K21Academy
Learn Cloud from Experts

Top 100+ Data Modelling Interview Questions and Answers

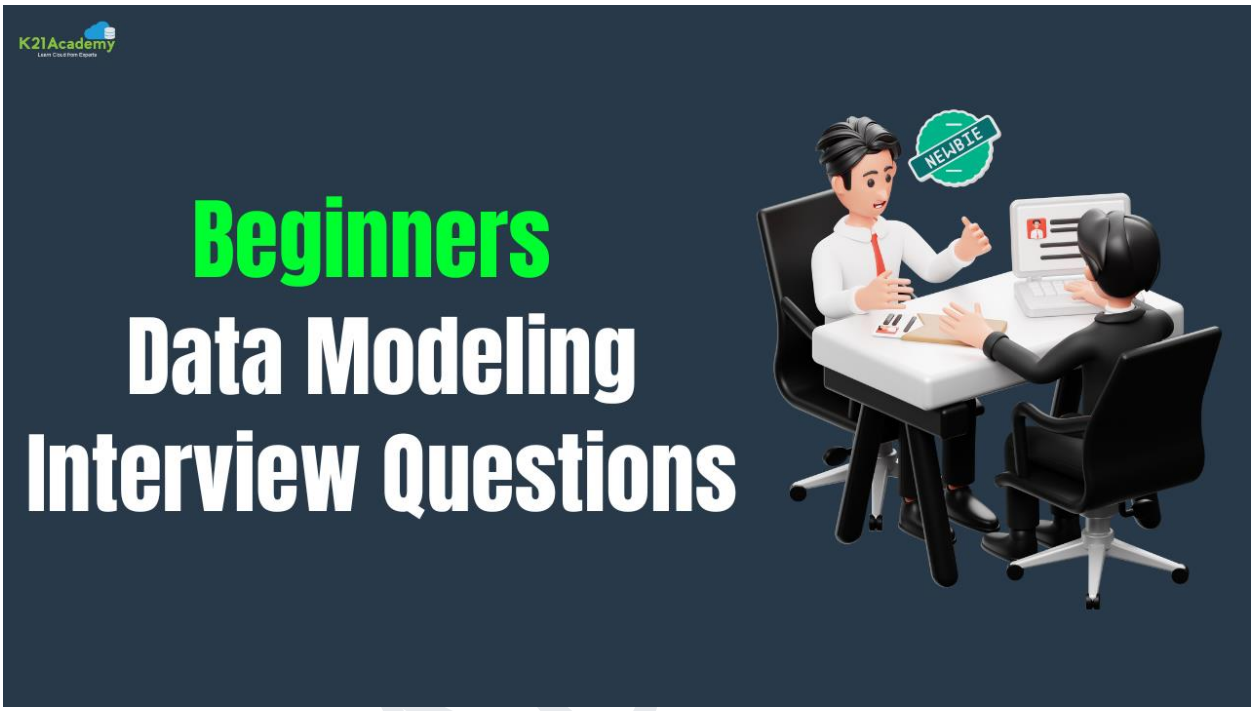
Atul Kumar
Author & Cloud Expert

[EDITION 01]

Table of Contents

1. Beginners Data Modeling Interview Questions

This section comprises some of the most popular beginner data modeling interview questions for anyone willing to start a career in data science. These questions are mainly relevant to the fundamental topics of data modeling.



Q1) Explain the three different types of data models.

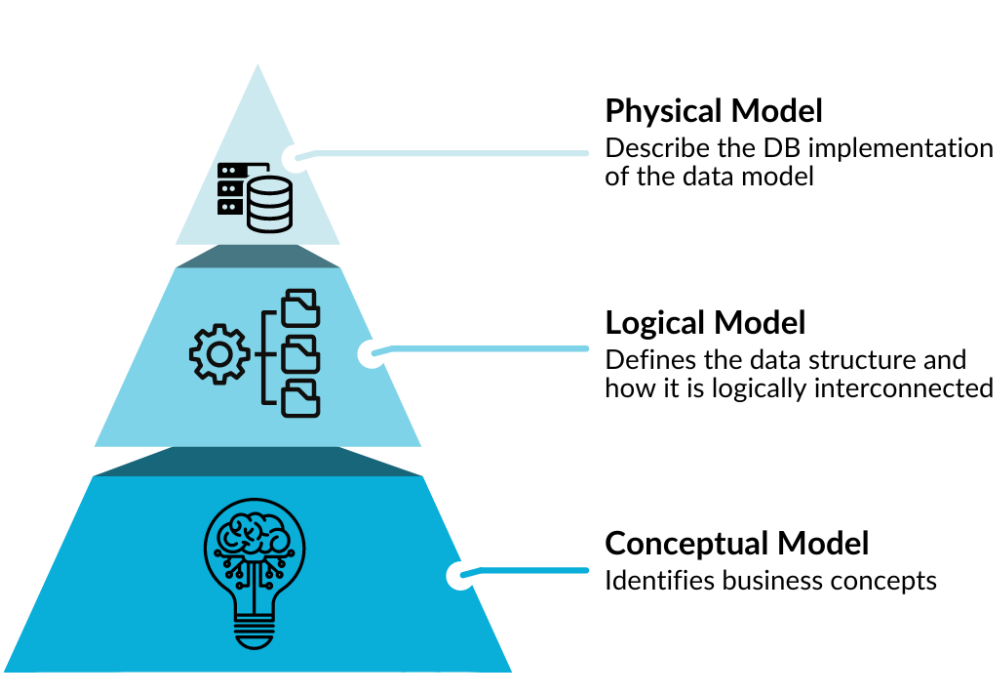
Ans) There are three data models-

- **Conceptual data model**:- The conceptual data model contains all key entities and relationships but does not contain any specific details on attributes. It is typically used during the initial planning stage. Data modelers construct a conceptual data model and pass it to the functional team for assessment. Conceptual data modeling refers to the process of creating conceptual data models.

- **Physical data model**:- The physical data model includes all necessary tables, columns, relationship constraints, and database attributes for physical database implementation. A physical model's key parameters include database performance, indexing approach, and physical storage. A table, which consists of rows and columns and is connected to other tables through relationships, is the most essential or principal object in a database. Physical data modeling is the process of creating physical data models.

- **Logical data model**:- A logical data model is a data model that defines an organization's business requirements (in its entirety or in part). This is the process of putting a

conceptual data model into action and extending it. Entity, Attribute, Super Type, Sub Type, Primary Key, Alternate Key, Inversion Key Entry, Rule, Relationship, Definition, etc., are all present in logical data models. The process of creating logical data models is known as logical data modeling.



Q2) How would you create a Data Model using SQL commands?

Ans) To build a data model, query the data with the SELECT statement and create the table structure with the CREATE TABLE statement. You can also use the INSERT command to fill your tables with data.

Q3) Differentiate between the logical data model and the physical data model.

Ans)

Logical Data Model	Physical Data Model
Entity	Table
Attribute	Column
Primary Key	Primary Key Constraint
Alternate Key	Unique Constraint
Inversion Key Entry	Non-Unique Indexes
Rule	Check Constraint, Default Value
Relationship	Foreign Key
Definition	Comment

Q4) What is the maximum number of null values that you can add to a column with a unique constraint?

Ans) In some databases, since one null value is not identical to another null value, numerous null values can be placed in a unique constraint field; in other databases, this is not the case.

Q5) What is the definition of a foreign key constraint?

Ans) A primary key constraint is applied to a column in the parent table, and a foreign key constraint is applied to a column in the child table. The child table's foreign key column value will always relate to the parent table's primary key values.

Q6) What are the benefits of employing data modeling techniques?

Ans) The following are some of the benefits of using data modeling in data warehousing:

- It facilitates the management of business data by standardizing and defining its properties.
- To eliminate data redundancy, data modeling brings together data from diverse systems.
- It allows for the creation of a database design that is both efficient and effective.
- Data modeling enables the organization's departments to work together as a unit.
- It makes data more accessible.

Q7) What does "data sparsity" imply?

Ans) The number of blank cells in a database is known as data sparsity. In a data model, it describes the amount of data that is available for a specific dimension. Due to a lack of information, saving aggregations consumes a lot of space.

Q8) What is the definition of a primary key?

Ans) A primary key is a column or set of columns in a relational database management system table that uniquely identifies each record. To avoid null values and duplicate entries, the primary key constraint is applied to the column data. Unique + Not Null is the primary key. For instance, a social security number, a bank account number, and a bank routing number are all examples of a primary key.

Q9) What is the goal of data modeling?

Ans) Data modeling is the process of defining the structure of data and how it will be stored in the memory. It depicts the system's objects and their relationships, as well as any rules that may apply to those within its bounds. Data modeling also enhances query performance by identifying which tables should be linked together.

Q10) What do you understand by tables in data modeling?

Ans) Tables are a form of database object that holds all of your table's data. Tables, like spreadsheets, logically organize information into rows and columns; each one constitutes a unique record with its own set of fields for storing any relevant details about them, including their ID values.

Q11) What do you understand by data models? What is data modeling?

Ans) A data model is a visual description of business needs (logical data model) or database objects (physical data model) needed for a database, and it's particularly useful for defining and interpreting such objectives and entities.

The practice of using words and symbols to represent data and how it flows in a basic representation of a software system and the data elements it contains is called data modeling. Overall, data modeling assists an organization in successfully using its data to satisfy business information needs.

Q12) Define normalization.

Ans) The method of effectively organizing data in a database is known as normalization. The normalization process helps in:

- removing redundant data (for example, storing data in multiple tables)
- ensuring data integrity.

Normalization is useful for minimizing data storage and logically storing data in multiple tables.

Q13) List some of the benefits of data modeling.

Ans) Professionals from several areas, such as software development, business analytics, data science, and others, use data modeling to visualize data and analyze the relationships between data items in a warehouse or database. The following are the key benefits of this process:

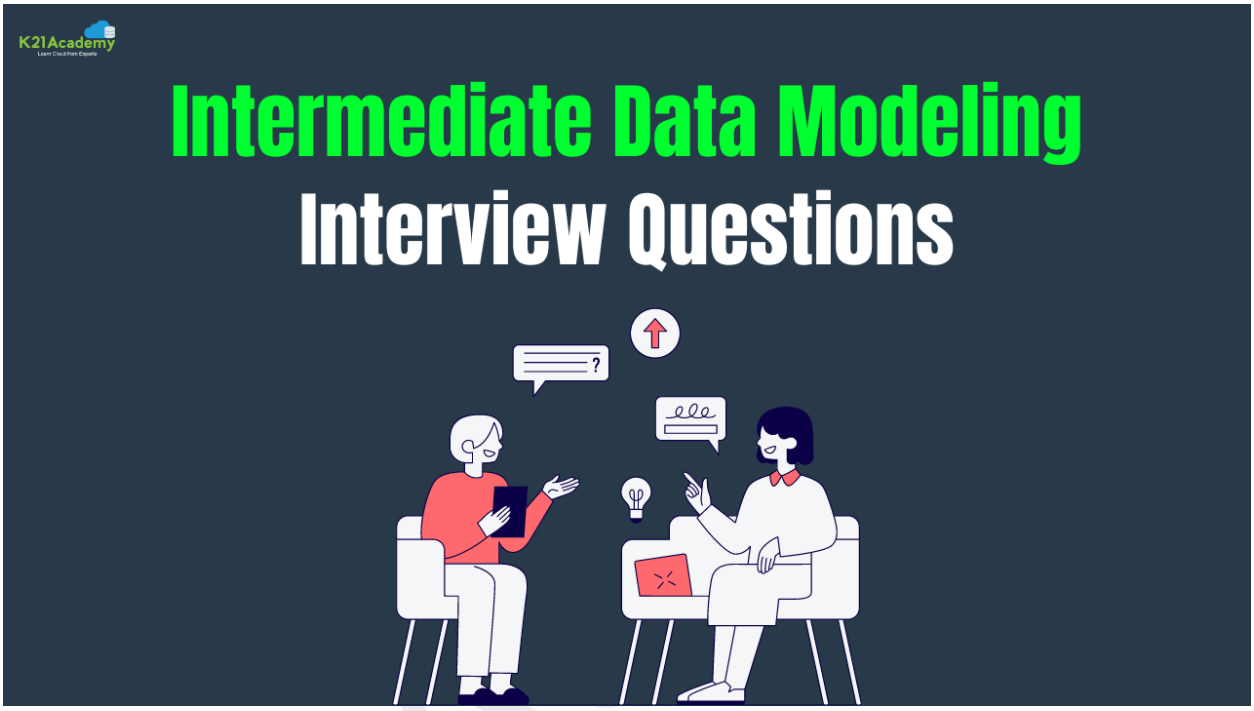
- **Faster product delivery:-** Data modeling allows application developers to focus on building products with minimal errors and meeting prior commitments without wasting time on additional data requirements. As a result, higher-quality products will be delivered sooner, the testing process will be easier, etc.

- **Lower development expenses:-** Data modeling detects errors and anomalies at the beginning of the project when they are easy and inexpensive to resolve.

- **Efficient performance:-** DBAs can use data modeling to analyze the database and configure it for optimal performance without having to sift through the code to find the schema.

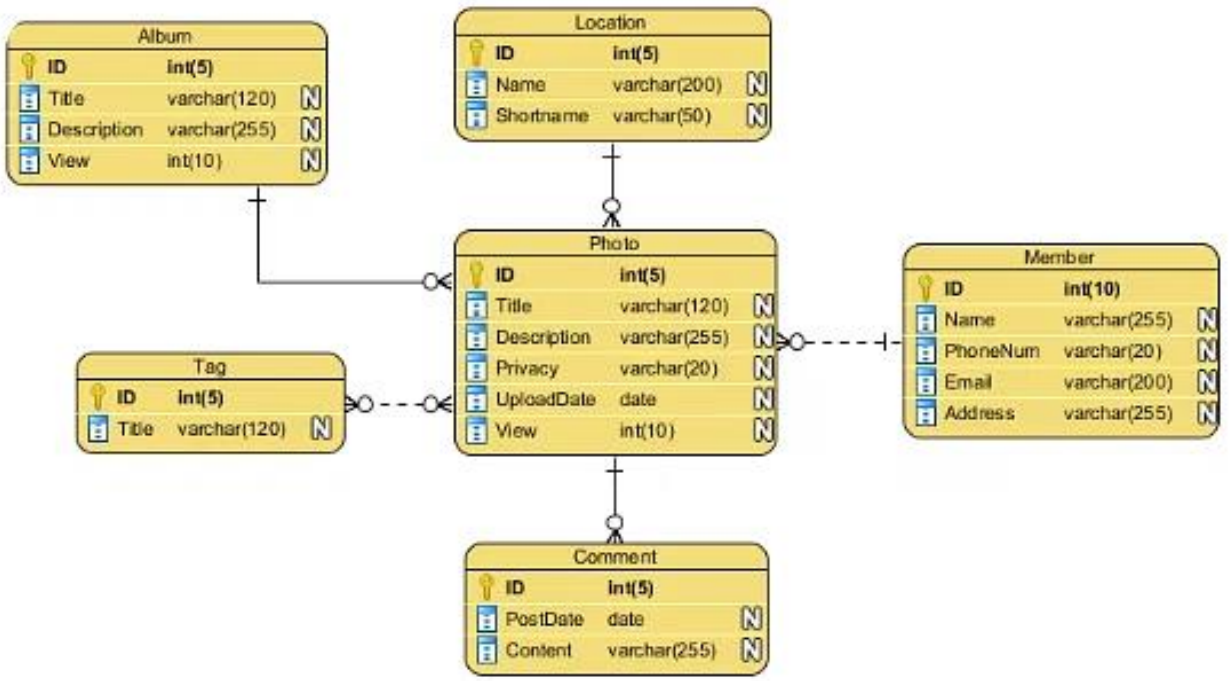
2. Intermediate Data Modeling Interview Questions

If you have been working in the data science industry for a few years and are planning to grab that next promotion in your office, then explore these intermediate-level data modeling interview questions.



Q14) Explain the ER (entity-relationship) diagram or ERD with an example.

Ans) An ER diagram is a graphical presentation of entities and their connections. Entities (tables) in a data model resemble a square or rectangular box that contains attributes and is connected by lines (relationship).



Q15) Explain the ER (entity-relationship) diagram or ERD with an example.

Ans) The following are the most prevalent data modeling errors:

- **Creating data models that are excessively broad:** When the number of tables in a table exceeds 200, the data model gets highly complex. As a result, the possibility of failure increases.

- **Unnecessary surrogate key:** A surrogate key is only necessary when the natural key is unable to perform its function as the primary key.

- **Building a data model with no purpose:** Sometimes, the user possesses no understanding of what the company's aim or mission is. Creating a specialized business model might be challenging if the modeler lacks the necessary knowledge.

Q16) Explain the concepts of subtypes and supertypes.

Ans)Entities can be split into sub-entities and classified according to their characteristics. Each sub-entity, also known as a subtype entity, has its own set of properties.

Some qualities are unique to each entity and are only found in higher or super-level entities. This is why they are referred to as supertype entities.

Q17) What is the role of an Erwin Data Modeler?

Ans)

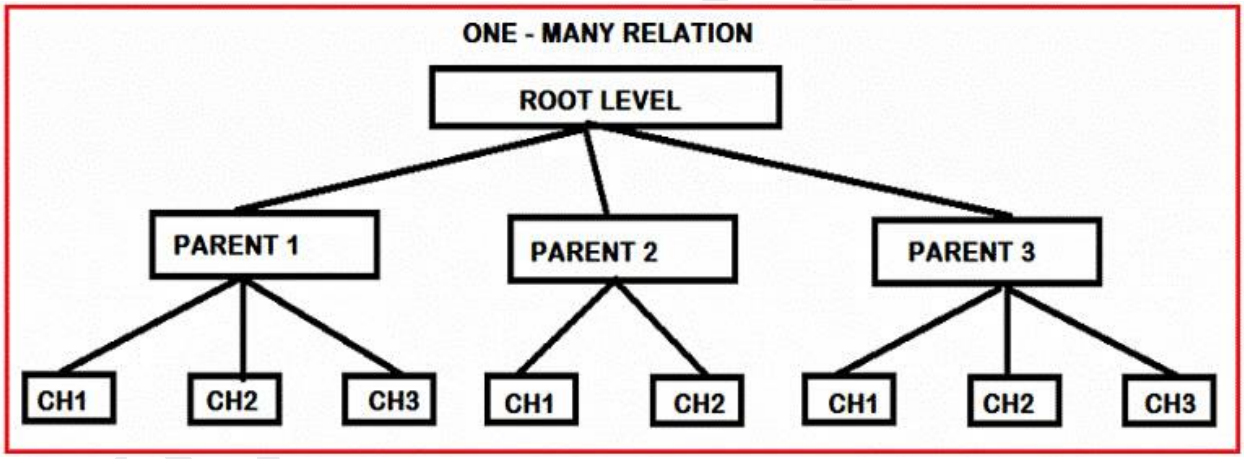
- Erwin Data Modeler (Erwin DM) is one of the data modeling tools that enable conceptual, logical, and physical models to assist management and technical users in the design of information systems and the databases that support them.

- To enhance business alignment, maintain data quality, and facilitate integration, Erwin Data Modeler streamlines and standardizes model design tasks, including complicated queries.

- Consolidate and develop hybrid architectures in the cloud and on-premises, combining conventional, NoSQL, and Big Data.

Q18) What is a hierarchical database management system (DBMS)?

Ans) A hierarchical database is a data model in which data is sorted into a tree-like structure, or parent-child hierarchy, in which one parent node can have numerous child nodes related through links.



A hierarchical database's key feature is its ease. The one-to-many data format makes database navigation simple and quick, which is appropriate for use cases like website drop-down menus or system directories in Microsoft Windows OS.

Q19) List a detailed comparison between a fact table and a dimension table.
Fact Table

Ans)

Fact Table	Dimension Table
A fact table stores the measurements together with the properties of dimension tables.	The qualities along which the fact table calculates the metric are stored in the dimension tables.
There are fewer attributes in a fact table, but there are more records.	The dimension table has fewer records and more attributes.
The primary key of a fact table is a concatenation of the primary keys of all dimension tables.	Dimension tables have individual primary keys.
You can create fact tables only when a dimension table is complete.	You must create dimension tables first, i.e., before fact tables.
There are fewer fact tables in a schema.	There are more dimension tables in a schema.
The data in a fact table can be numerical or textual.	A dimension table always has textual attributes only.

Q20) Briefly define factless fact tables in data modeling.

Ans)
A factless fact table does not include any facts. They only have dimensional keys, and they capture events that occur only at the information level, not at the computation level (just information about an event that happens over a period). The many-to-many links between dimensions are captured in a factless fact table, containing no numeric or textual facts. They're commonly used to document events or information about coverage.

Factless fact tables are useful for tracking a process or collecting data. There are two types of factless fact tables: one that describes occurrences and one that describes conditions.

Q21) What do you understand by a data mart? Discuss the various types of data marts available in data modeling.

Ans)
A data mart is a subject-oriented database that is usually a split portion of a larger warehouse. The data subset in a data mart usually relates to a certain business area, such as sales, finance, or marketing. Data marts speed up business operations by allowing users to access essential data from a warehouse or operational data store in very little time. A data mart is a cost-effective solution to efficiently acquire meaningful insights since it only comprises data relevant to a specific business area.

In data modeling, there are three types of data marts depending on their relationship to the data warehouse and the data sources-

- **Dependent Data Mart:-** An existing enterprise data warehouse helps establish a dependent data mart. It's a top-down method that starts with storing all enterprise data in a single area, and then extracting a well-defined piece of the data when it's time to analyze it.

- **Independent Data Mart:-** An independent data mart is a stand-alone system that concentrates on a single specific subject or business operation without the usage of a warehouse. These data marts store processed data from internal or external data sources (or both) until it is required for business analytics.

- **Hybrid Data Mart:-** A hybrid data mart is an integration of an existing warehouse and extra functional data systems. It combines the features of the bottom-up method's enterprise-level integration with the speed and end-user emphasis of a top-down approach.

Q22) Give a brief overview of the Critical Success Factor and its four types.

Ans)
A critical success factor (CSF) is a certain aspect or element that a team, department, or company must effectively implement and emphasize to meet its strategic objectives. CSFs lead to an increased value for products and services and beneficial outcomes. The significance of crucial success factors stems from the fact that they serve as a roadmap for a company. The only way to know what the outputs entail is to analyze and define a critical success factor; otherwise, they stay fictional.

Here's a breakdown of critical success factors-

- **Environmental CSFs:-** These are external elements that a company has no direct influence over, such as government policy, the economy, new technologies, and opponent activities. Business leaders must stay aware of and forecast changes to stay ahead of the curve.

- **Industry CSFs:-** Each company must perform certain tasks to stay competitive in its respective industry. These duties must be recognized by carefully examining the industry's affecting elements.

- **Temporal CSFs:-** The organization's permanent or long-term strategic goals are the most important critical success factors in strategic management. However, businesses must occasionally concentrate on and efficiently manage temporary challenges. Due to the COVID-19 epidemic, several companies had to increase their shipping and delivery processes temporarily.

- **Strategy CSFs:-** To create brand loyalty and sustain their market position, an industry leader will concentrate on crucial success factors in strategic management. A minor player will have to concentrate on CSF to improve their competitiveness. Peer orientation has a significant impact on organizational behavior.

Q23) What is a surrogate key? List a few benefits of using surrogate keys in relational databases.

Ans)
Surrogate keys are a form of primary key and are present in almost all relational database tables. It creates a simple, system-generated column that is useful for data analysis. A relational data model uses this column to identify individual rows instead of depending on existing data attributes.

Benefits of using surrogate keys in relational data modeling-

- Surrogate keys are system-generated and unique, which makes it difficult for the system to create and record a duplicate value.
- Surrogate keys follow a standard format since they are usually generated automatically.
- Surrogate keys can have any number of values.

Q24) What are some of the benefits of a data model?

Ans)
The following are some of the benefits of a data model-

Consistent representation:- A data model ensures whether the represented data objects are accurate.

Relationship between data objects:- A data model builds robust links between tables, stored procedures, primary keys, and foreign keys in different data objects.

Determine Data Sources:- A data model enables the detection of data sources to build the model.

Q25) What do you mean by dimensional modeling?

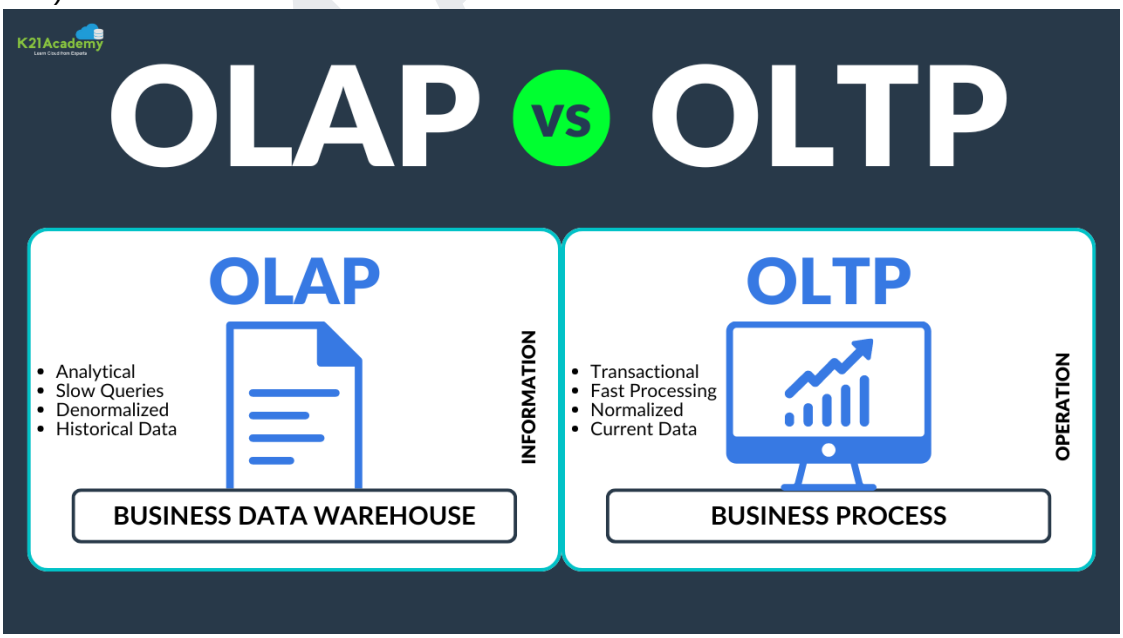
Ans)
Dimensional Modeling (DM) is a data model approach that is specifically useful for data storage in a warehouse. The goal of dimensional modeling is to make the database more efficient at retrieving data. The benefit of implementing a dimension model is that it allows us to store data in a manner that makes it easier to store and retrieve data once it has been saved in a warehouse. Many OLAP systems employ a dimensional model as their data model.

3. Advanced Interview Questions on Data Modeling

Are you a senior professional willing to take your career to the next level in the industry? The good news is we have some popular advanced data modeling interview questions for you!

Q26) Differentiate between OLAP and OLTP databases.

Ans)



Database SQL workloads can be divided into two categories:

Online-transactional processing (OLTP): These are simple queries with high concurrency and low latency that read or change a few records simultaneously while ensuring data integrity, such as bank account transactions.

Online Analytical Processing (OLAP): These workloads, similar to periodic reporting, are often complicated queries (including aggregates and joins) that necessitate scanning a large number of records.

Online-transactional processing (OLTP)	Online Analytical Processing (OLAP)
OLTP databases enable normalization in tables.	OLAP databases do not allow normalization in tables.
OLTP supports the traditional database management systems approach.	OLAP supports the data warehouse approach
OLTP has a response time of milliseconds.	OLAP has a response time of seconds to minutes.
OLTP databases comprise several quick online transactions.	OLAP databases comprise large data volumes.
OLTP is a real-time transaction processing system.	OLAP is a database that allows you to analyze business metrics by category and attribute.

Q27) What do you mean by the CAP Theorem? How does it work?

Ans)
The CAP theorem shows that no distributed system can ensure C, A, and P at the same time. To put it another way, it states that a distributed system cannot deliver over two of the three assurances.

Consistency:- After an activity, the data should stay consistent. After upgrading the database, for example, all of our queries should return the same result.

Availability:- There should be no downtime with the database; it should always be accessible and active.

Partition Tolerance:- Even if communication between the servers is not consistent, the system should remain functional.

Q28) What is metadata and why is it important?

Ans)
Metadata is data-related information that identifies the type of data stored in the system and its purpose and intended audience. Depending on the objective, there are various forms of metadata, including

Technical Metadata:- This type specifies the database system and tables' names, table sizes, values, attributes, data types, etc. It also contains constraint details such as primary, foreign, and indexes.

Business Metadata:- This data is business-specific and identifies data rights, business norms and standards, policies, and so on.

Descriptive Metadata:- Descriptive Metadata delivers information on a folder, file, image, book, or movie. Title, date, size, author, etc., are all examples of information.

Operational Metadata:- This type of metadata contains information on any corporate operation and is needed by managers and executives to accomplish any activity.

Q29) What are the different critical relationship types in a data model?

Ans) A relationship mainly connects parent and child tables. In a data model, there are three different types of critical relationships:

Identifying:- In this type, the primary keys of the parent tables include a reference column of child tables, which is connected by a thick line and helps identify the entries. A thick line is frequently used to replace this relationship. Because the foreign key forms part of the primary key, they are referred to as identifying keys. There can't be a child object if there isn't a parent object.

Non-identifying:- A child table's reference field does not exist in the parent table's primary key. A dotted line represents this relationship. Depending on the situation, this relationship can be optional or mandatory. It signifies that NULL can be allowed (if optional) or not allowed (if obligatory) in foreign key columns.

Self-recursive:- In this case, a single column is linked to a table's primary key. This is a relationship that exists between different column objects belonging to a similar entity.

Q30) Explain the different types of dimensions with examples.

Ans) There are usually five different categories of dimensions-

A. **Conformed dimensions:-** A conformed dimension is used in multiple domains. You can use it with multiple fact tables in a particular database or even across multiple data marts/warehouses.

B. **Role-Playing Dimensions:-** These are the dimensions in a database that serve various purposes.

C. **Slowly Changing Dimension (SCD):-** SCDs are the most essential dimensions. These are the dimensions where the value of an attribute changes over time. SCDs are classified into five categories: type 0, type 1, type 2, type 3, and type 4.

D. **Junk Dimension:-** This is a dimension table that contains attributes that don't belong in the fact table or any of the existing dimension tables. These dimensions have typical attributes such as flags or indications.

E. **Degenerated Dimension:-** A degenerated dimension is not a fact yet appears as a primary key in the fact table. It doesn't have its own set of dimensions. A single attribute dimension table is another name for it.

4. Data Modeling Interview Questions for Experienced

Q31) Differentiate between primary and foreign keys.

Primary Key	Foreign Key
A clustered index serves as the primary key and allows physical data sorting in the DBMS table according to the clustered index's sequence.	**A foreign key cannot generate an index, clustered or non-clustered, independently. You can, however, establish an index on the foreign key individually.**
The primary key allows you to identify a record in the database by its unique ID.	**A foreign key is a table field that is the primary key of another table.**
Null values are never acceptable in the primary key.	**You can add null values in the foreign key.**
There can be only one primary key in a table.	**You can have multiple foreign keys in a single table.**

Q32) How many types of normalization exist? Mention the rules for the second normal form and third normal form.

Ans)
Normalizations are classified as follows:

1) first normal form,
2) second normal form,
3) third normal form,
4) Boyce-Codd fourth normal form, and
5) fifth normal form.

Second-normal form rules are as follows:

- It must already be in the first normal form.
- It doesn't have any non-prime attributes that are functionally dependent on any subset of the table relation's candidate keys.

Third-normal-form rules are as follows:

- It must already be in the second normal form.
- There are no transitive functional dependencies.

Q33) Describe the features of a database management system.

Ans)

- A multi-user environment is enabled by DBMS, allowing users to explore and process the data simultaneously.
- It adheres to the ACID principle (Atomicity, Consistency, Isolation, and Durability).
- It ensures security and eliminates redundancy.
- It enables multiple data views.
- Tables are built by DBMS by combining entities and their relationships.
- Data sharing and multiuser transaction processing are made easier using DBMS.

34) List the differences between a data mart and a data warehouse.

Ans)

Data Mart	Data Warehouse
A data mart concentrates on a specific business subject.	The focus of a data warehouse is on various business areas.
It helps to make strategic business decisions.	It aids the strategic decision-making of business owners
The bottom-up approach is used in a data mart.	The data warehouse is organized in a top-down manner.
Data comes from a single data source.	The data source is made up of multiple heterogeneous data sources.

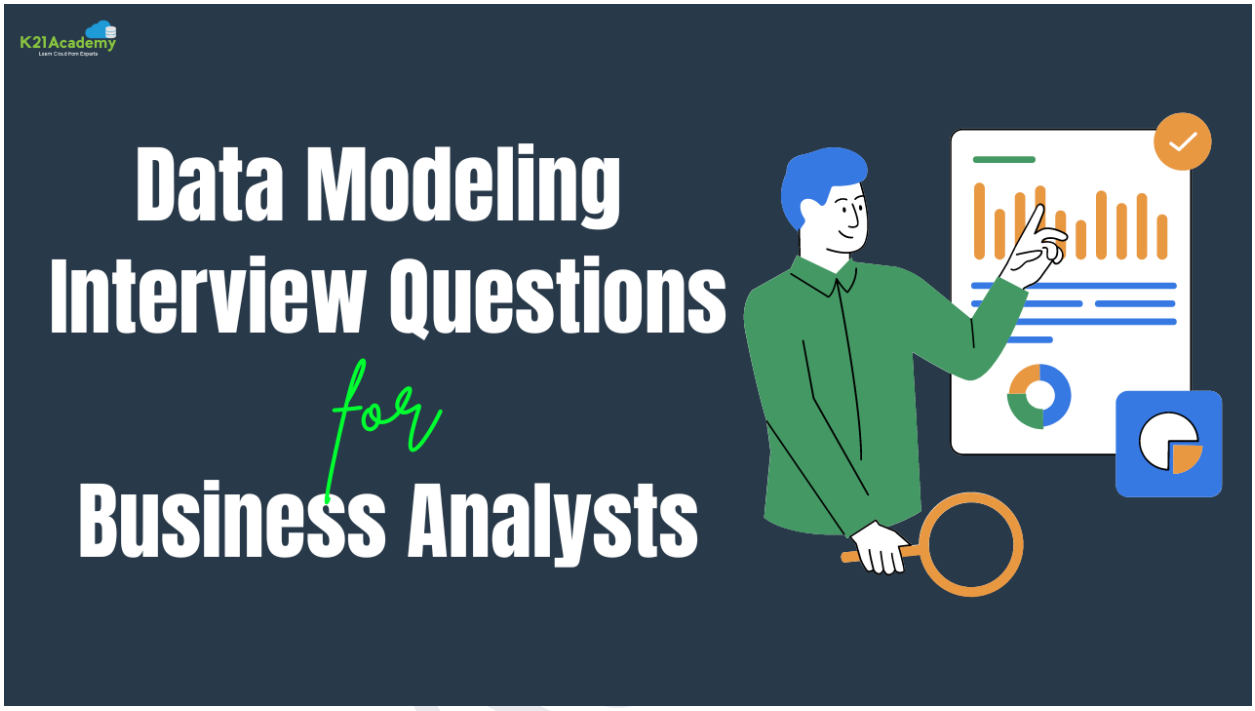
35) How does the Star Schema differ from the Snowflake Schema?

Ans)

Star Schema	Snowflake Schema
You can use Star Join Query Optimization for efficient query performance. Also, you can use multiple dimensions to connect tables.	A centralized fact table represents the Snowflake Schema with no numerous dimensions attached to it.
You can establish the link between the fact table and any dimension table in a star schema by using only a single join.	You will require numerous joins to retrieve data from a snowflake schema.
It has a fact table with dimension tables surrounding it.	A fact table is encircled by a dimension table, which is encircled by another dimension table.
The dimensional tables keep track of the hierarchy of the dimensions.	Each hierarchy requires an individual table.
Star schema involves a denormalized data structure and also enables faster query execution.	The Snowflake schema features a normalized data structure.
It offers a basic database design and also, high data redundancy.	It has a complicated database design along with a low level of data redundancy

5. Data Modeling Interview Questions for Business Analysts

Planning to switch your career and become a business analyst? Well well! Don't worry. We have some typical data modeling interview questions for business analysts that will ease your job transition!



Q36) What do you mean by a network model?

Ans)
Similar to a hierarchical model, the network model permits multiple relationships to link records, implying that it has more than one relationship. It allows you to create a set of parent and child entries. Each record can belong to several sets, which allows you to create complicated table relationships.

Q37) Mention some of the fundamental data models.

Ans)
Fully-Attributed (FA):- This is a third normal form model that provides all of the data for a specific implementation approach.

Transformation Model (TM):- Specifies the transformation of a relational data model into a structure that is suitable for the DBMS in use. The TM is no more in the third normal form in the vast majority of cases. The structures are optimized depending on the DBMS's capabilities, data levels, and projected data access patterns.

DBMS Model:- The DBMS Model contains the database design for the system. For the complete integrated system, the DBMS Model can be at the project or area level.

Q38) Define an entity and an attribute.

Ans)
Entity:- An entity is a collection of real or abstract objects (people, locations, events, etc.) that share common features or properties. Entities can be self-contained or interdependent.

Attribute:- A form of trait or property associated with a collection of real or abstract elements (people, places, events, etc.).

Q39) Briefly define Normalization and the three normal forms- 1NF, 2NF, and 3NF.

Ans) In relational database design, normalization is the process of organizing data in a relational structure to reduce redundancy and non-relational structures. By deleting any model structures that allow various methods to know the same information, you can control and eliminate data redundancy by following the normalization criteria.

First Normal Form (1NF):- The entity E is in first normal form (INF) if and only if all underlying values include only atomic values. You must remove any recurring sets (such as those found in legacy COBOL data structures).

Second Normal Form (2NF):- If an entity E is in 1NF and every non-key attribute depends entirely on the primary key, it is in 2NF. In other words, there are no partial key dependencies; rather, dependence is on the full key K of E, not on a specific subset of K.

Third Normal Form (3NF):- If an entity E is in 2NF and no non-key attribute of E depends on another non-key attribute, it is in 3NF. Another way to think about it is that if an entity E is in 2NF and every non-key attribute is non-transitive relying on the main key, it is in 3NF.

Q40) What are the various types of measures available for fact tables?

Ans) There are mainly three different types of measures available-

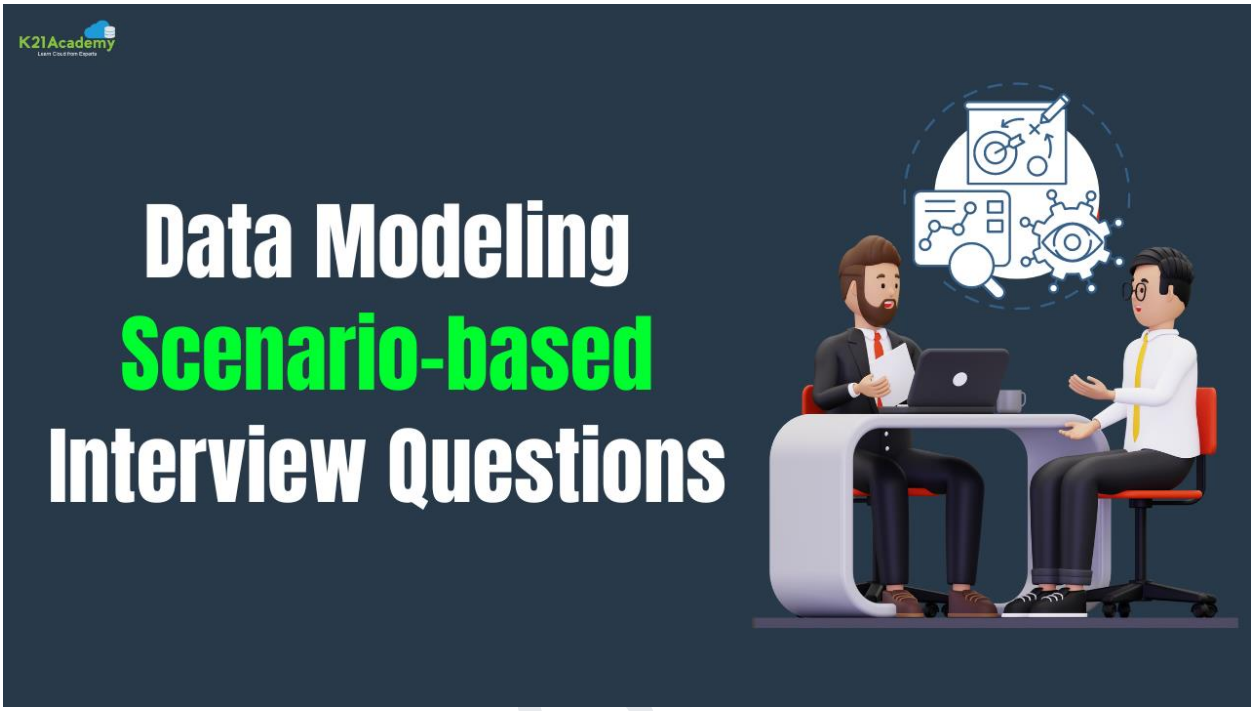
Additive measures:- The most popular type of measure is additive, which can be aggregated throughout all of the dimensions in the fact table. Additive measures are used to sum data across multiple dimensions.

Semi-additive measures:- Semi-additive measures are the ones you can aggregate across some dimensions, but not all. Semi-additive measures, for example, include head counts and inventories.

Non-additive measures:- You cannot aggregate non-additive measures over any of the dimensions. You can't logically combine these measures across records or fact rows. Non-additive measures are mostly calculated using ratios or other mathematical methods. For such a measure, the only analysis possible is to count the number of rows.

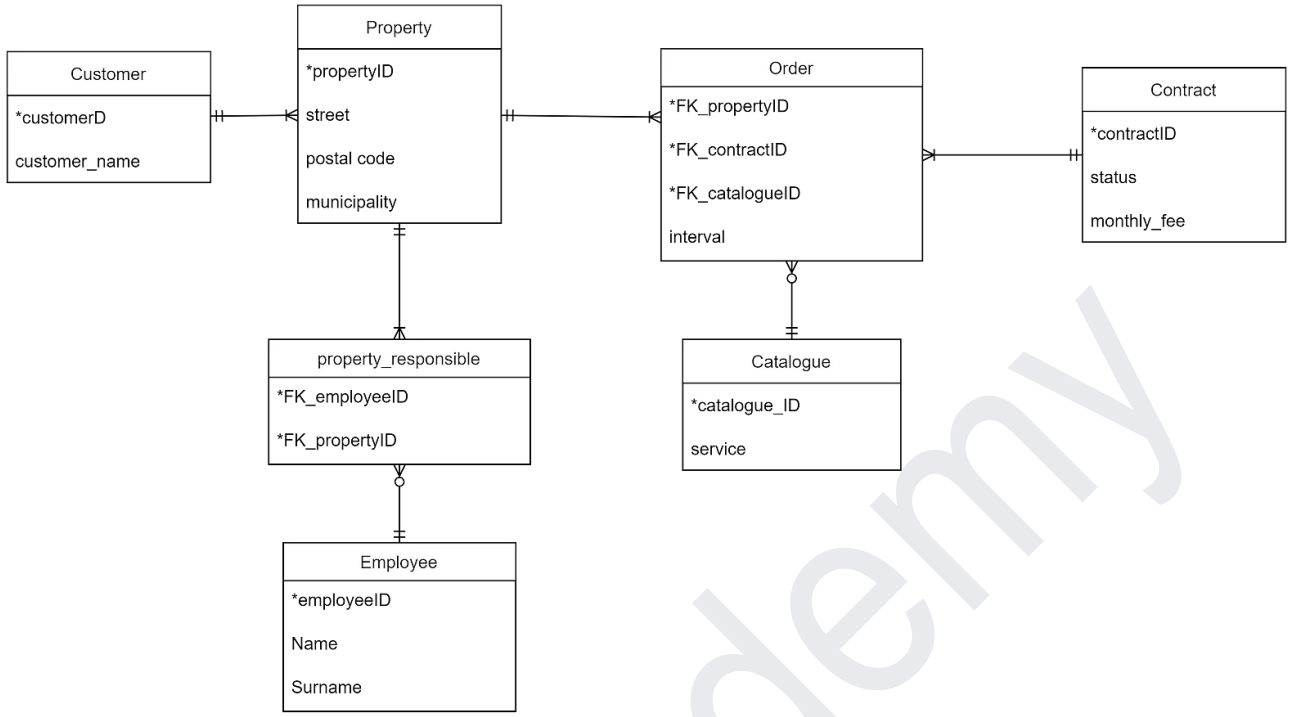
6. Scenario-based Data Modeling Interview Questions

Here are some examples of scenario-based interview questions to help you enhance your skills and gain expertise



Q41) Lisa is the owner of CleanAlp, a cleaning service. Property management businesses, that maintain residential complexes, are her company's primary clients (properties). For each property, clients book one or more services. CleanAlp and the respective customer sign a contract for this purpose. In some situations, there may be multiple contracts for the same property (for example, following termination and a new assignment) or temporary additional cleaning services. Every cleaning service has an agreed-upon period (weekly, fortnightly, or monthly) in the contract. The term "cleaning service" refers to a catalog of Cleanalp's cleaning services (e.g. sweeping the stairwell, cleaning windows). Lisa has a large workforce. Cleaning services are carried out by a single employee for each house. During the contract period, the responsible person can change, but it's important to keep track of who was liable and when. Customers receive monthly invoices detailing the amount due, as specified in the contract. Build an ER diagram for the given scenario.

Ans)



Q42. How do you model a set of entities in a NoSQL database using an optimal technique?

Ans)

Begin with a 3NF model in which the entity "Order" has the main key "order id" and the entity "Detail" has foreign keys:

order_id	order_date	customer_id	detail_id	item_id	qty	unit_price	sales_total
12345	1/21/2019	123456	12001	2345	2	157.75	315.5
12345	1/21/2019	123456	12002	2110	1	75.25	75.25
12345	1/21/2019	123456	12003	1760	3	55	165

The above 3NF Model looks like this after denormalization:

Order

order_id	order_date	customer_id
12345	1/21/2019	123456

Detail

detail_id	item_id	unit_price	qty	sales_total	order_id
12001	2345	157.75	2	315.5	12345
12002	2110	75.25	1	75.25	12345
12003	1760	55	3	165	12345

Finally, you can turn the denormalized model into a NoSQL entity model with embedded parent-child relationships:

```
{
   "order_id": "12345",
   "order_date": "1/21/2019",
   "customer_id": "123456",
   "order_details": [
      {
         "item_id": "2345",
         "qty": 2,
         "unit_price": 157.75,
         "sales_price": 315.5
      },
      {
         "item_id": "2110",
         "qty": 1,
         "unit_price": 75.25,
         "sales_price": 75.25
      },
      {
         "item_id": "1760",
         "qty": 3,
         "unit_price": 55,
         "sales_price": 165
      }
   ]
}
```

7. Data Modeling Interview Questions Amazon

Worrying about your Data Scientist dream job at Amazon? It is time for you to sit back and relax! Meanwhile, don't forget to go through these data modeling interview questions from Amazon.

Q43) What is Amazon's RDBMS service?

Ans) The Amazon Relational Database Service is a web service for setting up, managing, and scaling a relational database in the cloud. Standard relational database engines are supported by Amazon RDS, which is controlled, extended, and available on demand. RDS handles time-consuming management operations, allowing you to focus on your application rather than your database.

Q44) What is Amazon Aurora?

Ans) Amazon Aurora is a high-availability, automated failover relational database engine that supports MySQL and PostgreSQL. To put it another way, Amazon Aurora is a hybrid of MySQL and Postgres. Aurora is a multi-threaded, multiprocessor database engine that prioritizes performance, availability, and operational efficiency.

Amazon Aurora differs from typical MySQL and Postgres database engines in that it does not employ a Write-Ahead Log, but it does feature a crash recovery scheme. Amazon also provides

the feature to duplicate data from an Amazon Aurora database to a MySQL or Postgres database on the same instance, as well as vice versa, allowing you to employ the two databases together.

Q45) What is the function of Amazon Redshift?

Ans) Amazon Redshift is a cloud-based data warehousing solution that is quick, fully managed, and extends to petabytes.

It enables you to examine all of your data quickly and effectively using your existing business intelligence tools.

Amazon Redshift allows you to access data using conventional SQL and integrates with a variety of business intelligence applications, including Tableau, MicroStrategy, QlikView, and others.

Q46) What is the difference between Amazon DynamoDB and other NoSQL databases?

Ans) Amazon DynamoDB is a cloud-based database with no physical location. It is scalable enough to suit the needs of high-performance applications. You can set up security, back up all in case of a crash or issue, save data in various locations, and even export all of your data if you want to.

Amazon DynamoDB is a hosted database for applications that require regular updates without the assistance of a database administrator. DynamoDB takes care of all database administration on its servers, minimizing the need for manual tasks like backups and replication.

Q47) Is MongoDB better than PostgreSQL in terms of performance?

Ans) It's difficult to determine whether MongoDB is significantly faster than PostgreSQL since database performance depends on numerous parameters. SQL is a Relational Database, while MongoDB is a Document Database. They significantly differ in terms of design and function. Although MongoDB is preferable for keeping unstructured data, it's better to use PostgreSQL for structured data. MongoDB is more user-friendly, but PostgreSQL is more reliable. Both are free and open-source software.

8. Facebook Data Modeling Interview Questions

Q48) What significance does the third normal form have?

Ans) When a relation does not include transitive dependencies for non-prime attributes as they do in the second normal form, it reaches the third normal form. The third normal form (3NF) mainly eliminates data duplication and anomalies.

Q49) Define Recursive Relationships.

Ans) When an entity holds a relationship with itself, recursive relationships develop. These complex relationships demand more advanced data conversion methods into a schema. Think about a situation where a doctor is listed as a care provider in the medical database. Now, if a doctor is sick, he must visit another doctor as a patient, which creates recursive relationships. This is handled by adding a foreign key to the health center number in each patient's file. In such entity relations, care must be taken to ensure the recursion has a way out.

Q50) What happens if we attempt to insert two nulls into a column that has a unique constraint on it?

Ans) Inserting two null values in a column with a unique constraint does not result in an error because a null value cannot be equal to another null value and only inserts the data.

9. Accenture Data Modeling Interview Questions

Q51) What types of techniques are available for visual data modeling?

Ans) There are two different categories of visual techniques used in data modeling.

Entity-Relationship (ER) Model:

This data modeling technique helps build conventional databases and supports data normalization by minimizing data redundancy. It helps in the visualization of top-level data views in database structure.

UML (Unified Modeling Language):

UML is a general-purpose language used in software engineering for database creation, visualization, and modeling. It consists of several diagrams depicting the software systems, with the Class Diagram being one of them, and class diagrams and ER diagrams are very similar.

Q52) What exactly is an artificial (derived) primary key? When should you use it?

Ans) A derived key is a key you build up on purpose. A natural key is already present in the database. It is against the stability rule to use a name as the primary key. The social security number might be an appropriate option, but a foreign employee might not have one. This is where a derived primary key is preferable to a natural one.

Q53) Explain the differences between the DataStage Server's Hashed file stage and the Sequential file stage.

Ans) We use a hashed file on the Datastage server to store information on the hash algorithms and the hash key value. On the other hand, a sequential file lacks a key value for data storage.

Reading or writing data from one or more flat files becomes easier using a sequential file stage. The hashed file stage is helpful for data extraction within a DataStage job. Many inputs or outputs may be present at each stage of the hashed file process.

10. SAP Hana Data Modeling Interview Questions

Q54) Are aggregates or indexes necessary for SAP HANA?

Ans) There is no need for separate main indexes in SAP HANA since every column is stored as an index. For OLTP applications like the Business Suite, secondary indexes with multiple columns are feasible and used. To make multi-column joins effective, HANA will additionally self-generate helper indexes.

Q55) What license keys exist in the HANA system?

Ans) Temporary License key- Setting up the HANA database automatically allows you to install the temporary license keys. These keys are valid for 90 days after installation, so you should get permanent license keys from the SAP Market Place before that period expires.

Permanent License Key- Permanent License keys remain valid until the specified expiration date. The license keys specify the memory allotted for the intended HANA installation.

Q56) What does a restricted user in the SAP HANA system mean?

Ans) Users who access the HANA system through specific applications and do not have access to the SQL database are known as restricted users. The HANA database's restricted users cannot create new objects or modify their Schemas. Since their profiles don't have the generic Public role assigned like regular users, they cannot read any data in the database.

11. Salesforce Data Modeling Interview Questions

Q57) Define the term Data Skew.

Ans) When working with a large business with more than 10,000 records, you will come across the issue of data skew. Experts refer to a situation as "ownership data skew" when a single person owns many records. Due to "data skew," these users will have performance issues when performing updates. This occurs when the majority of the records for a specific object are owned by a single user or members of a single role.

Q58) What happens to the detail record after deleting the master record? What happens to the child record when the parent record is removed?

Ans) When a master record is deleted in a master-detail relationship, the detail record is also immediately deleted (Cascade delete).

The child record in a Lookup relationship won't be removed, even if the parent record is.

Lookup vs Master Detail Relationships

Lookup Relationships	Master Detail Relationships
Loose Relationships	Tight Relationships
Parent Not Required	Parent Record Always Required for Creation
Independent Sharing & Security	Parent Access Controls Child Access
Independent Owner	Parent Owner
Deletion Removes Value in Lookup Field	Parent Deletion results in Child Deletion
Lookup Field Optional	Lookup Field Required on Page
Roll Up Summary Fields Not Possible	Roll Up Summary Fields Possible
Standard Object Can be On Many sides	Standard Object Cannot be Child
Maximum of 40 per Object	Maximum of 2 per Object

Q59) Is it possible to have a roll-up summary field in a master-detail relationship?

Ans) Yes. A roll-up summary is an option if there is a master-detail relationship. However, a lookup relationship is an exception.

12. Oracle Data Modeling Interview Questions and Answers

Q60) What is the relation between a database, a tablespace, and a data file?

Ans) A tablespace refers to a logical storage unit in an Oracle database. In the Oracle database, each tablespace includes one or more files, known as data files. The entire database's data is stored across these tablespaces. The datafiles are the physical structure that validates the operating system that Oracle software is running.

Q61) What exactly does the ANALYZE command do?

Ans) You can modify an index, table, or cluster using the "Analyze" command. The analyze command helps find rows in a table or cluster chained and migrated. It is used to validate the object's structure. Additionally, this command helps gather statistics about the user-used object, which are subsequently added to the data dictionary.

Q62) What do you mean by mirroring Redo Log files?

Ans) Creating a replica of Redo log files is known as mirroring. This is executed by compiling a set of log files, ensuring that the current online redo log group's members will receive it automatically from the LGWR. When a group fails, the database immediately moves on to the next group, which degrades database performance.

13. SQL Data Modeling Interview Questions

Q63) Name some popular DBMS software.

Ans) The following are some of the popular DBMS software:

- MySQL
- PostgreSQL
- Dbase
- Microsoft SQL Server
- FoxPro
- Microsoft Access
- Oracle
- SQLite
- IBM DB2

Q64) Define an Apex transaction.

Ans) An Apex transaction is a set of actions that execute together. The DML activities in charge of querying records are included in this list of actions. If even one record-saving error happens during a transaction's DML processes, the entire transaction is either successful or rolled back.

Q65) Briefly define a NoSQL database.

Ans) A NoSQL database is a unique database that does not need a fixed schema. To achieve greater efficiency, it is non-relational and non-relational and typically minimizes joins in preference of schema variation. It is also very simple to scale following the demands of the application. NoSQL facilitates distributed storage of large amounts of data.

14. Snowflake Data Modeling Interview Questions

Q66) How does one gain access to the Snowflake Cloud data warehouse?

Ans) The following methods exist for accessing Snowflake's data warehouse:

- **ODBC Drivers** (A connection driver for Snowflake).
- **JDBC Drivers** (A driver that enables the communication between a Java application and a database).
- **Python Libraries** (for developing Python applications that connect to Snowflake and execute common operations).
- **Web User Interface** (This can be used for practically every activity you can execute using SQL and the command line).
- **SnowSQL Command-line Client** (Command-line interface based on Python for connecting to Snowflake from Windows, Linux, and macOS).

Q67) Briefly explain snowflake clustering.

Ans) Clustering is a form of data partitioning used in Snowflake, where unique cluster keys are given for every table. Cluster keys are subsets of a table's columns that allow you to group data within the table. Re-clustering is the process of managing clustered data in a table.

Q68) What does Snowflake's data retention period mean?

Ans) The data retention period is a key element of Snowflake Time Travel. Snowflake maintains the previous state of the data in a table when it is modified, such as when data is erased or objects containing data are eliminated. Data retention defines how long past data will be kept to support Time Travel operations (SELECT, CREATE, CLONE, UNDROP, etc.).

The default retention period for all Snowflake accounts is one day (24 hours). The default data retention duration for standard objectives is one day, ranging from 0 to 90 days for enterprise editions and higher accounts.

15. Power BI Data Modeling Interview Questions

Q69) How do Power BI Desktop and Power Pivot for Excel handle data modeling differently?

Ans) There is just one import mode, one computed column, and one one-to-many relationship supported by Power Pivot for Excel. Security, computed tables, bidirectional cross-filtering connections, and various import options are supported by Power BI Desktop.

Q70) Is it possible to have multiple functional relationships between two tables in Power Pivot?

Ans) No. In a Power Pivot data model, numerous inactive relationships exist but not more than one active relationship between two tables. Continuous lines indicate active relationships, whereas dot lines indicate inactive ones.

Q71) What is bi-directional cross-filtering?

Ans) Bidirectional cross-filtering enables data modelers to choose how they want their Power BI Desktop filters to flow for data using the relationships between tables. A second related table that is present on the other side of any given table relationship receives the filter context. Data modelers can use this method to address the many-to-many problem without using complex DAX calculations.

Q72) Define the Power Pivot data model. What analytics platform does Power Pivot use?

Ans) The relationships established between several data tables to efficiently structure the data are known as data models. There are specified data types, columns, associations, tables, relations, etc., in a data model. In contrast to calculated tables in Power BI, the data models in Power Pivot only support single-direction relationships that are one-to-many.

Power Pivot uses the SSAS in-memory Vertipaq compression engine. The client computer's in-memory storage holds the data models.

Q73) Define query parameters.

Ans) In the Power BI query editor, query parameters are the parameters that can be used with or as queries. Power BI Desktop allows us to generate new parameters, which we can then use for queries, data models, and reports. Query parameters include a subset of the dataset's data values.

16. Redshift Data Modeling Interview Questions

Q74) What does Amazon Redshift's AQUA stand for?

Ans) With the help of the hardware-accelerated cache, Advanced Query Accelerator (AQUA), Redshift can operate up to ten times faster than any other enterprise cloud data warehouse. For processing, all the data in a warehouse architecture with centralized storage must be sent to computing clusters. By performing a major fraction of the data processing in place on the cutting-edge cache, AQUA is used to move the computation to storage.

Q75) Define materialized views in Redshift.

Ans) An SQL query on one or more base tables is the foundation of a materialized view, which stores a precomputed result set. The same SELECT queries you use to query other tables and views in the database can query a materialized view.

Q76) What options for database querying does Amazon Redshift offer?

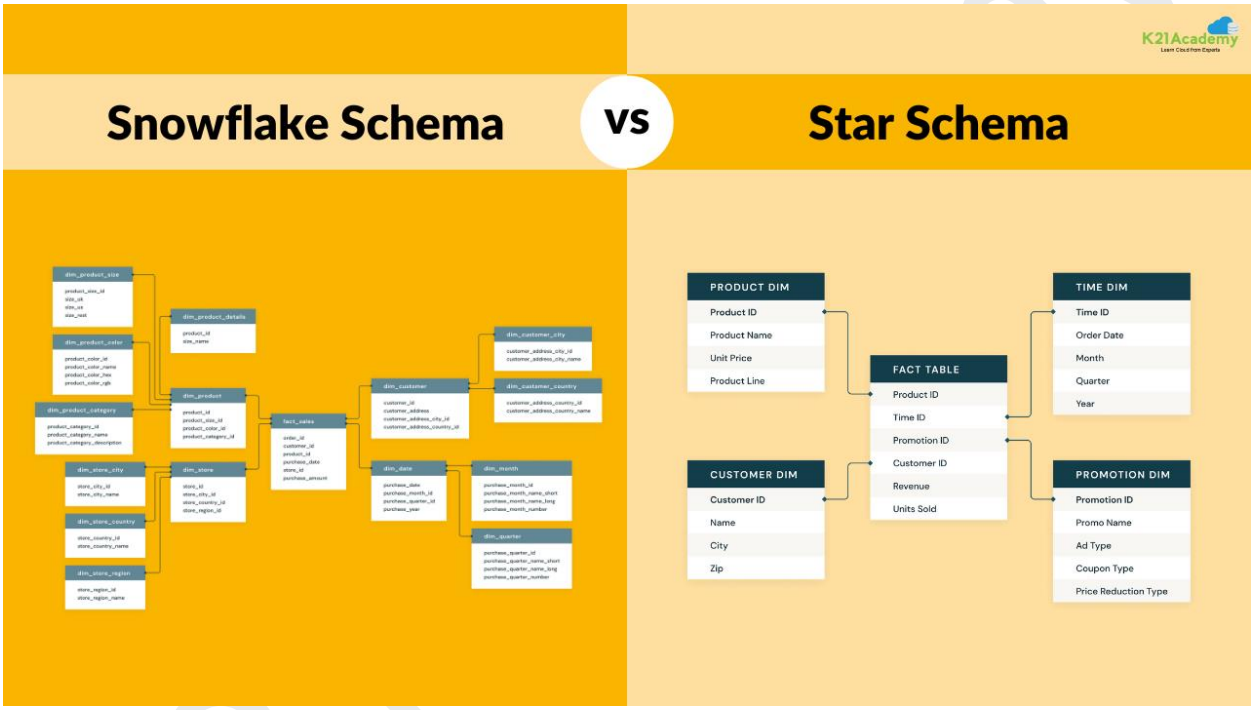
Ans)

- Using a SQL client application, employ common ODBC and JDBC connections to connect to your cluster.
- Run queries using the Query Editor on the AWS Management Console by connecting to your cluster.

17. Database Modeling Interview Questions

Q77) What exactly do you mean by Star schema?

Ans) A star schema consists of a central fact table and several connecting dimension tables. The central fact table's foreign key serves as the primary key for the dimension tables. Since the entity-relationship diagram can be seen as a star with points that branch from the main fact table, this schema is known as a star schema.



Q78) Briefly explain the snowflake schema database design.

Ans) A snowflake schema has one or more dimension tables that should interact with the fact table through other dimension tables but do not directly connect to it.

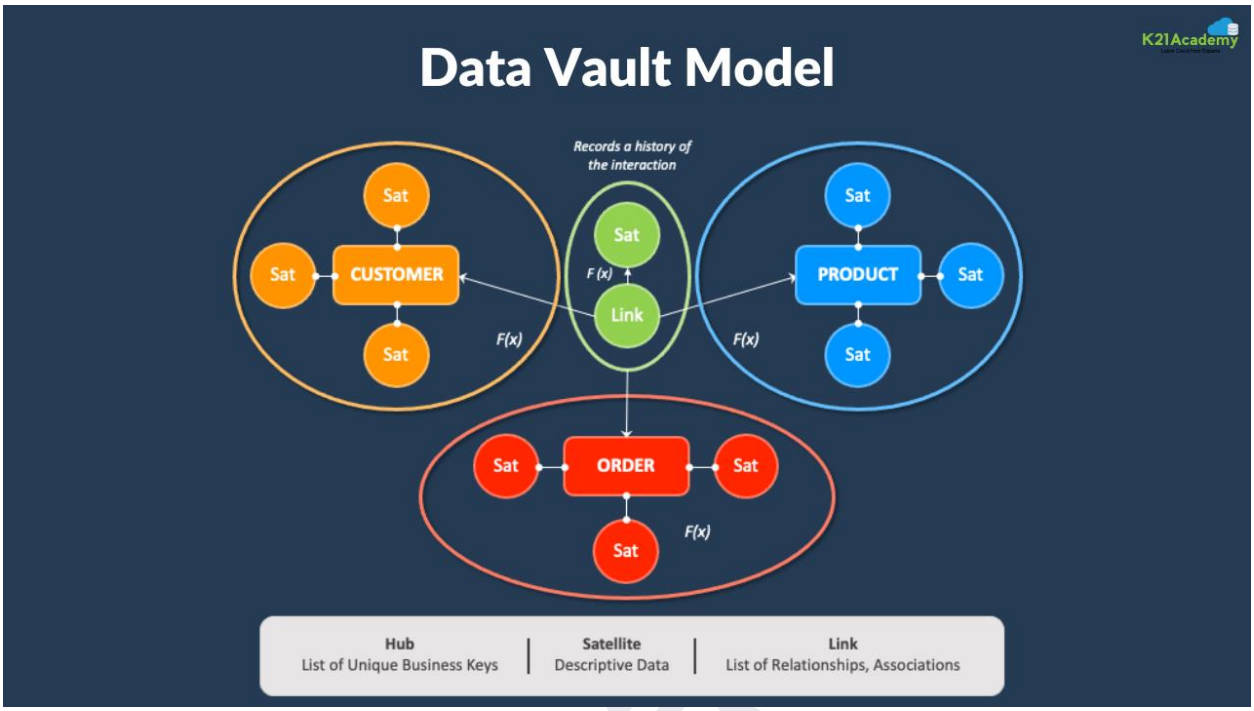
Snowflaking is the technique of normalizing dimension tables in the STAR schema to have the fact table at the center of a snowflake-like structure. Each dimension level in the snowflake schema corresponds to a level in the hierarchy, and the tables in the schema are normalized to the 3NF form. Multiple dimension tables connecting to other dimension tables via many-to-many or many-to-one relationships are connected to the core fact table. The schema helps improve query performance since it only requires a small amount of disk space and allows for the joining of smaller lookup tables.

Q79) What does forward and reverse engineering mean in terms of a data model?

Ans) Forward Engineering is the method for generating DDL scripts from the data model. Data modeling tools can connect to different databases and generate DDL scripts. These scripts further help in creating databases.

Reverse engineering is a method that helps to build data models from scripts or databases. Reverse engineering a database into a data model is possible using data modeling tools that enable connections to databases.

18. Data Vault Modeling Interview Questions



Q80) What is Data Vault Modeling?

And) The data vault modeling is a hybrid strategy that uses dimensional modeling and a third normal form to address the logical enterprise data warehouse. The design is adaptable to the organization's demands and is also extensible, flexible, and reliable. The data vault concept can be used with big, structured, unstructured data sets because it is designed as a bottom-up, incremental, and modular approach.

Q81) How many types of tables exist in data vault modeling?

Ans)
There are three basic categories of tables in data vault modeling:

- **Hubs:-** A distinct set of business keys for business objects.
- **Links:-** A unique set of relationships and transactions that act as the basic building blocks of a business process.
- **Satellites:-** Descriptive information for hubs and links.

Q82) Do Data Vaults function well with Star Schemas?

Ans) Staging, vault, and mart layers are all present in Data Vault 2.0. The Mart layer is host to star schemas, exposing a part of the vault to a specific user group. Typically, hubs and their satellites create dimensions, whereas links and their satellites create facts.

19.Data Warehouse Modeling Interview Questions

Q83) Define Virtual Data Warehouses.

Ans) A virtual data warehouse provides a unified overview of the final data. Historical data is missing from a virtual data warehouse. It is defined as a logical data model with metadata.

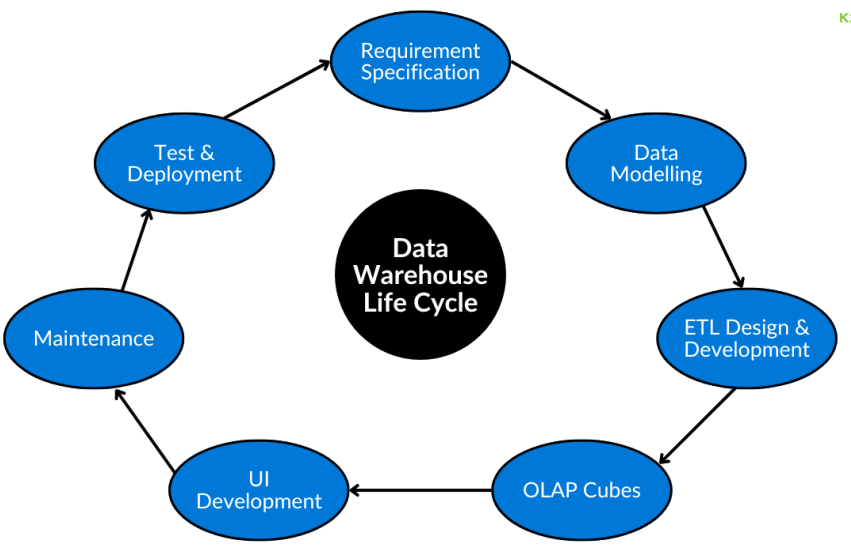
Q84) Explain column store databases and in-memory analytics.

Ans) A database can now store and retrieve data using a column store. A database's column store is a contemporary method for storing and retrieving data. These days, column store (or columnar store) databases use column-oriented storage, which stores the data vertically. Since the database reads the columns, a query contains, fewer hard disk reads are necessary to retrieve data.

Caching the whole database in RAM is known as in-memory analytics. This approach was typically overly expensive in the past because RAM was so expensive. Since RAM is relatively inexpensive, in-memory analytics are now being used by the industry at much higher rates.

Q85) What is real-time data warehousing?

Ans) Real-time data warehousing refers to a system that continuously reflects the status of the warehouse. Suppose you execute a query on the real-time data warehouse to find out more about a specific feature of the warehouse or entity it describes. In that case, the result will reflect that entity's status when the query is run. Most warehouses contain highly latent data representing the business during a specific period. Low latency current (or real-time) data is provided through a real-time data warehouse.



20. Data Science Modeling Interview Questions

Q86) Briefly define the Data Modeling Development Life Cycle.

Ans)

- Identify business requirements and get accurate data from users, subject matter experts, business analysts, etc.
- Make a project plan (Data Modeling Development Life Cycle Plan).
- Develop guidelines for data modeling development.
- Develop Conceptual data models.
- Develop Logical data models.
- Build a physical data model.
- Generate Lookup Data
- Build lookup scripts.
- Add more entities and attributes.
- Generate DDL scripts, which are necessary to create a database.
- Execute DDL scripts and lookup scripts in the database used for data modeling development, and if all works smoothly, share the results with DBAS.
- Compare Data Model and Database. Release the data model version if everything goes smoothly.

Q87) When should you use dimensional modeling as opposed to normalized modeling?

Ans)

Normalized Modelling:

- Predictable data usage.
- Usually, Enterprise-Wide use cases.
- When building a data store requires more effort and complexity than it does to maintain it.

Dimensional Modeling:

- Usually for Line of Business use cases.
- When a data store needs to be quickly set up.
- When you have more use cases for analytics or when querying is more crucial.

Q88) Why do you need an Enterprise Data Model?

Ans)

- Bringing common data across the organization - reduce duplication.
- Enhanced Analytics& Reporting. More informed decision-making.
- Ease of Regulatory Compliance.
- Helps with Data Discovery by breaking silos.

21. ERwin Data Modeler Interview Questions

Q89) What does "Name Hardening" mean in terms of ERWIN?

Ans) The Name Hardening Wizard offers a technique to prevent name inheritance and avoid name changes or resets inside a model. The object types and instances you need to harden logical names, physical names, or both within the model, as opposed to using the inherited name values, can be determined using the wizard.

Q90) What are the steps to enforce Data Type Standards in the Erwin Tool?

Ans)
Follow the steps below to apply data type standards:

- Select a method for applying standards.
- Use Model Explorer to define data type standards.
- (Optional) Generate a mapping.
- Lastly, apply data type standards to the model.

Follow these steps to import data type standards from an earlier version:

- Import data type standards.
- Apply data type standards to the model.

Use a model template to bind data type standards:

- Bind a template model.

Q91) What function does Report Designer perform in the ERwin Tool?

Ans) Report Designer enables you to generate a wide range of reports on the current data present in your data models. You can generate Mart administration reports and cross-model reports using the Workgroup Edition. You can export reports in HTML and CSV formats and create report templates for widely used reports. You can build report solutions with a variety of report templates.

Q92) What does the Complete Compare feature do?

Ans) With the helpful tool Complete Compare, you can identify and resolve any anomalies between two data models or between a data model and a database or script file. To help you determine the variations between the models, database, or script file, the Complete Compare wizard offers a wide range of compare criteria.

Q93) How can you build domains in the ERwin Tool?

Ans) Along with the pre-existing domains, you can add new domains using the Domain Editor. When you are in logical edit mode or a physical model, you can build a domain (physical edit mode). Depending on the edit mode, several choices are available in the editor.

22. Data Modeling Exercises With Answers

Q94) What entities & attributes would you start with when creating a data model for a bank's customer data?

Ans)

- Customer- Name, Date of Birth.
- Contact Information- Address and Phone number.
- Products - Auto-Loan, Debit Card, and Credit Card
- Risk Profile- Risk Score, Risk Appetite, Customer Segment.

Q95) Determine the potential flaws and drawbacks of the dimensional model below:

Ans)

ABC Logistics Shipment Dimensional Model
FACT_Shipment

Shipment_No	Date	Product	FCLLCL	Port_To_Port	FRREV	FRCOST	FUEL	Custom	Insurance	Security	Special_Handling	Other_Charges
AC229-02010	4/30/2017	Ocean Export	FCL	CNXGG - NLRT	467.23	-312.75	-25	-81	-8.95	-4.95		-4.95
TT52-001812	4/30/2017	Ocean Export	FCL	CNSHA - NLRT	446.73	-275.65	-31	-80	-8.95	-4.95	-9.95	
CW86-001287	11/12/2017	Ocean Export	FCL	USHOU - BHBA	396.98	-248.3	-28.5	-72	-6.95	-3.95		
BP72-001265	11/12/2017	Ocean Export	FCL	USHOU - BHBA	435.25	-327.12	25.73	-82.25	-8.95	-4.95		-14.95

The model should be separated into a FACT Header and a Fact Detail table, rather than extending charge line items horizontally. For in-memory/column storage standard practices, the charge line item column names should be combined into a single "Charge Code" dimension.

FACT_Header

Shipment_No	Date	Product	FCLLCL	Port_To_Port
AC229-02010	4/30/2017	Ocean Export	FCL	CNXGG - NLRT
TT52-001812	4/30/2017	Ocean Export	FCL	CNSHA - NLRT
CW86-001287	11/12/2017	Ocean Export	FCL	USHOU - BHBA
BP72-001265	11/12/2017	Ocean Export	FCL	USHOU - BHBA

FACT_Detail

Shipment_No	Charge_Code	Amount
AC229-02010	FRREV	467.23
AC229-02010	FRCOST	-312.75
AC229-02010	FUEL	-25
AC229-02010	Custom	-81
AC229-02010	Insurance	-8.95
AC229-02010	Security	-4.95
AC229-02010	Other_Charges	-4.95
TT52-001812	FRREV	-446.73
TT52-001812	FRCOST	-275.65
TT52-001812	FUEL	-31
TT52-001812	Custom	-80
TT52-001812	Insurance	-8.95
TT52-001812	Security	-4.95
TT52-001812	Special_Handling	-9.95
CW86-001287	FRREV	-396.98
CW86-001287	FRCOST	-248.3
CW86-001287	FUEL	-28.5
CW86-001287	Custom	-72
CW86-001287	Insurance	-6.95
CW86-001287	Security	-3.95
BP72-001265	FRREV	435.25
BP72-001265	FRCOST	-327.12
BP72-001265	FUEL	-25.73
BP72-001265	Custom	-82.25
BP72-001265	Insurance	-8.95
BP72-001265	Security	-4.95
BP72-001265	Other_Charges	-14.95

Q96) How do you model tables with the same dimensions in a data warehouse? (Consider the below three tables relevant to the hiring process)

Ans)

Table 1 shows the information on candidates who have already been accepted.

DEPARTEMENT	SPECIALIZATION	CLASSIFICATION	NATIONALITY	APPLICANT	PLACE OF RESIDENCE	CONTRACT STATUS

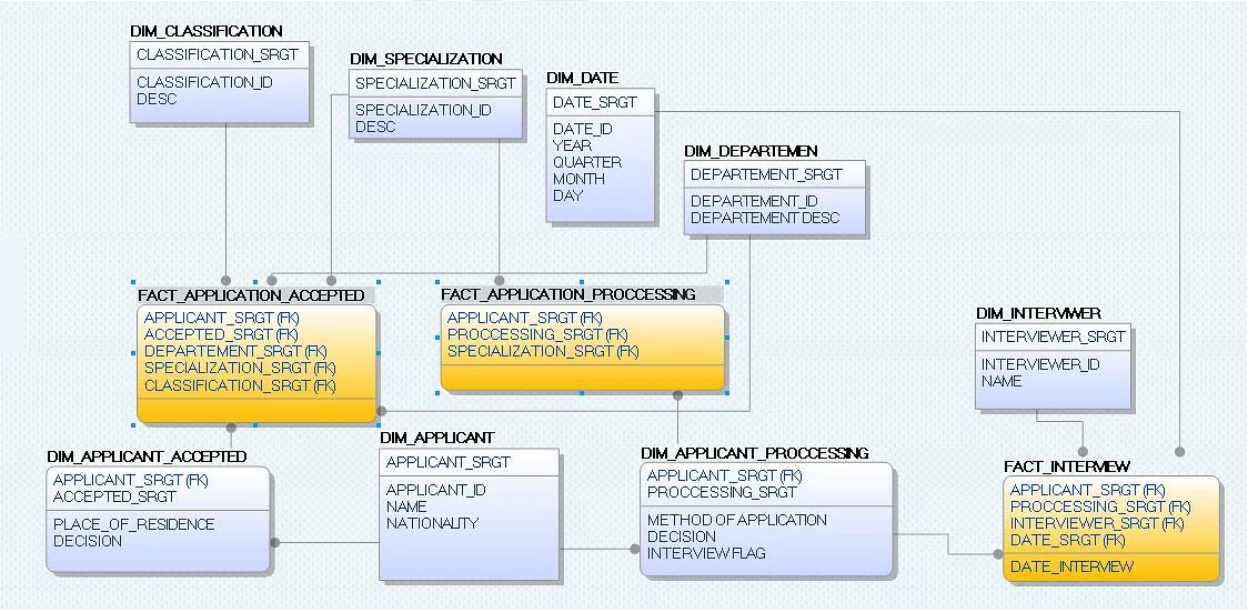
Table 2 shows information about the candidate who is currently under evaluation

DEPARTEMENT	SPECIALIZATION	NATIONALITY	APPLICANT	METHOD OF APPLICATION	DECISION	INTERVIEW FLAG

Table 3 shows the results of the interviews.

DATE OF INTERVIEW	INTERVIEWER	APPLICANT	DEPARTEMENT	NATIONALITY	DECISION

Here's how we can model the given tables with similar dimensions-



23. FAQs on Data Modeling Interview Questions

Q1) What are the 4 different types of data models?

Ans)
The four different types of data models are-

- **Relational Model:-** The relational model, which is the most popular, organizes data into tables, or relations, which consist of columns and rows.
- **Hierarchical Model:-** The hierarchical model organizes data into a tree-like layout with a single parent or root for each record.
- **Network Model:-** The network model extends the hierarchical model by enabling many-to-many links between connected records, indicating the presence of several parent records.
- **Entity Relationship Model:-** This model, like the network model, represents relationships between real-world entities, although it isn't as closely linked to the database's physical design.

Q2) How to prepare for data modeling interview questions?

Ans) If you want to prepare for data modeling interview questions, think about what types of questions the interviewer can ask about the data modeling process so you can come up with quality responses that reflect your knowledge and experience. You will find several blogs and videos on the internet that will help you get a better idea of some of the common data modeling interview questions. Also, you can strengthen your data modeling skills by working on relevant projects available on GitHub, ProjectPro, etc.

Q3) What is the purpose of data modeling?

Ans) The purpose of data modeling is to show the different types of data that are used and stored in the system, their relationships, how they might be grouped and structured, and their formats and attributes. Data modeling aims to provide high-quality, accurate, structured data that helps implement business applications and generate consistent outputs.

Q4) What is data modeling in Excel?

Ans)

- Data modeling in Excel makes it simple to build links between simple metrics and their underlying data sources.
- It simplifies data analysis.
- Creating relationships between matching columns enables the unification of data from various tables across different worksheets.

Q5) What is data modeling in SQL?

Ans) Data modeling is a method of organizing and connecting data to perform data analysis. We use data modeling to arrange data for numerous tables.

FREE Consultation Call

Whether you're struggling to get a job or confused about where to even begin to get a thriving career, book a free clarity call with one of our experts at the time that works best for you.

BOOK A CALL NOW

THANK YOU

www.k21academy.com

K21Academy
Learn Cloud from Experts