Banking Data Engineering – Data Modeling Reasoning (Interview Ready)

1. Why Star Schema?
In our banking project, we use a star schema to clearly separate transactional data from descriptive data. Transactions are high-volume events, while customer, merchant, and category data change slowly. This separation improves query performance, reduces data duplication, and makes analytics easier for business users.

2. Fact vs Dimension
Fact tables store measurable business events such as transactions and amounts. Dimension tables store descriptive attributes such as customer risk, merchant type, and category. In our project, transactions are stored in a fact table, while customers, merchants, and categories are modeled as dimensions.

3. Why Surrogate Keys?
We use surrogate keys instead of business keys because business keys can change across systems. Surrogate keys are stable, system-controlled identifiers that ensure reliable joins and historical tracking. This is a common enterprise banking practice to decouple analytics from source system volatility.

4. Why SCD Type 2 for Customer Dimension?
Customer attributes such as risk rating change over time. Using SCD Type 2 allows us to preserve history by maintaining effective start and end dates. This ensures point-in-time accuracy so historical reports remain correct and auditable.

5. Why Not SCD Type 1?
SCD Type 1 overwrites old data and loses history. In banking, historical accuracy is critical for compliance and audits. Therefore, SCD Type 2 is preferred to prevent retroactive changes to past reports.

6. Why Not Snowflake Schema?
Snowflake schemas increase join complexity and reduce query performance. In our project, dimensions are small and manageable, so a star schema provides better performance and simpler analytics. Banks prioritize fast and predictable reads over overly normalized designs.

7. How Facts Join to Dimensions
Facts join to dimensions using surrogate keys and, in the case of SCD dimensions, effective date ranges. This ensures each transaction is associated with the correct version of a dimension at the time of the event.

Summary
The data model in this project is designed for scalability, auditability, and analytical performance. Star schema, surrogate keys, and SCD Type 2 together form a robust modeling approach suitable for enterprise banking systems.