrastructure details
Columns
road_id (PK)
trafficway_type
alignment
road_defect
traffic_control_device
- Helps evaluate:
Road design safety
Signal effectiveness

DIM_SEVERITY
Purpose
Explains crash seriousness and root cause
Columns
severity_id (PK)
crash_type
first_crash_type
most_severe_injury
prim_contributory_cause
- Used for:
Cause analysis
Injury severity patterns

STEP 4: Define Relationships (Data Integrity)
Relationship Rules (Sr. Analyst Standard)
One-to-Many (1:N) only
Dimension → Fact
Single-direction filtering
No bi-directional unless absolutely required

Relationships Table

Dimension	PK	Fact FK
Dim_Time	date_id	date_id
Dim_Location	location_id	location_id
Dim_Weather	weather_id	weather_id
Dim_Road	road_id	road_id
Dim_Severity	severity_id	severity_id

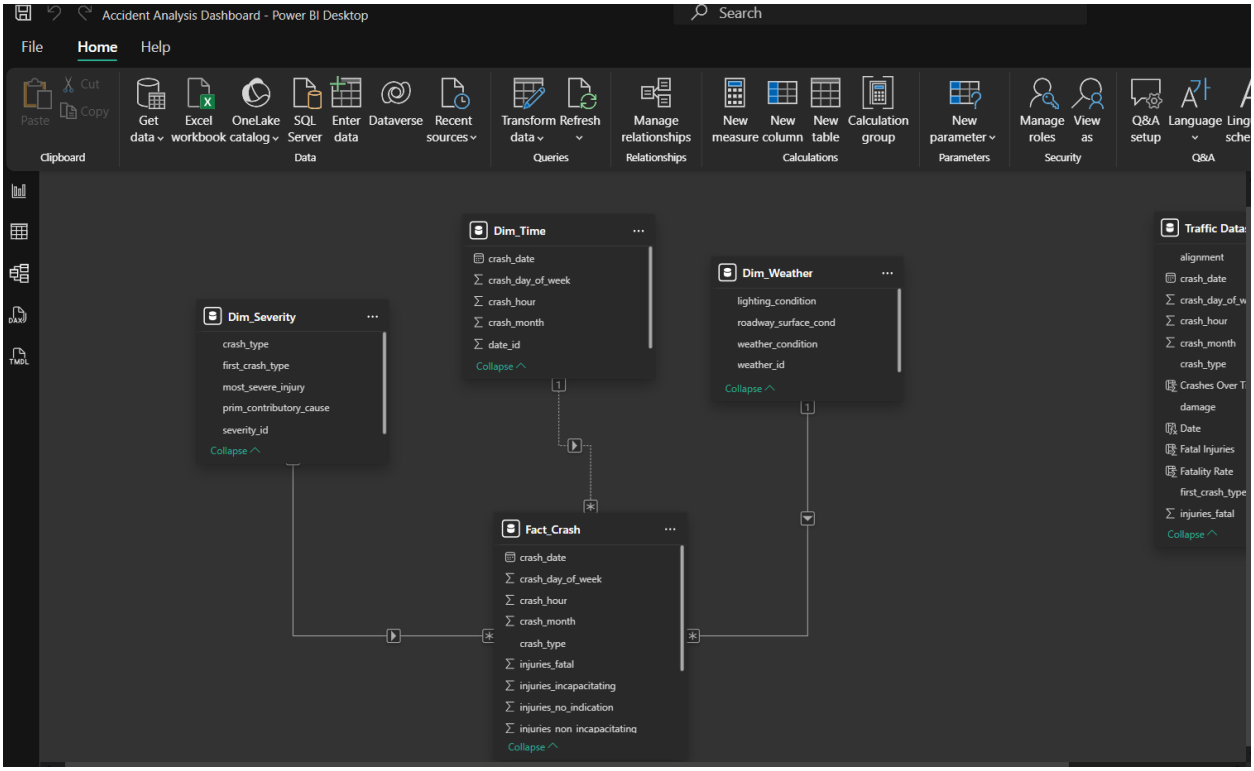
-This ensures:

No duplicate counting

Correct aggregations

Clean drill-down

STEP 5: Data Model Diagram (Visual Representation)



STEP 6: Hierarchies (Must-Have for BI Tools)

⏳ Time Hierarchy

Year → Month → Day → Hour

📍 Location Hierarchy

Region → Street → Intersection

📌 Enables:

Drill-down
Roll-up
Interactive dashboards