# Data Modeling Intro

## What is data modeling?

* Its about organizing data by applying certain formal techniques. It involves:

**Data Gathering**

* Structured process aligned with a business need

**Best Practices**

* The effective use of data

**Storage**

* Systems used to store, edit and access data

**Ethics & Legality**

* Appropriate use of data

## Why Learn Data Modeling?

* To understand data and its structure and how it needs to be structured in order to get to the answers you are going to need to get to

1. Overcome limitation in analysis
   1. We have limitations on the size of our worksheets in excel or google sheets and the amount of data that can be stored in computer memory is limited. Therefore it limites our ability to run quick and effective analysis.
2. Overcome systemic limitations
3. Simplify calculation
4. Gain insight faster
5. Develop a modeling mindset
6. Develop a modeling skillset
   1. Not enough to understand how data is organized but we need to also know how to do it.

## 

## Terminology

### Field (or fact)

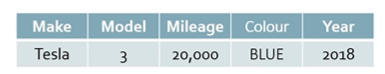
* Atomic level of storage
* Simple fact or data point
* Meaningless on its own
* Requires context to interpret

Example the table has a fact call blue but it means nothing



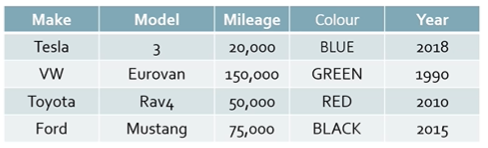
### Row

* Collection of fields (or facts)
* Related facts that are descriptive in some way
* Describe one object in the larger collection
* Rows represent one of the data objects
* Begins to add context (and therefore understanding) to the data



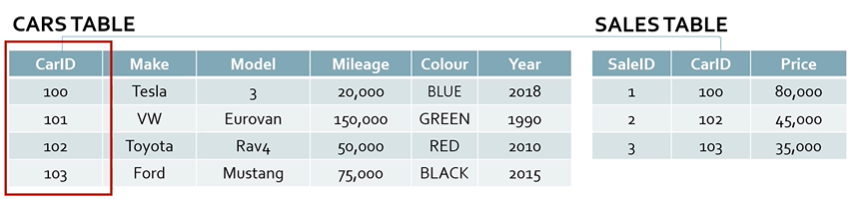
### Table

* Collection of rows
* Tables collect objects into groups
* Each table contains one collective concept of the database (“Entity”)
  + All the cars in the database
* Table entities may be physical (cars) or virtual (sales)



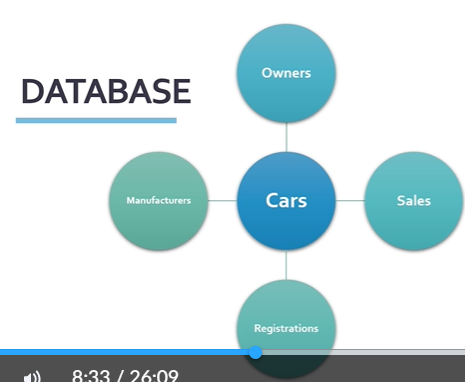
### Primary Key

* Also foreign key
* Contains a field for ID. For each row/record in the table in CarID there is a unique identifier. It is auto incremented. This is important to identify the rows or records within our data sources so we can relate the data to other sources of data.



* Cna see in the sales table there is there is CarID listed again but it is a foreign key. The primary key is SaleID
* When we refer to a primary key in another table that is known as a foreign key. This is to build relationships between the data

**Database**



* Collection of tables
* Tables are joined by relationships
* Tables do not need to be in the same database (virtual or linked tables)
* **Data modeling** is the act of gathering many data sources into one related set of data.
* This is where power query is really important in excel because it enable you to gather many data sources into one related set of data and using excel as an example that is called the **data model**
* For example we can extend the Cars into different sources such as ownership, sales, manufacturers, registrations

## Introducing the Data model

### Data Modeling

* We want to highlight tools in excel because it is ubiquitous
* Excel has research rich functionality that not many people know how to use.

**Why did they do this?**

* Because as powerful our computers are, we still are challenged by the resources that are available to us.
* When we open an app it runs in computer memory including whenever you are running data/data models.
* Excel power query makes this easier on your hardware in order to get the insight in a timely manner.
* If we think of limitations in excel and our computer if we have data we dump it into excel and model excels understand tables and formatting or pivot tables/charts. There are some limitations. We can only have 1 million rows of data within a table/excel.
* When you are dealing with big data it is easy to go past 1 million rows.
* We store data externally. We use database and database technologies ot access and shape data but for business users who do have access to excel. There is another way to access this data. Even if its on a website/database, excel etc we connect using power query or power pivot. We use this tool to shape and format and connect data from multiple sources. This is what is known as data modeling.

## Using Power Pivot to Overcome Limitations

* No limit to table size loaded in model
  + Power Query holds model (full dataset)
  + Excel reports from model (grouped and filtered dataset)
* Power Pivot (2010/2013) to Excel Data Model (2016+)
  + Same tool. Different name.
* You are continually benefiting from updates in excel
* Even updates in Power BI
* Microsoft invests in this because they know there is so much data given to us and business users need to access and pull the data into a tool they can quickly use to shape, format and connect to other data sources. What this requires is that we summarize the data to an extent and require the user to consider the granularity of that data.

## Granularity

* Refers to the level of detail in the dataset’s critical value
* Single table models require a determination at creation, which restricts analysis
* Granularity can in dimension (columns) or value (rows)
* “Highly granular” data had more detail
  + More columns generally means more detail to describe
  + Leads to more rows of data
* Granularity must be determined at the critical level of detail (just right) to report correctly
* Aggregation diminishes granularity
* In order to get the insight that we need to get we need just the right level of granularity. Not too much or else it will be cumbersome. If we don't have the right level then we wont get to our answer.

## Normalized vs Denormalized

### Normalized

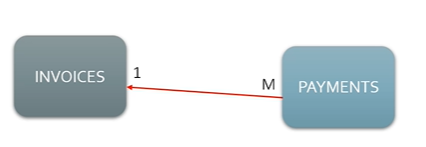
* Databases optimize storage using the rules of normalization
  + One table per concept
  + No repeating data
  + Unique identifiers

### Denormalized

* Business intelligence reporting works from denormalized data
  + Queries consolidate data from the database into a single virtual table
  + Reports group and summarize from this virtual table

## Relationships

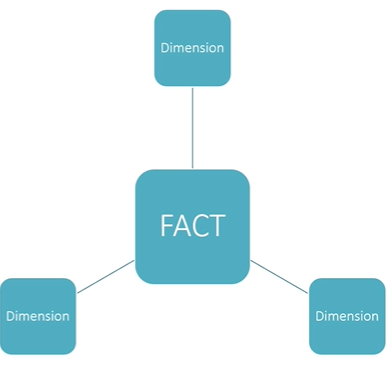
* Databases resolve the connection between data concepts by building relationships.
* Most common relationship type is One-to-Many (1:M)



* It shows that there is one invoice but there can be many payments made.

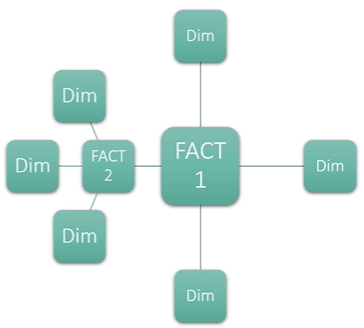
## Star Schemas

* Defined by a central measurement table surrounded by additional descriptive detail
* The **Fact** table contains critical numeric measures (the key metric that we wish to record in the database)
* Dimension tales contain categorical detail, allowing for grouping, filtering, and aggregation (slicing the data)
* Where is your data and how is it layed out



## Snowflake Schemas

* What if a **Fact** table connected (related) to another star schema?
* Snowflake schemas are often needed to model several different “facts” which are closely related, but conceptually distinct
* Snowflake schemas can be useful but pose special problems for reporting



## The Importance of Names

### Names Should..

* Be appropriately descriptive
* Be clear and concise
* Follow convention
  + Tables contain conceptual objects; plural
  + Fields contain data describing a record; singular
  + Avoid space & special characters; CamelCase