# Data Modeling: The Star Schema

Marco Sanchez-Ayala · Follow
4 min read · Apr 5, 2020



Data modeling is a crucial step in modern data workflows, as its purpose is to organize raw data into convenient and efficient forms. Data analysts and scientists will find their jobs much easier if a usable dataset is readily accessible. Quicker analytics and predictions will then lead to faster insight for business decisions.



The first step to modeling is often to normalize the data, which is a process of organization that increases database flexibility by reducing inconsistent dependencies and redundancies). I'd suggest reading up on this and/or looking up some videos if you're unfamiliar! The problem with a normalized database is that any really interesting insights from the data will require many `JOINS`, which can significantly slow down the speed of our query as the size of our database increases. For instance, looking at the schema below, most tables are not directly related. This means that to connect information from two tables like `Orders` and `Location` we'd need a minimum of 4 `JOINS` (`Orders -> Employment -> Person -> Phone_Number -> Location` is one way to arrive there)

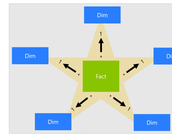
https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/bp-relational-modeling.html

Now what if we need data from even MORE tables after that 4-part `JOIN`? It would be madness. Not to mention, it would be an absolute headache to just write the query without any errors.

Furthermore, real-life databases can have far more tables than just the ones shown in the example above. As you can imagine, it becomes increasingly difficult to even understand the relationships between tables as our schema grows.

## The Star Schema

One solution to this problem is to perform a denormalization step of data modeling to create a simpler and easy-to-understand schema optimized for certain queries. The process of creating a star schema involves distilling down our full schema into just relevant features for a particular analytic propose. The general structure of the star schema is as follows:




https://docs.microsoft.com/en-us/power-bi/guidance/star-schema

The star schema consists of two types of tables:

1. *Facts:* Metrics of a business process. These are generally numeric and additive (e.g. amount of an invoice or the number of invoices), or quantitative. The fact table also contain keys pointing to relevant dimension tables. There is just one fact table at the center of the star schema.
2. Dimensions: The where, when, what, etc. (e.g. date/time, locations, goods sold). These typically contain qualitative information. There are multiple dimension tables in the schema, all of which are related to the fact table.

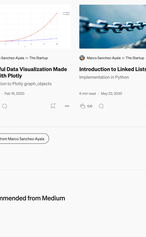
### Advantages

- A simplified schema means that we won't have to write confusingly long queries every time we want some information from our database.
- We optimized for reads. Now that we can write fewer `JOINS`, our results will be returned more quickly.
- Also, it will business logic for reporting. We won't have to explain to stakeholders all the crazy joins that went into making the schema, just maybe .

### Disadvantages

- Denormalizing our data means that data anomalies could arise from one-off inserts or updates. In practice, star schemas are implemented via "trickle feeds" or batch processing to compensate for this issue.
- We have limited analytical flexibility. A star schema is generally designed for a particular purpose. Since we have fewer features in the star schema than in the full database, we are restricted to just what this star schema contains.

### Example




https://www.guru99.com/star-snowflake-data-warehousing.html

Let's consider a database of sales for a store. We have a fact table, `Revenue`, at the center of the schema, and four dimension tables.

The fact table is comprised of a composite primary key, which is the combination of the dimension table primary keys. Fact table non-primary keys `Units_Sold` and `Revenue` are the facts we're interested in, and the dimensions such as `Product_Name` and `Name` (branch name) allow us to understand more information about the goods sold.

For instance, the following query would allow us to calculate the total revenue by product in the year 2010:

```
SELECT
    p.Product_Name AS product,
    SUM(r.Revenue) AS total_revenue
FROM
    Revenue r
JOIN
    Product p ON (r.Model_ID = p.Model_ID)
JOIN
    DateDim d ON (r.Date_ID = d.Date_ID)
WHERE
    d.Year = 2010
GROUP BY
    p.Product_ID
```

• • •

The star schema is widely used and incredibly useful for business applications. It helps us speed up queries that we may run often and clean up what could otherwise be very messy queries, among other things.

There are other schemas such as the snowflake and galaxy schemas that are simple extensions of the star schema. If you like the star schema, I recommend checking the others out too!

Data Engineering | Data Modeling | Sql | Star Schema | Analytics



Written by Marco Sanchez-Ayala
94 Followers
Programmer, dancer, and musician.



More from Marco Sanchez-Ayala



Marco Sanchez-Ayala in Level Up Coding
**Calculating Median in MySQL**
While sharpening my SQL skills online, I was asked to calculate the median of a column in...
6 min read · Mar 16, 2020
354 3



Marco Sanchez-Ayala in The Startup
**Regular Expressions in MySQL**
Text data often requires complex searching to extract relevant and/or meaningful...
6 min read · Dec 18, 2019
140



Marco Sanchez-Ayala in The Startup
**Beautiful Data Visualization Made Easy with Plotly**
Introduction to Plotly graph_objects
8 min read · Feb 16, 2020
331



Marco Sanchez-Ayala in The Startup
**Introduction to Linked Lists**
Implementation in Python
6 min read · May 23, 2020
108

See all from Marco Sanchez-Ayala

## Recommended from Medium



Nuhad Shaabani
**Practical Introduction to Data Vault Modeling**
The Data Vault modelling is used to model the enterprise Data Warehouse Core layer. The...
10 min read · Jul 15, 2023
87



Nam Huynh Thien
**Building a Dimensional Data Warehouse Using dbt**
Introduction
10 min read · Aug 5, 2023
110

Lists



Apple's Vision Pro
7 stories · 41 saves

Predictive Modeling w/ Python
20 stories · 774 saves

ChatGPT
23 stories · 388 saves

data science and AI
38 stories · 97 saves



Deepanshu Sunal
**Introduction to Data Modeling—2024 Guide With Problems**
Data modeling is the process of creating the conceptual representation of data and its...
5 min read · Jan 2
138 2



Tarandi Kiyur in Code Like A Girl
**Slowly Changing Dimensions: Unlocking the Potential.**
Strategies and Best Practices for Maintaining Data Consistency and Accuracy Over Time
7 min read · Jul 19, 2023
3



Sanjeev Dora in Dev Genius
**Introduction to Data Build Tool (dbt)**
Once upon a time, in the mystical realm of data management, two powerful forces...
10 min read · Aug 6, 2023
102 2



Carlos Costa
**Yet Another Data Modeling Approach for the Data Lakehouse**
If you were missing a framework for modeling data inside the Data Lakehouse, I hope this...
12 min read · Jul 22, 2022
185 2

See more recommendations