ERD Connection Traps

by Sophia

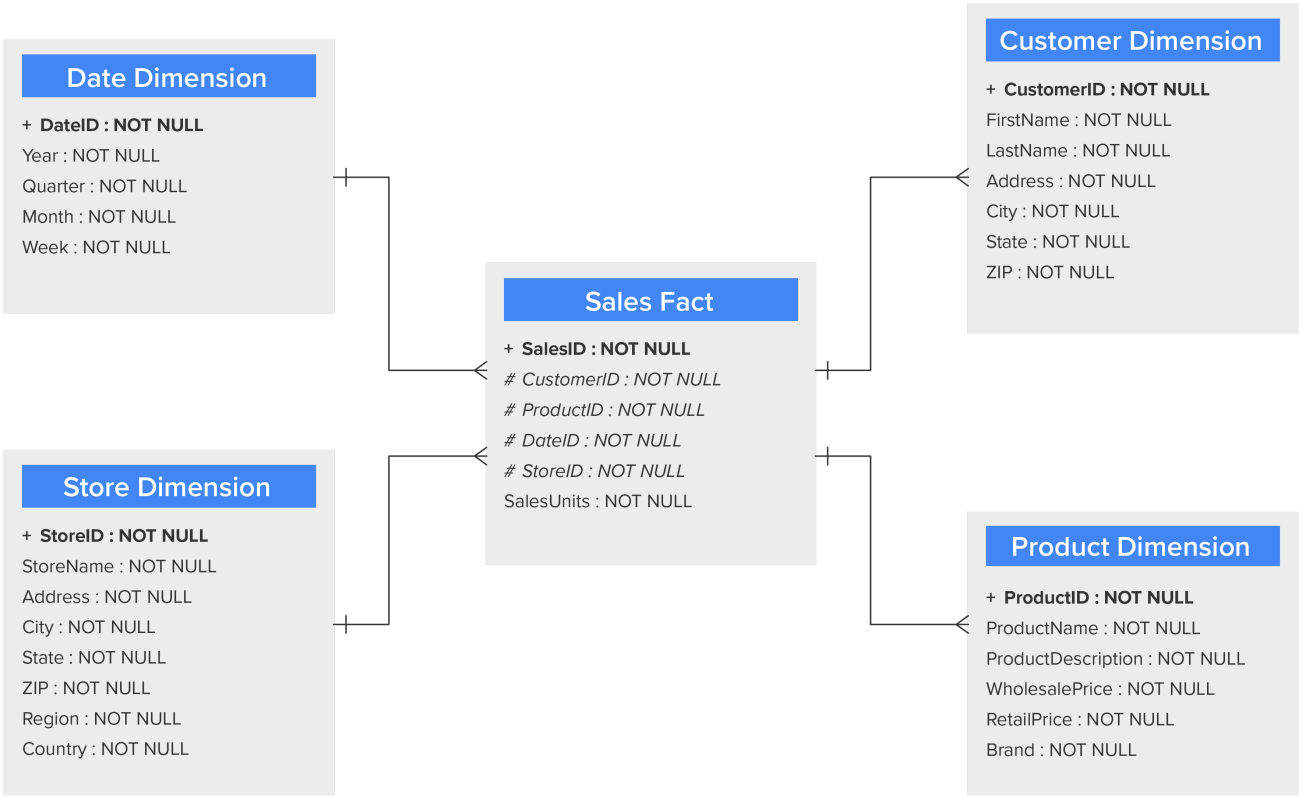
WHAT'S COVERED

In this lesson, you will explore connection traps that can be created in entity-relationship models, in four parts. Specifically, this lesson covers:

1. Star and Snowflake Schema

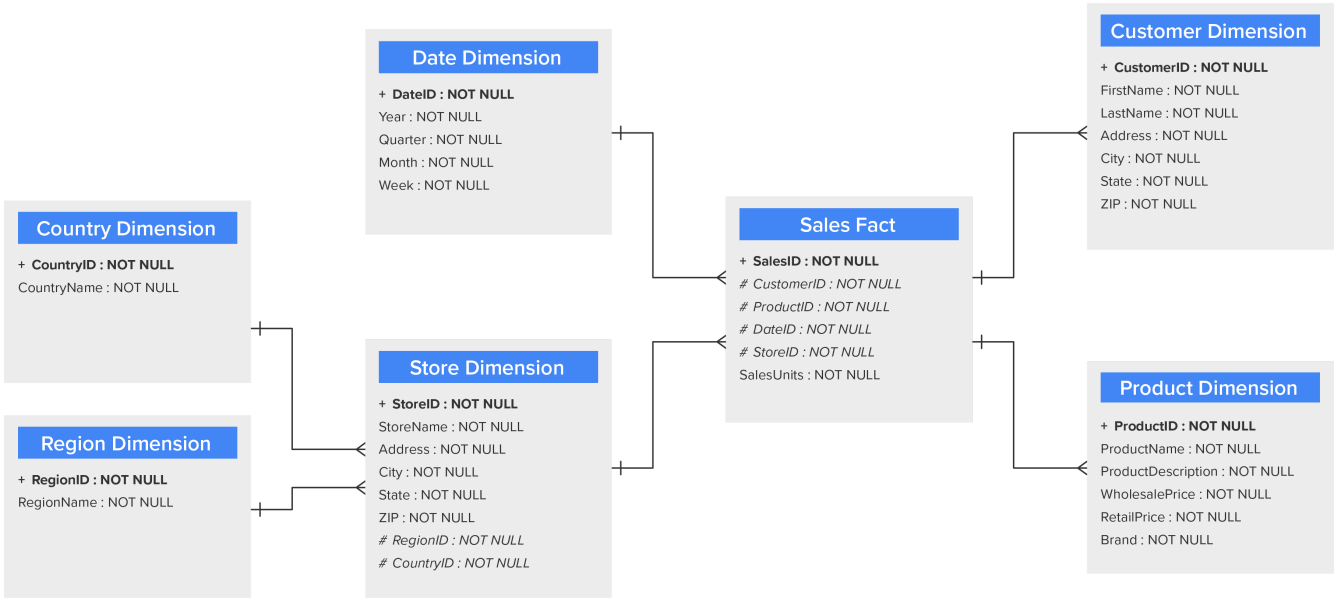
Data warehouse database systems, which should be optimized for analysis, often use a **star schema**. There is a central table, called a **fact table**, which contains quantitative data such as sales figures or other measurable metrics. This table serves as the primary focus for analytical queries.

Then there are smaller tables, called **dimension tables**, which are joined to it. The primary key of each dimension table is joined to a foreign key in the fact table. These relationships establish the context and provide descriptive attributes for the numeric data. Examples of dimension tables for a Sales fact table are Date, Customer, Store, and Product, as shown below. A complex schema might have several fact tables, each with connections to multiple dimension tables.



Some dimension tables have a hierarchical structure, allowing for easy drilling down or rolling up in analysis. For example, a Date dimension might have hierarchy levels like Year, Quarter, Month, and Week. This would enable someone querying the database to run reports based on any of those time intervals.

Sometimes one of the dimension tables needs its own dimension table for better normalization, so you end up with a **snowflake schema**. A snowflake schema is a star schema with multiple levels of dimension tables. In the following example schema, Store is a dimension table, and it has Country and Region as its own dimension tables.



Star and snowflake schemas are optimized for query performance, which is what is needed in an analytical database. The modular nature of a star or snowflake schema makes it very scalable. New dimensions can be added without impacting the existing structure.

TERMS TO KNOW

Star Schema

A schema in which a central fact table serves as the primary focus with support from dimension tables.

Snowflake Schema

A schema in which some of the dimension tables in a star schema have their own subordinate dimension tables.

Fact Table

A table that contains quantitative data such as sales figures or other measurable metrics.

Dimension Table

A table that supports a fact table by storing additional details about its data.

2. Connection Traps

A **connection trap** is a situation where a complex relationship pattern can lead to ambiguous or incorrect query results and challenges in maintaining data integrity.

Two common types of connection traps are a chasm trap and a fan trap. The key difference between them lies in the way relationships are structured. You will learn about those differences when you study each type individually later in this lesson.

Both kinds of connection traps occur in scenarios where there are at least three tables involved, and they cause problems (such as delivering inaccurate query results) when a query requests information from multiple interconnected tables.

The root cause of both traps is often related to the way relationships are structured among the tables. Resolving both types of traps typically involves introducing additional tables or adjusting existing relationships to create clearer paths for queries to execute their logic.

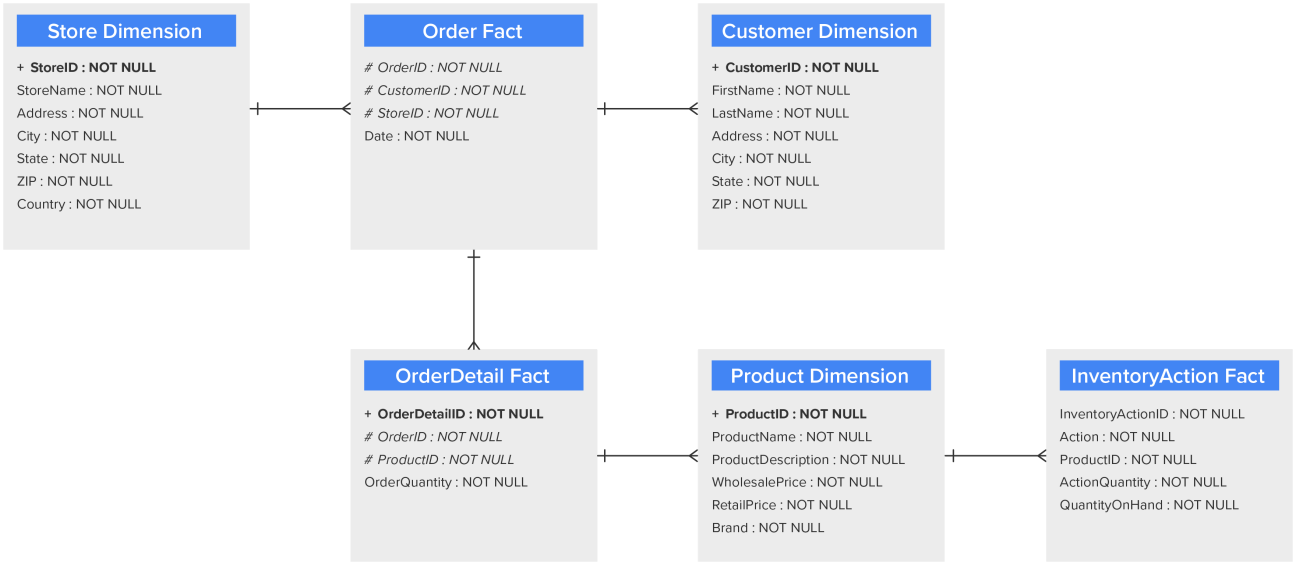
TERM TO KNOW

Connection Trap

A situation where a complex relationship pattern can lead to ambiguous or incorrect query results and challenges in maintaining data integrity.

2a. Chasm Traps

A **chasm trap** can occur when two or more fact tables have no direct relationship to each other except through a shared dimension table. For example, in the following diagram, the OrderDetail and InventoryAction fact tables are connected only through the Product dimension table. It's called a "chasm" because there's an empty space, a lack of relationship, that must be bridged in order to make the connection.



The main problem with a chasm trap's existence is that joining tables across a chasm trap can enable a **Cartesian product** to occur if the query does not specify appropriate join conditions. In other words, each row from the first fact table will be joined to each row from the other fact table. This may cause over-counting in a query that computes aggregates or counts when the query includes data from both of the "many" ends of their respective joins. In the case of the above diagram, this would be the case if you wrote a query that asked for aggregate values from both OrderDetail and InventoryAction.

TERMS TO KNOW

Chasm Trap

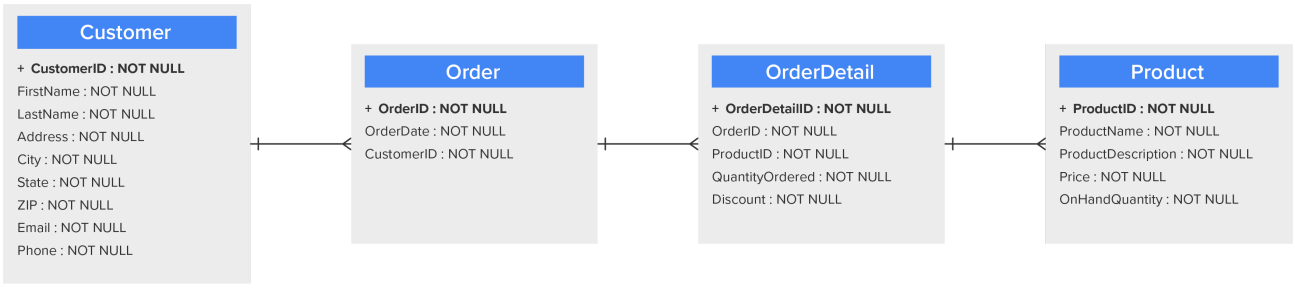
A condition in which two one-to-many joins converge on a single table.

Cartesian Product

An over-counting of data in a query that computes aggregates or counts when the query includes data from multiple tables that are not directly connected.

2b. Fan Traps

Whereas chasm traps occur when two relationships converge on a central table, a **fan trap** occurs when two or more one-to-many relationships fan out from the same table. A fan trap represents a relationship chain where there is a one-to-many relationship between two tables, and then the table on the “many” side of that relationship has another one-to-many relationship in which it is the “one” side. This fanning-out effect can cause queries to return incorrect results when the query draws data from tables that are not adjacent in that chain. For example, in the following figure, the fan chain is from Customer to Order and then from Order to Order Details.



As with a chasm trap, a fan trap can also enable the possibility of a Cartesian product situation where a query that pulls aggregate data from multiple tables can show inflated aggregate values because of errors in the aggregation.

TERM TO KNOW

Fan Trap

A condition in which two one-to-many relationships follow one another in a parent-child form.

3. Resolving Connection Traps

To resolve connection traps in a database, you must adjust the database schema by introducing additional tables or modifying existing relationships. The steps to take to resolve connection traps require more SQL knowledge than you currently have, so we will confine this discussion to an overview of the process.

Here is a high-level overview of the process for resolving a chasm trap:

1. Create a bridge table (also called a link table) that establishes a direct connection between the two outer tables causing the chasm trap. This bridge table should contain foreign keys that correspond to the outer tables’ primary keys.
2. Adjust relationships to use the bridge table as a connector.

Here is a high-level overview of the process of resolving a fan trap:

1. Create a new table to provide separate paths for the two one-to-many relationships. This is essentially the same as the bridge table when resolving a chasm trap, except in this case, it is called a resolver table.
2. Adjust relationships to use the resolver table as a connector.

KEY CONCEPT

It's important to note that while both of these types of connection traps introduce the possibility of Cartesian product errors, they do not inherently cause such errors. You can prevent Cartesian product errors even in the presence of connection traps by making sure to use explicit join conditions in your queries. You will learn about structuring queries that include joins in upcoming units of this course.

After making adjustments to your schema, test your queries to make sure that the adjustments have resolved the connection trap, and the query results are accurate and meaningful. Then document the changes you made. This documentation helps others who may work on the database to understand its structure and relationships.

If the database structure is complex, consider seeking the expertise of a database designer to ensure optimal resolution of connection traps.

SUMMARY

In this lesson, you learned about a **star schema**, in which a single fact table is supported by one or more dimension tables. A **snowflake schema** is similar to a star schema, but some of its dimension tables have their own subordinate dimension tables.

Next, you learned about **connection traps** and how they can result in weaknesses in a database that permit queries to return inaccurate data if the queries are not carefully constructed. The most common type of query error that can result is a Cartesian product, in which a query that includes data from multiple tables that are not directly connected may over-count data when computing aggregates.

You then learned about two common connection traps. In a **chasm trap**, two or more fact tables are connected to a single-dimension table, but no direct relationship exists between them.

Fan traps occur when multiple dimension tables are connected to a single fact table without any direct relationship between them.

Resolving the traps requires careful data modeling to ensure that the relationships between the fact tables and dimensions are correctly represented, enabling accurate and meaningful data analysis and reporting. Creating additional bridge or resolver tables is usually required. The queries themselves can also be fine-tuned to use explicit join conditions that prevent aggregation errors.

Source: THIS TUTORIAL WAS AUTHORED BY DR. VINCENT TRAN, PHD (2020) AND Faithe Wempen (2024) FOR SOPHIA LEARNING. PLEASE SEE OUR TERMS OF USE.

TERMS TO KNOW

Cartesian Product
An over-counting of data in a query that computes aggregates or counts when the query includes data from multiple tables that are not directly connected.

Chasm Trap
A condition in which two one-to-many joins converge on a single table.

Connection Trap
A situation where a complex relationship pattern can lead to ambiguous or incorrect query results and challenges in maintaining data integrity.

Dimension Table
A table that supports a fact table by storing additional details about its data.

Fact Table
A table that contains quantitative data such as sales figures or other measurable metrics.

Fan Trap
The fan trap occurs when you have two one-to-many joins that follow a parent-child form. You will probably get incorrect results if you try to aggregate both measures simultaneously.

Snowflake Schema
A schema in which some of the dimension tables in a star schema have their own subordinate dimension tables.

Star Schema
A schema in which a central fact table serves as the primary focus with support from dimension tables.