📘 Data Modeling

🎯 Purpose of Data Modeling

To design structured data systems that support analytical needs, especially for OLAP (Online Analytical Processing) use cases like:

- Dashboards
- KPIs and trend reports
- Forecasting & ML models
- Data warehousing & reporting

🧩 OLAP Concepts: Fact & Dimension Tables

✅ Fact Table

- Stores **quantitative, event-based data** (measurable metrics)
- Examples: `sales`, `trips`, `payments`, `logins`
- Often grows very fast (append-only)
  Contains:
  - Foreign keys to dimensions
  - Numeric metrics like amount, time, distance

🛠️ **Example**: `trip_facts`

- Columns: `trip_id`, `driver_id`, `city_id`, `fare`, `duration`, `rating`

✅ Dimension Table

- Stores **descriptive information** related to facts
- Examples: `drivers`, `cities`, `riders`, `vehicles`, `date`
- Changes slowly (managed via SCD)

- Used for **filtering, slicing, and grouping**

🛠️ **Example**: `driver_dim`

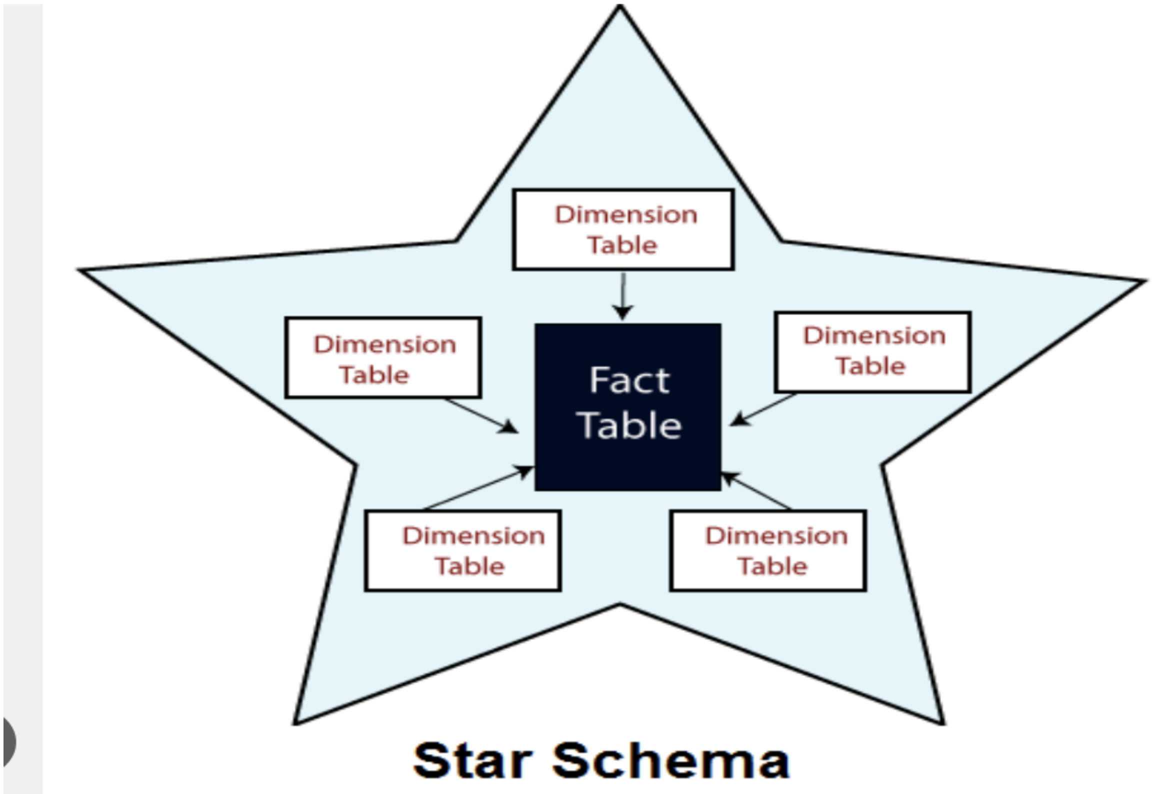
- Columns: `driver_id`, `name`, `city_id`, `rating`, `vehicle_type`

⭐ Star Schema

- 1 central **fact table**
- Surrounding **denormalized** dimension tables
- Simplified structure, faster for querying and BI tools

🧩 **Example**: Uber's Trip Data

- `trip_facts`: trip_id, driver_id, city_id, time_id, fare
- `driver_dim`, `rider_dim`, `time_dim`, `city_dim`

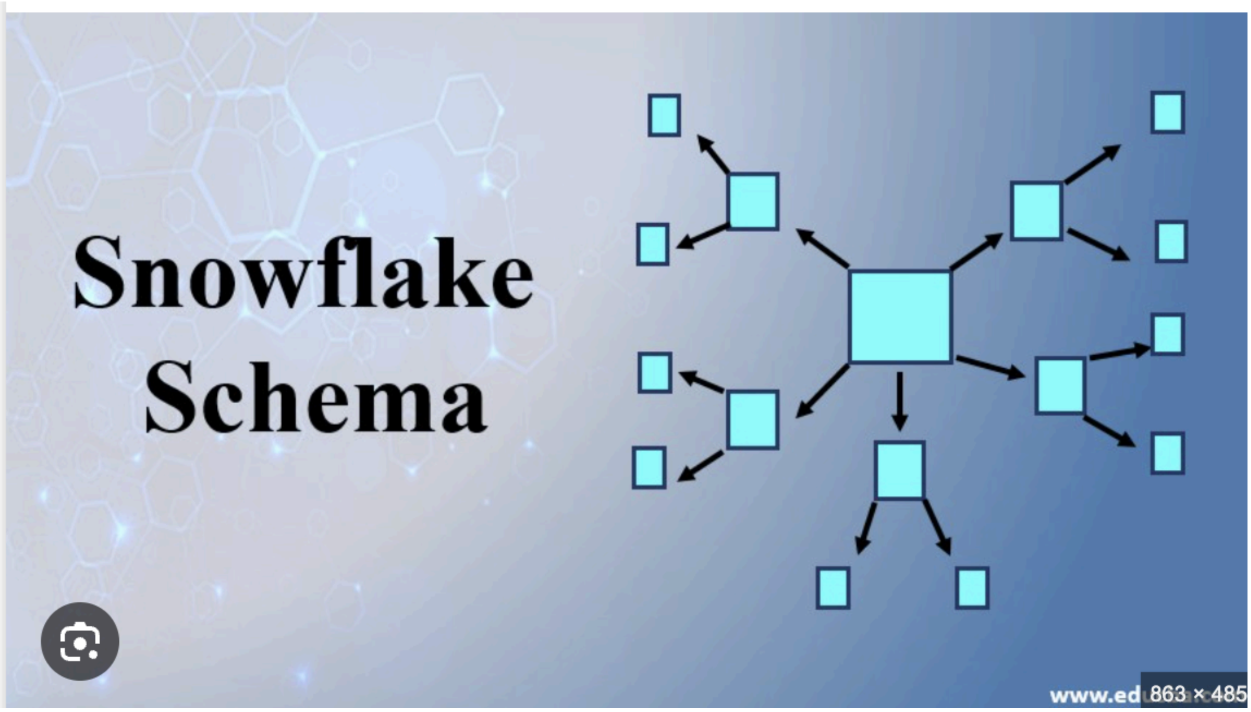


❄️ Snowflake Schema

- **Normalized dimensions** → further broken into sub-dimensions
- Reduces redundancy and improves data consistency
- Slightly slower queries due to more joins

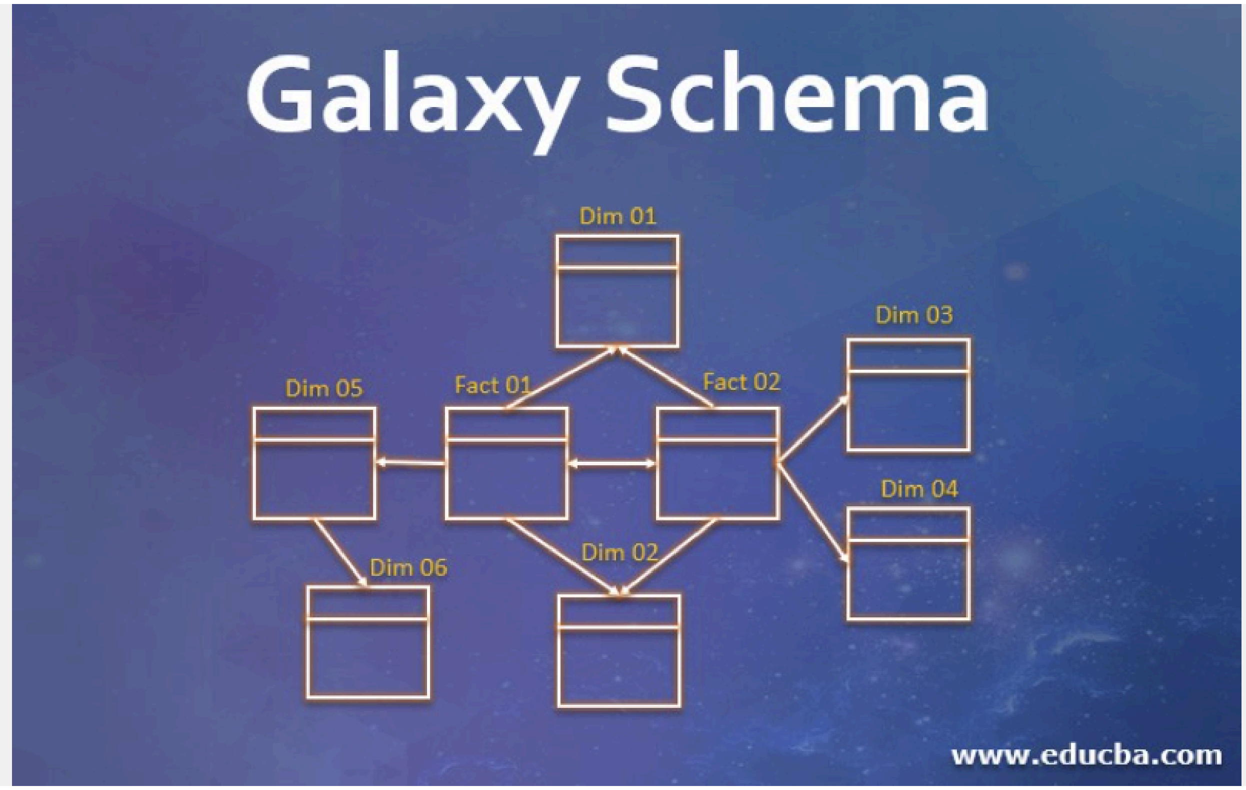
🧩 Example:

- `trip_facts → driver_dim → location_dim → region_dim`



🌌 Galaxy Schema (Fact Constellation)

- **Multiple fact tables** share common dimensions
- Used for multi-domain models like Uber's:
  - `trip_facts`, `payment_facts`, `support_ticket_facts`
- Shared dimensions: `driver_dim`, `city_dim`, `time_dim`



🐢 SCD – Slow Changing Dimensions

SCD Type	Description	Keeps History?	Example
SCD 0	Fixed – never changes	✅	Country code, gender
SCD 1	Overwrite existing data	❌	Update phone number
SCD 2	Add new row with version/timestamp	✅	Address or rating change
SCD 3	Add new column for current/previous	⚠️ Partial	Old vs new pricing tier
SCD 4	Separate current + historical table	✅	Active vs archived coupons
Hybrid	Mix of SCD1 + SCD2	✅	Overwrite contact info, version rating

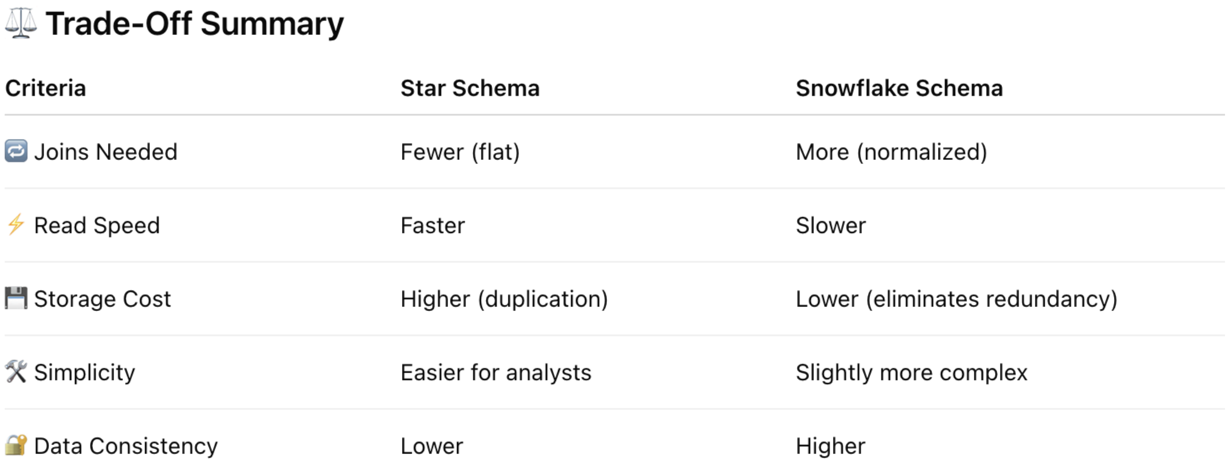
---

🧠 Summary Tips (For Interview)

- Mention **grain** of fact tables (e.g. 1 row per trip):

> "The grain of my `trip_facts` table is **one row per completed ride**.
> This means each row represents a unique trip, and includes metrics like fare, distance, trip duration, and foreign keys to dimensions like driver, rider, city, and time."

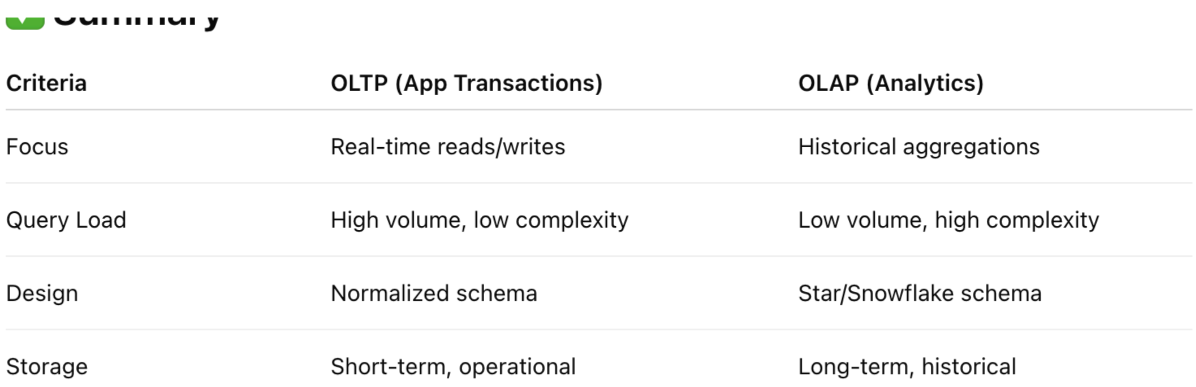
- Star schema is **faster**, Snowflake is **normalized**
- Choose schema based on **read performance vs storage cost**



⚖️ Trade-Off Summary

Criteria	Star Schema	Snowflake Schema
🔁 Joins Needed	Fewer (flat)	More (normalized)
⚡ Read Speed	Faster	Slower
💾 Storage Cost	Higher (duplication)	Lower (eliminates redundancy)
🛠️ Simplicity	Easier for analysts	Slightly more complex
🔐 Data Consistency	Lower	Higher

- **SCD type selection** is key in modeling slowly changing attributes
- Always consider use case: analytics (OLAP) vs app transactions (OLTP)



✅ Summary

Criteria	OLTP (App Transactions)	OLAP (Analytics)
Focus	Real-time reads/writes	Historical aggregations
Query Load	High volume, low complexity	Low volume, high complexity
Design	Normalized schema	Star/Snowflake schema
Storage	Short-term, operational	Long-term, historical