🟢 DAY 1 – DATA MODELING (SIMPLE NOTES)

1️⃣ Grain

Meaning

👉 **Grain = what ONE row in a table represents**

Always define grain using:

> **“One row represents …”**

Why grain is important

* Avoids double counting
* Decides what you can COUNT / SUM
* Wrong grain = wrong reports

Examples

* Orders table → one row = **one order**
* Order items table → one row = **one product in one order**
* Daily sales table → one row = **one seller per day**

Rule to remember

> **If one row tries to describe two different things, grain is wrong and table must be split.**

2️⃣ Fact vs Dimension

Fact table

* Stores **events or measurements**
* Things that happen
* Has numbers

Examples:

* Orders
* Payments
* Violations
* Ratings

Dimension table

* Stores **descriptive information**
* Who / What / When / Where

Examples:

* Customer

* Seller
* Product
* Date

Memory trick

> **Fact = action**
> **Dimension = description**

3️⃣ Why Dimensional Modeling

Dimensional modeling is used **for analytics**, not transactions.

Benefits

* Easy to understand
* Simple SQL
* Fewer mistakes
* Predictable joins
* Business-friendly reporting

Joins clarification

* More joins than one flat table ❌
* **Fewer and simpler joins than normalized relational models** ✅

4️⃣ Event vs Snapshot Tables

Event table

* Stores **every action**
* One row per event
* History preserved

Examples:

* Order placed
* Page viewed
* Violation occurred

Grain:

> One row = one event

Snapshot table

* Stores **state at a point in time**
* Daily / monthly summaries

Examples:

* Daily inventory
* Monthly salary

* Daily account balance

Grain:

> One row = one entity per time interval

Decision rule

Ask:

> **“Do I need every change or just the value at intervals?”**

5️⃣ Slowly Changing Dimensions (SCD)

SCD Type 1

* Overwrite old value
* No history

Use when:

* History doesn’t matter

Example:

* Fixing spelling error

SCD Type 2

* New row for every change
* History preserved

Use when:

* Past reports must remain correct

Examples:

* Seller country change
* Product category change

Decision rule

Ask:

> **"If this value changes, will old reports become wrong?"**

* Yes → SCD2
* No → SCD1

6️⃣ Indexing (Basics)

What is indexing

👉 **Index = faster way to find rows**

* Avoids full table scan
* Improves read performance
* Slows down writes

Common index columns

* Primary keys
* Foreign keys
* Columns in WHERE / JOIN

Interview line

> **“Indexes speed up reads but add overhead to writes, so they should be chosen based on query patterns.”**

⭐ ONE COMPLETE DATA MODELING QUESTION (NORMAL VERSION)

Question

Design a data model to track product compliance violations by sellers.

Business wants to:

* Analyze violations over time
* See reports by seller, product, rule, and date

Step 1: Define Grain

> **One row represents one compliance violation event for a product by a seller at a specific time.**

Step 2: Identify Fact & Dimensions

Fact table

fact_product_violations

* seller_id
* product_id
* rule_id
* violation_date
* violation_timestamp
* severity
* status

Why fact?

* A violation is an **event**
* Can be counted and analyzed

Dimension tables

* **dim_seller** (SCD Type 2 – seller details change)
* **dim_product** (SCD Type 2 – category/brand change)
* **dim_rule** (SCD Type 1 – rule text corrections)
* **dim_date**

Step 3: Event vs Snapshot

* Violations are **event-based**
* Daily summaries can be built later as snapshot tables

Step 4: Queries Supported

* Violations per seller per month
* Repeat violators
* Rule-wise violation trends
* High-risk sellers

Step 5: Performance (High level)

* Index foreign keys
* Partition fact table by date (warehouse)

Interview Closing Line

> **"This dimensional model uses an event-level grain, separates facts and dimensions clearly, preserves history where required using SCD Type 2, and supports scalable analytics."**

✅ DAY 1 SUMMARY (LOCK THIS)

* Grain defines one row
* Every table has a grain
* Facts = events
* Dimensions = context
* Dimensional modeling is for analytics
* Event vs snapshot matters
* SCD2 preserves history
* Indexing avoids full scans