GENERAL ASSEMBLY

DATA SCIENCE
DATABASES AND SQL

I. INTRODUCTION TO DATABASES
II. STAR SCHEMAS
III. WHY DO DATA SCIENTISTS NEED DATABASES?
IV. LEARNING SQL WITH CODE

I. INTRODUCTION TO DATABASES

WHAT IS A DATABASE?

- An organized collection of data
- Organized overall by a schema (like a blueprint of a database)
- Organized into tables with different sets of data
- If each family is a set of data, a house would be the table, and the neighborhood would be the schema.
- Think many Excel sheets/pandas dataframes, but without limitations

WHY USE A DATABASE?

- You can ask questions of the data
- Has a nice, structured language
- Produce reproducible code
- Access large amounts of data relatively quickly
- Reliable and scalable
- Many are ACID compliant — ensures your transactions are safely processed or that you're notified otherwise

RELATIONAL VS. NOSQL

- Relational
  - Traditional rows and columns data – like dataframe
  - Strict structure
  - Entire column for each feature
- NoSQL
  - No well defined data structure
  - Works better for unstructured data
  - Commodity hardware

RELATIONAL VS. NOSQL

- Today, we'll be talking about relational databases
- Most widely used and most appropriate for many types of data
- Popular names
  - MySQL
  - Oracle
  - Postgres
  - Microsoft SQL Server
  - SQLite

SQL

- Structured Query Language
- Used to ask questions of the database
- Many different functions for creating, adding, retrieving, transforming, aggregating, and deleting data
- Standard language with some differences among “dialects”

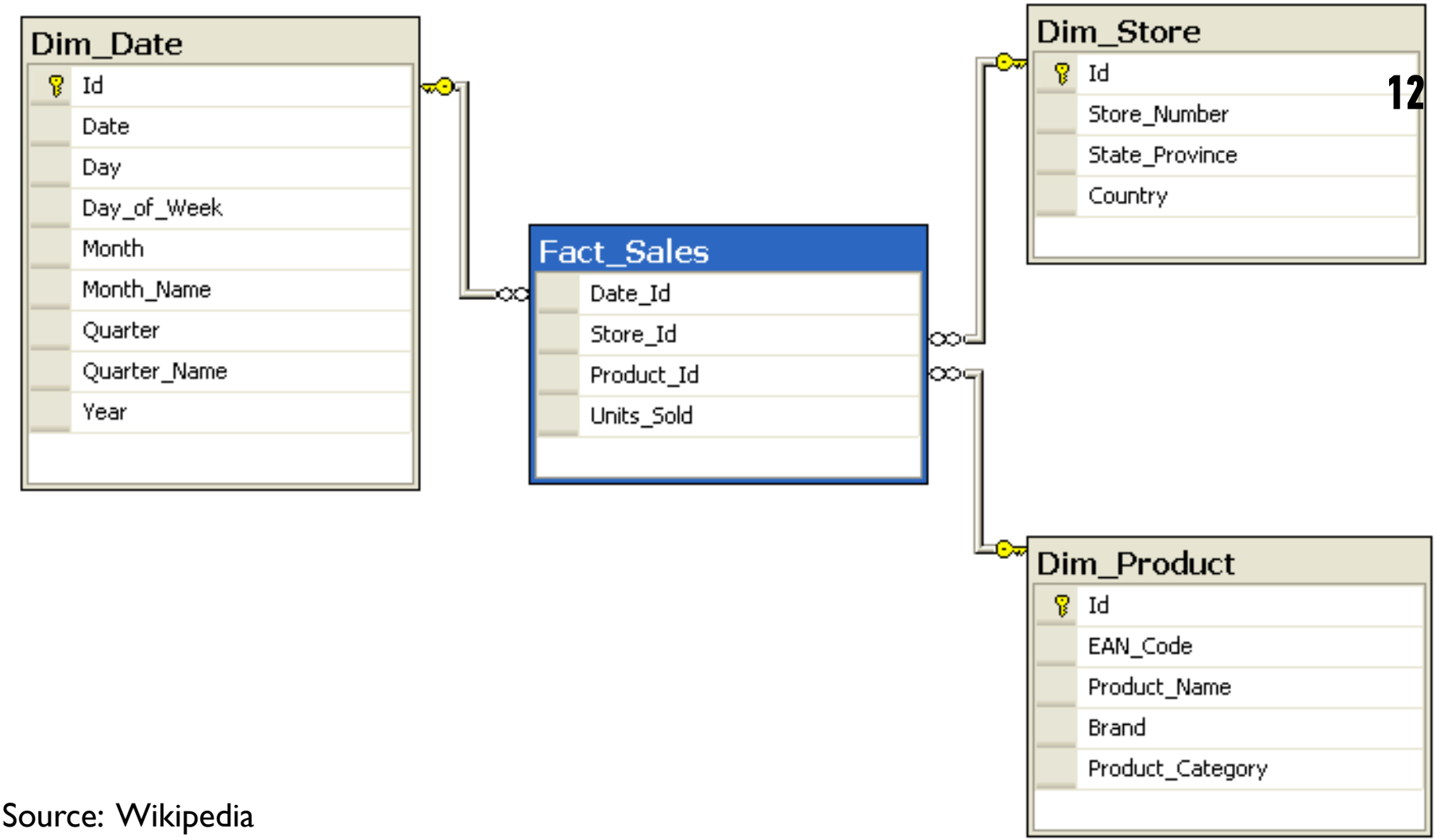
DATA TYPES

- BOOLEAN/TINY INT– 0/1
- INT – any whole number
- FLOAT(<n>,<m>) – number with n digits before the decimal and m digits after the decimal
- DATETIME, TIMESTAMP, and DATE – various date and time combinations
- CHAR(<length>) – text with a fixed length
- VARCHAR(<length>) – text with a given maximum length
- And many more...

II. STAR SCHEMAS

- The star schema consists of one or more fact tables referencing any number of dimension tables.
- A fact table contains “event” data. You can think of this as the type of information that we are really measuring (“measurements, metrics, or facts of a business process”).
- A dimension table contains meta data or information that enhances “event” data (“structured labeling information”).

Source: Wikipedia



Source: Wikipedia

III. WHY DO DATA SCIENTISTS NEED DATABASES?

- In business, data doesn't often live in flat files like CSV's or TXT's.
- Data lives in databases.
- You don't need to know how to build one, just get data out of one.
- This opens up the amount of data you can work with.
- Looks great on your resume!
- This doesn't change modeling approaches or anything else. It only changes where you get your data from.

IV. LEARNING SQL WITH CODE