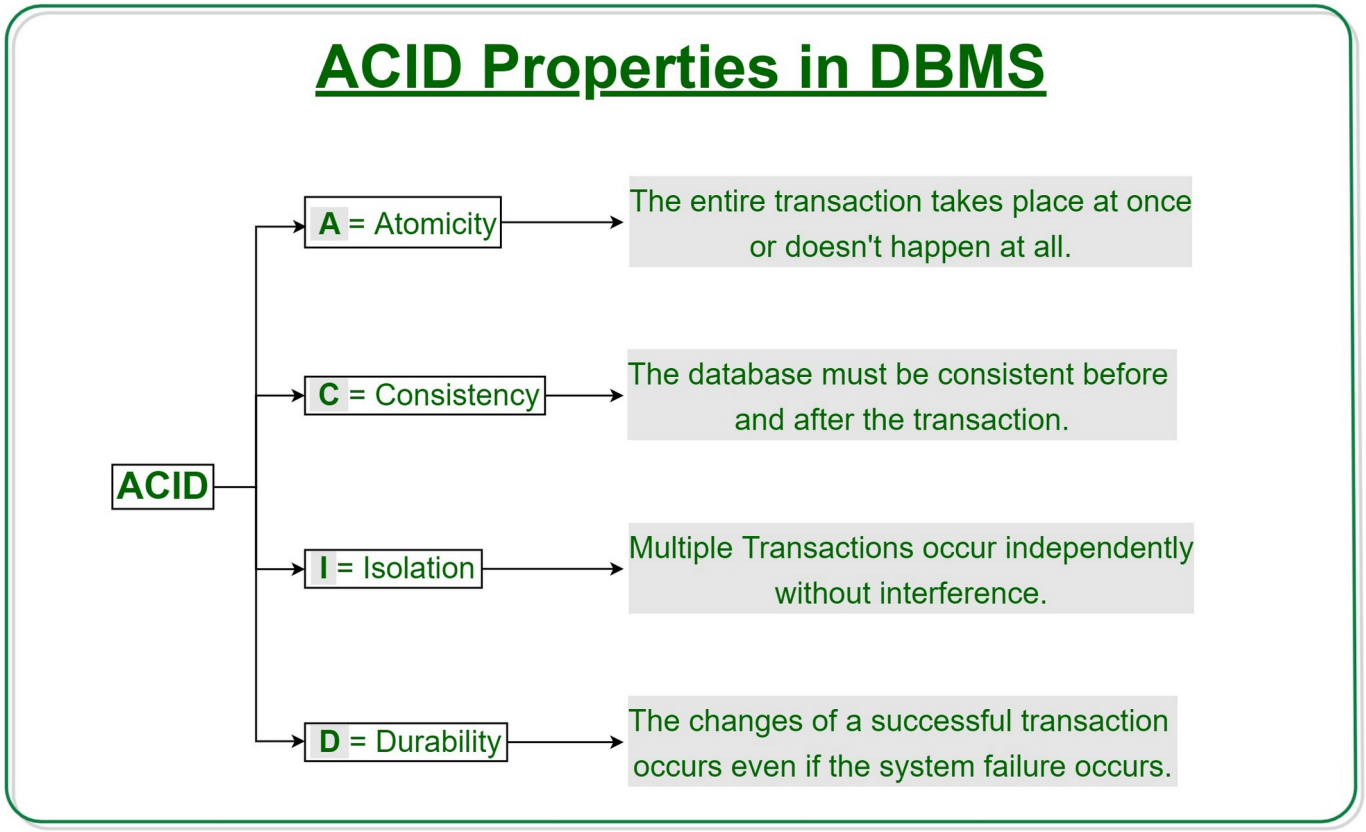
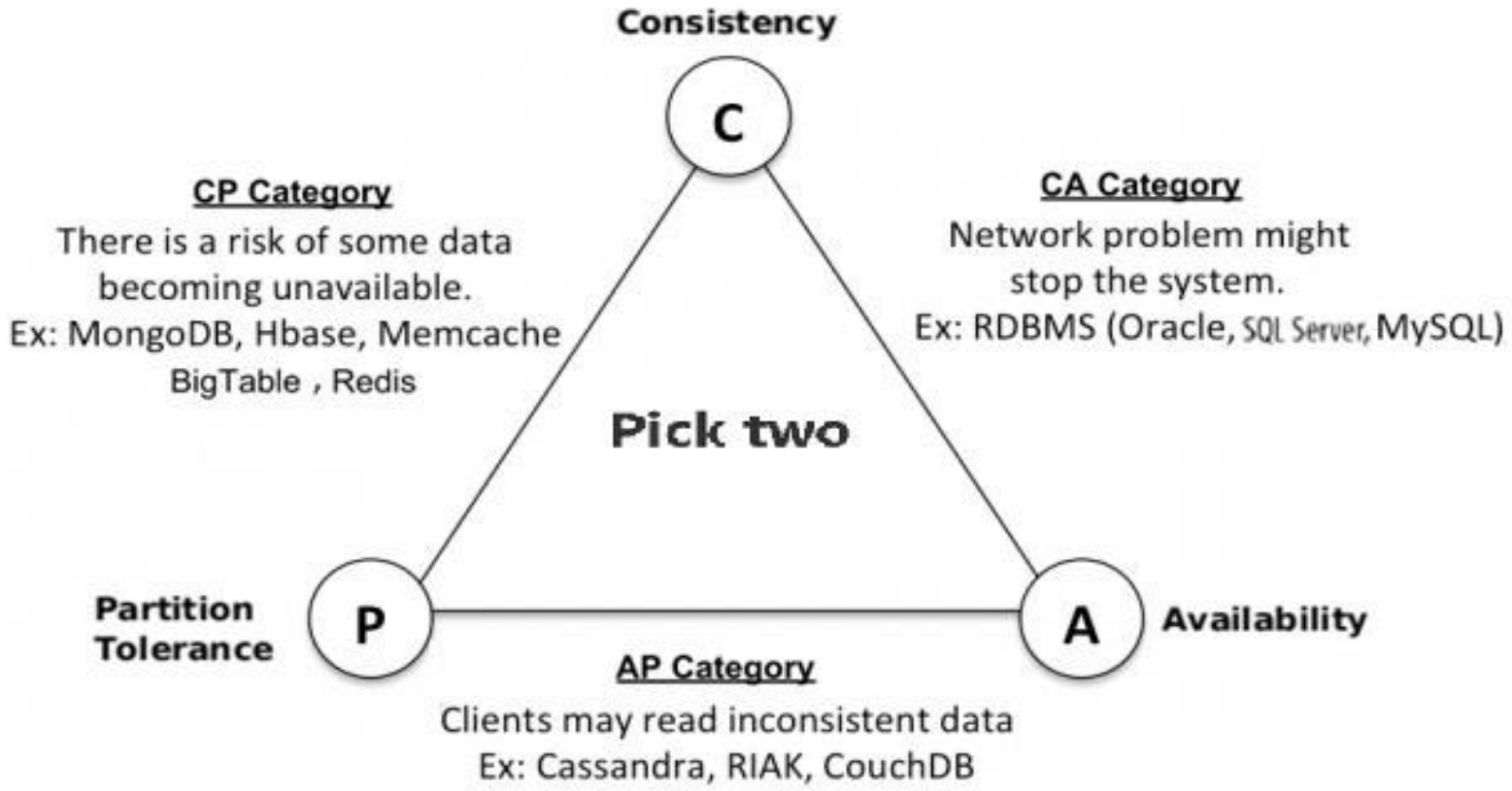
Josh Bodyfelt, Ph.D.

Theory of Databasing

The Electric Kool-Aid ACID Test



Distributed Datastores: CAP Theorem



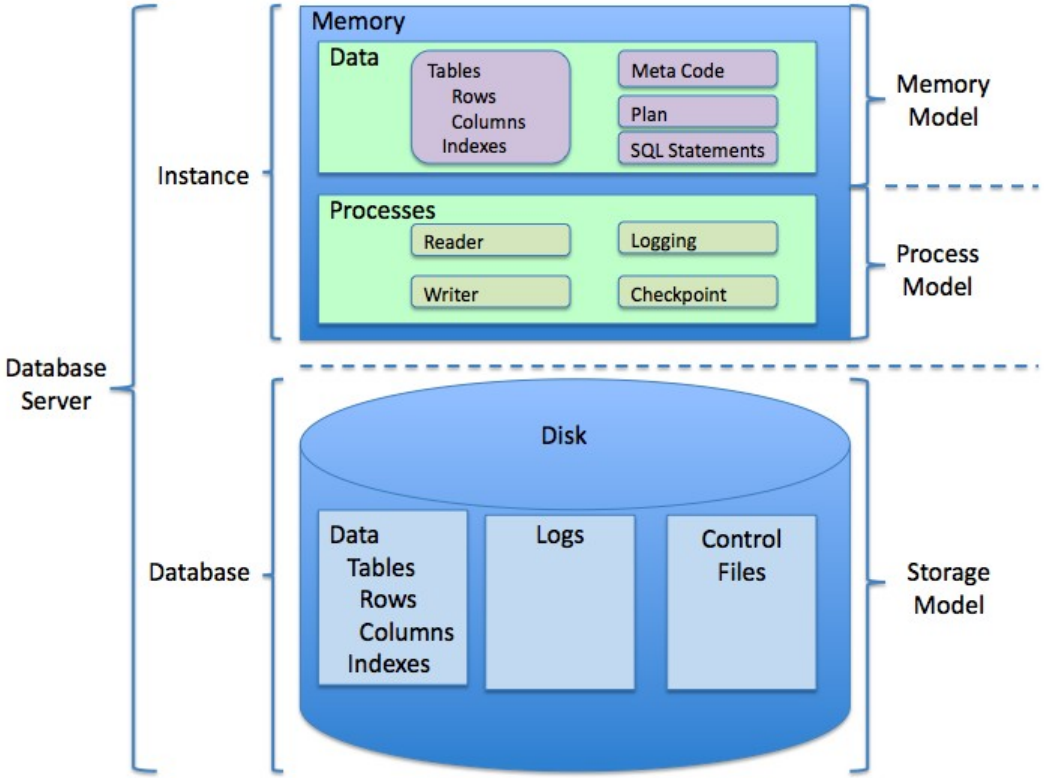
DBMS: Ways to Store Data

- Flat File Database
  - Comma Separated Values (CSV)
- Document Database
  - MongoDB, CouchDB, ElasticSearch
- Key-Value Database
  - Redis, Memcache, Aerospike, LevelDB, Zookeeper
- Relational Database
  - Apache Hive, Oracle, MariaDB, Postgres,... 

- Graph Database
  - AllegroDB, Neo4J, Neptune
- Blockchain Database
  - Sia, BigchainDB, Storj

Relational Database = Tables & Links

- Connect to Database
- Create/Edit Schema
  - *Data Definition Language* (DDL)
- Create/Read/Update/Delete Records
  - *Structured Query Language* (SQL)



Theory of Databasing: Tables

- A **table** consists of *fields* (columns) and *records* (rows):

mov_id	mov_title	mov_year	mov_time	mov_lang	mov_dt_rel	mov_rel_country
901	Vertigo	1958	128	English	1958-08-24	UK
902	The Innocents	1961	100	English	1962-02-19	SW
904	Lawrence of Arabia	1962	216	English	1962-12-11	UK

Theory of Databasing: Tables

```
mov_id |          mov_title          | mov_year | mov_time |  mov_lang  | mov_dt_rel   | mov_rel_country
---------+--------------------------+------------+-------------+--------------+----------------+-----------------
```

- A table and its fields created by DDL:

```
CREATE TABLE movie (
        mov_id int,
        mov_title varchar(255),
        mov_year int,
        mov_time int,
        mov_lang varchar(20),
        mov_dt_rel date,
        mov_rel_country varchar(2)
);
```

Theory of Databasing: Primary & Foreign Keys

- **Primary Key**: Unique Record Identifier – ID Number (*employeeID)*
- **Foreign Key**: Link to another table – usually other table's Primary

TABLE MOVIE

mov_id	mov_title	mov_year	mov_time	mov_lang	mov_dt_rel	**mov_rel_country**
901	Vertigo	1958	128	English	1958-08-24	UK
902	The Innocents	1961	100	English	1962-02-19	SW
904	Lawrence of Arabia	1962	216	English	1962-12-11	UK

TABLE COUNTRY

country_code	country_name
UK	United Kingdom
US	United States
SW	Sweden
CN	China

Theory of Databasing: Primary & Foreign Keys

- **Primary Key**: Unique Record Identifier – ID Number (*employeeID)*
- **Foreign Key**: Link to another table – usually other table's Primary

TABLE MOVIE

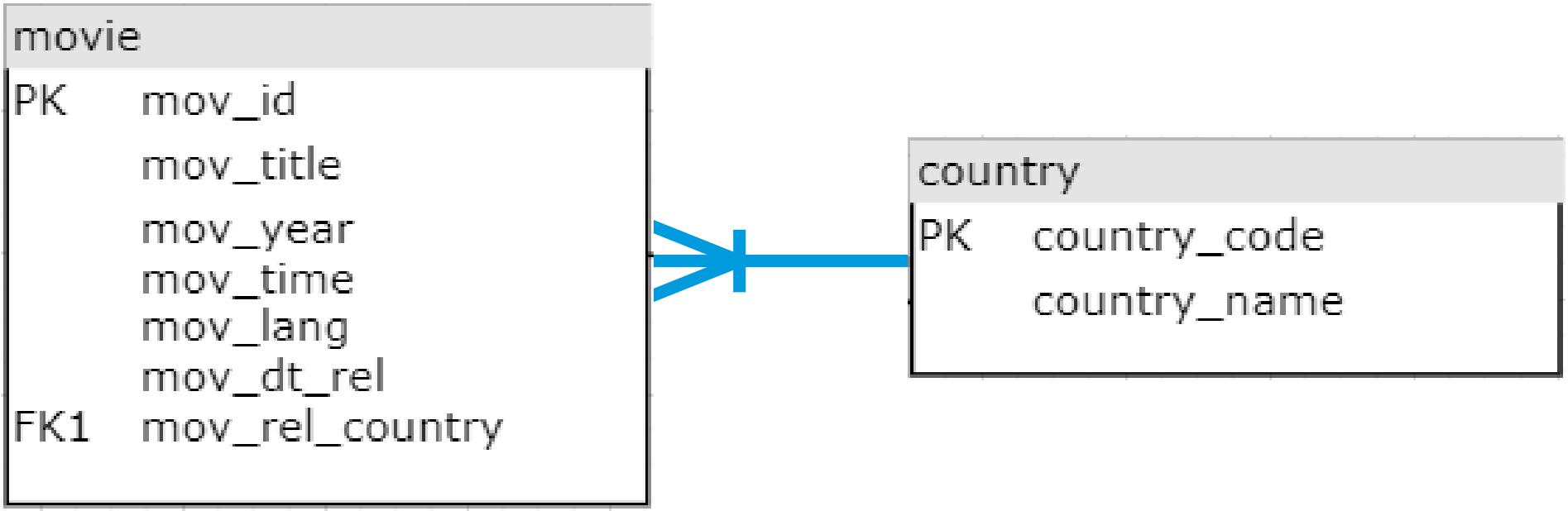
mov_id	mov_title	mov_year	mov_time	...	mov_rel_country
901	Vertigo	1958	128	...	UK
902	The Innocents	1961	100	...	SW
904	Lawrence of Arabia	1962	216	...	UK

TABLE COUNTRY

country_code	country_name
UK	United Kingdom
US	United States
SW	Sweden
CN	China

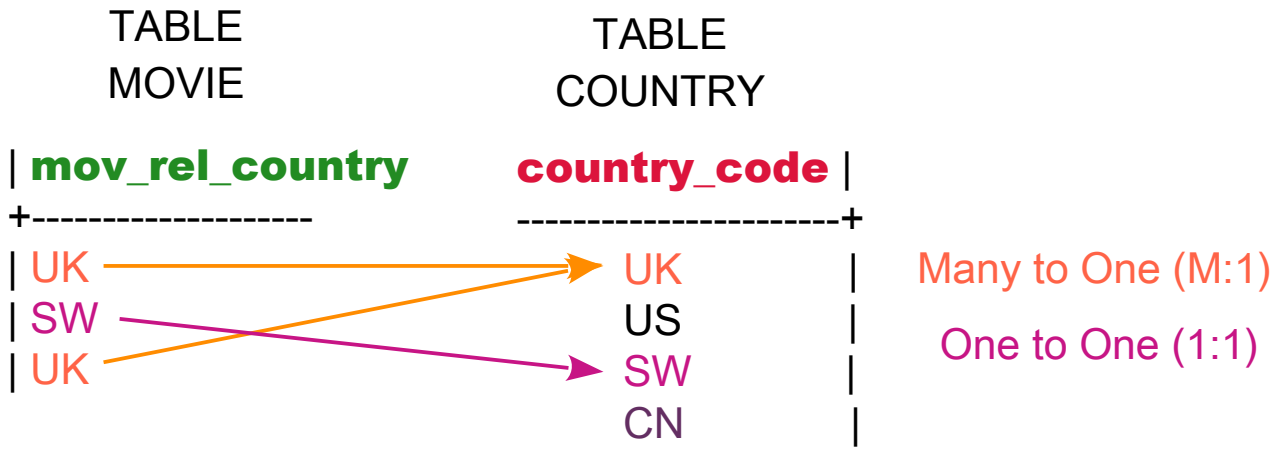
Entity-Relationship Diagram (ERD)

- A **schema** is the structure of tables and relationships:

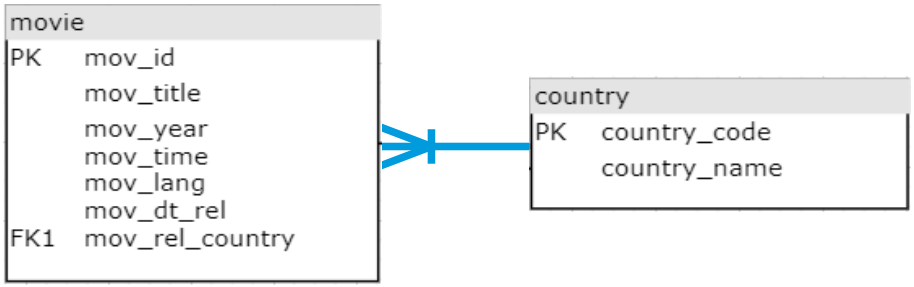


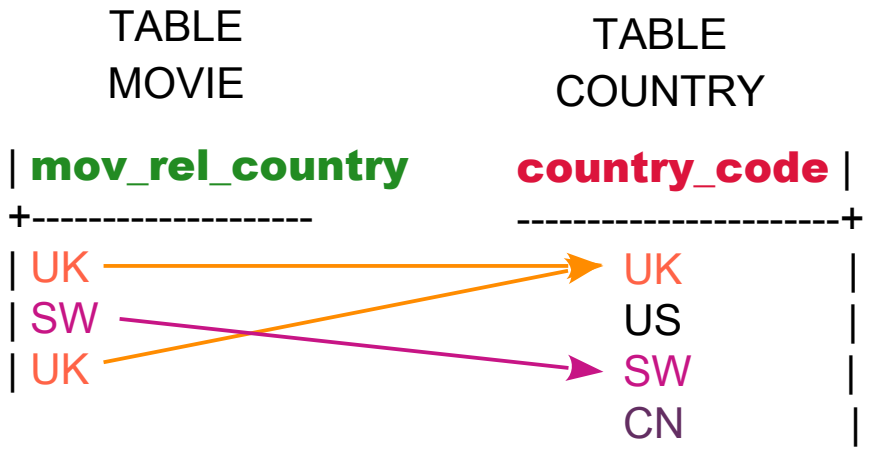
ERD: Cardinality & Ordinality

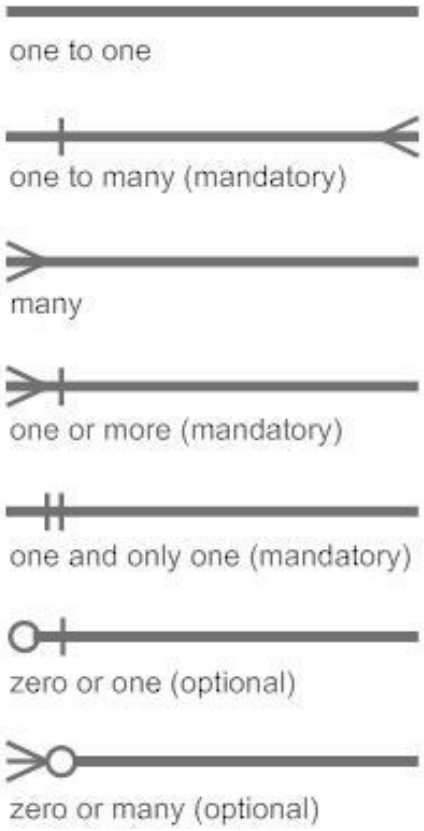
- How single record in table relates to record(s) in other tables
- **Cardinality** – *Maximum* number of records
- **Ordinality** – *Minimum* number of records

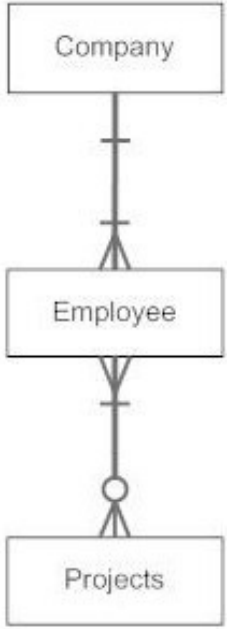


ERD: Cardinality & Ordinality

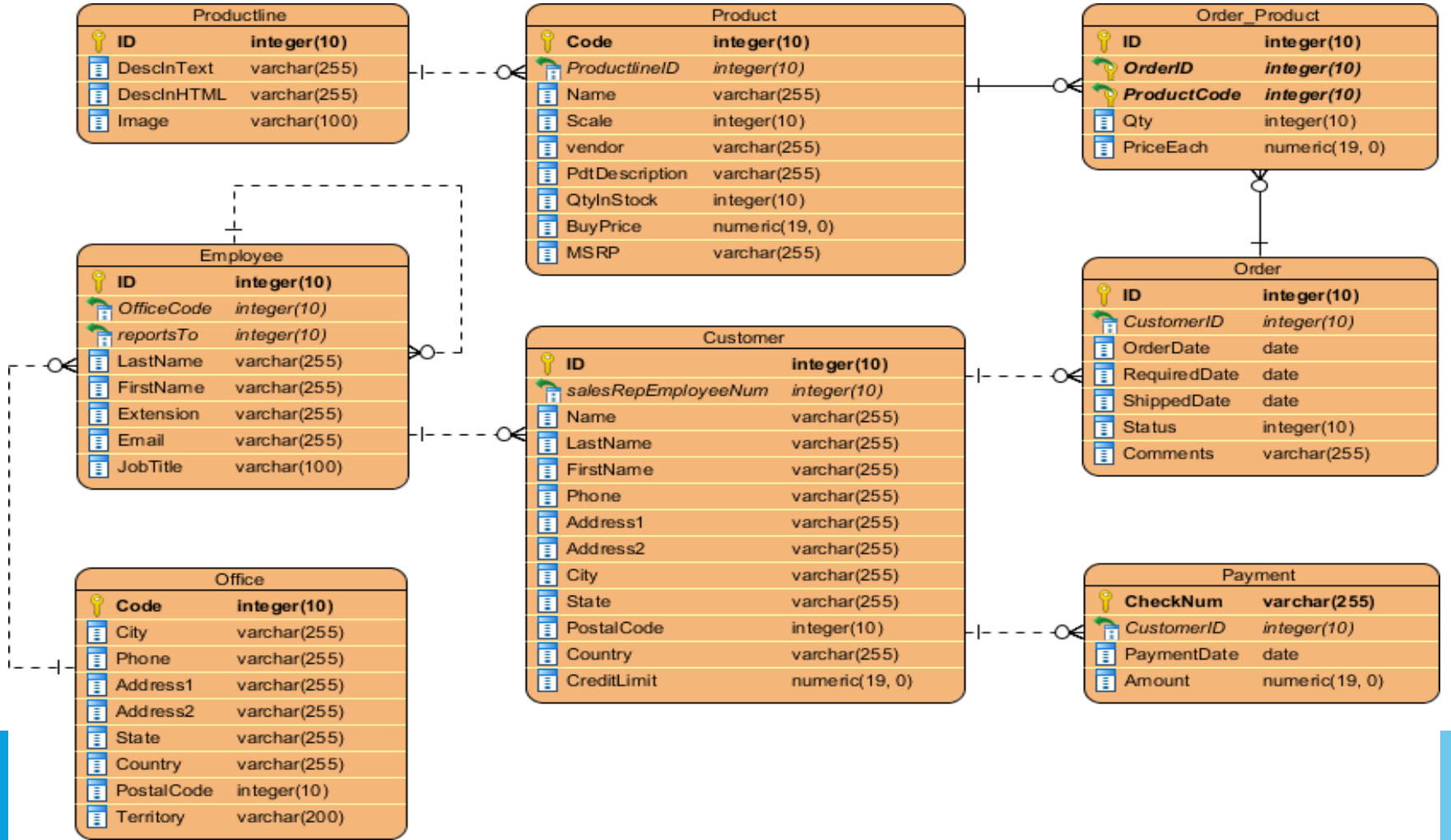








ERD: An Advanced Schema



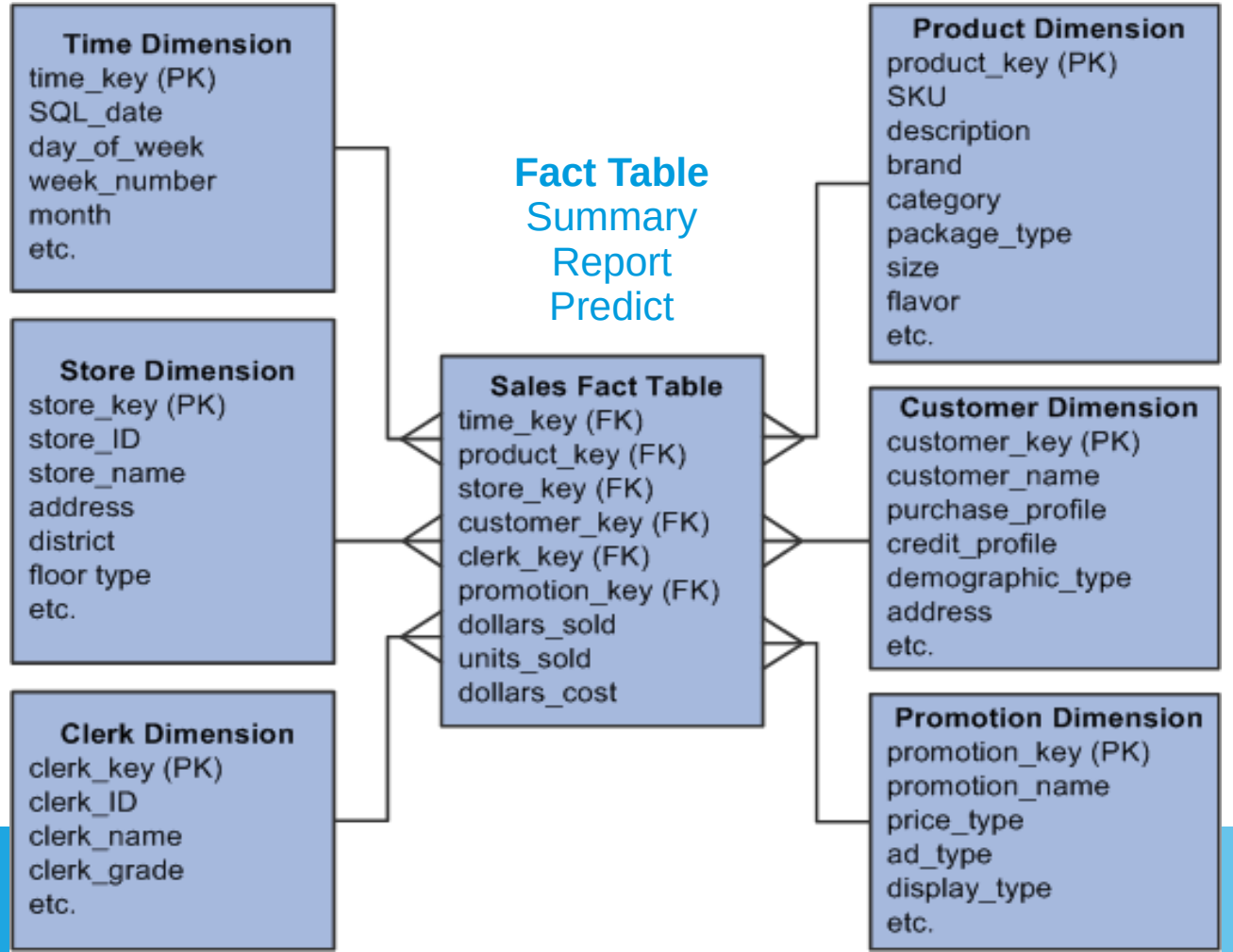
Normalization: Rules to Schema Design

- **Unique:**
  - One table per tracked entity
  - Every table has a primary key
- **Type:** Entity can be physical or logical
- **Conserve:**
  - Only store data once
  - If data can be calculated, don't store
- **Link:** Every foreign key to another table's primary key
- **Simplify:** Many-to-many as its own table

Customer *Order*
Product *Payment*

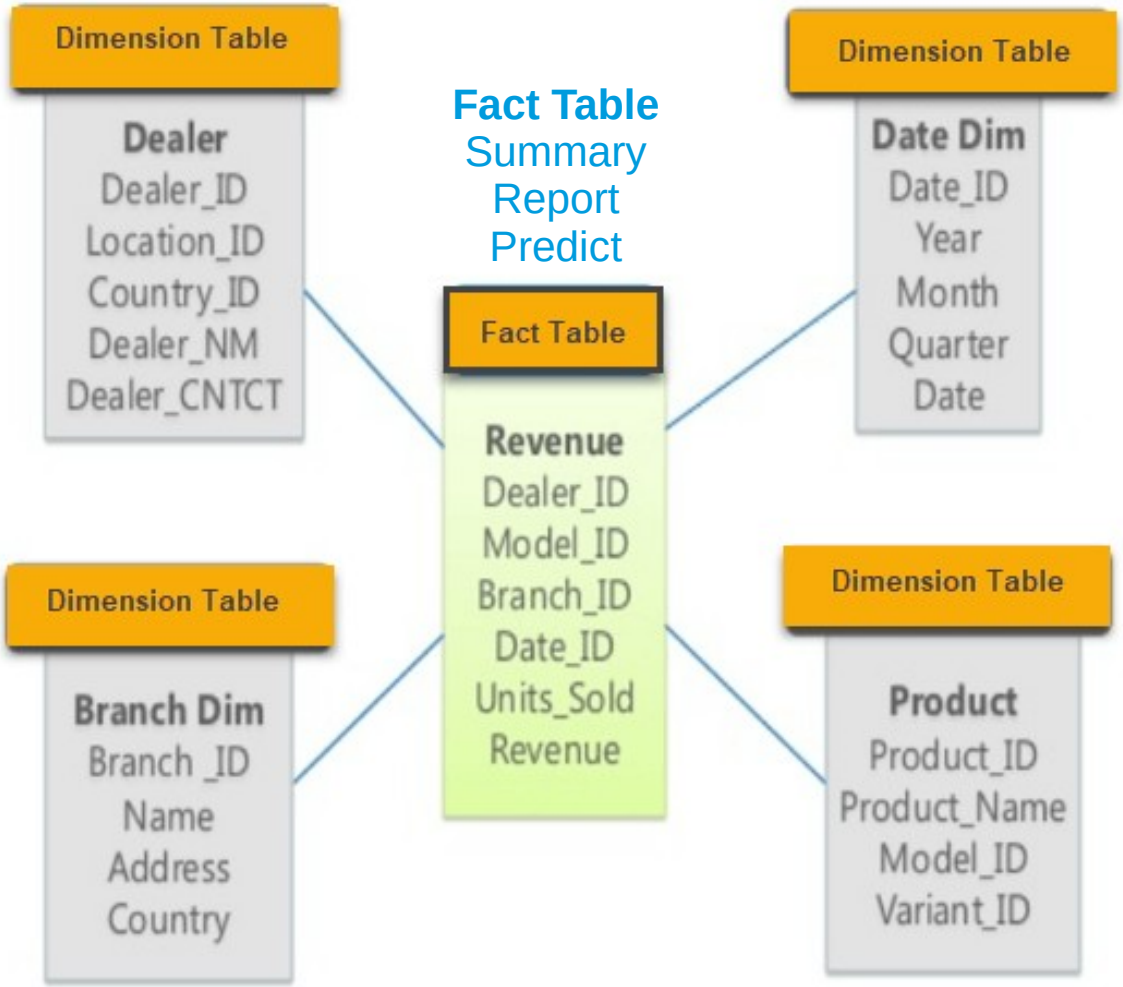
Star Schema: Data Store for Reporting

Dimension Table
A Specific Factor
Relating to Fact



Star Schema: Data Store for Reporting

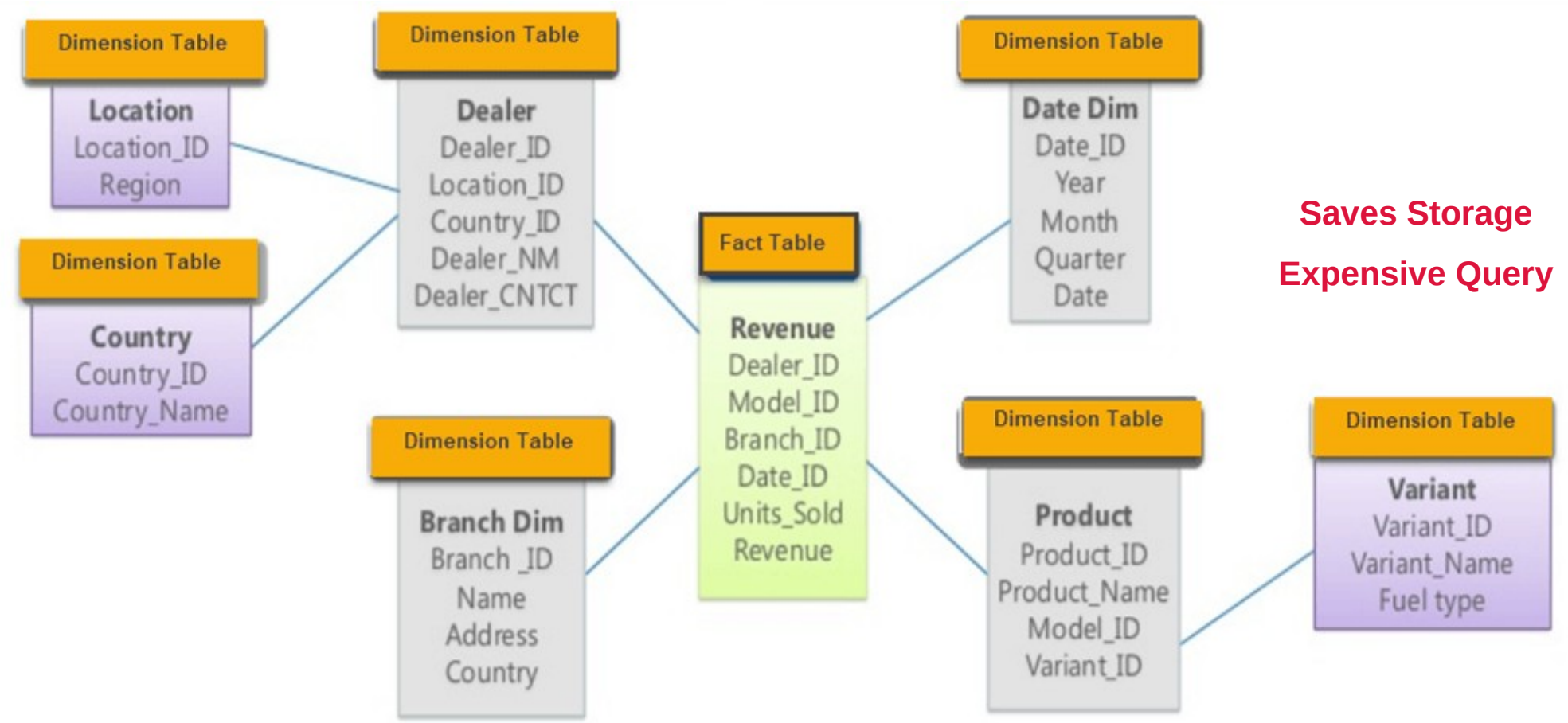
Dimension Table
A Specific Factor
Relating to Fact



Denormalized

Quick Query

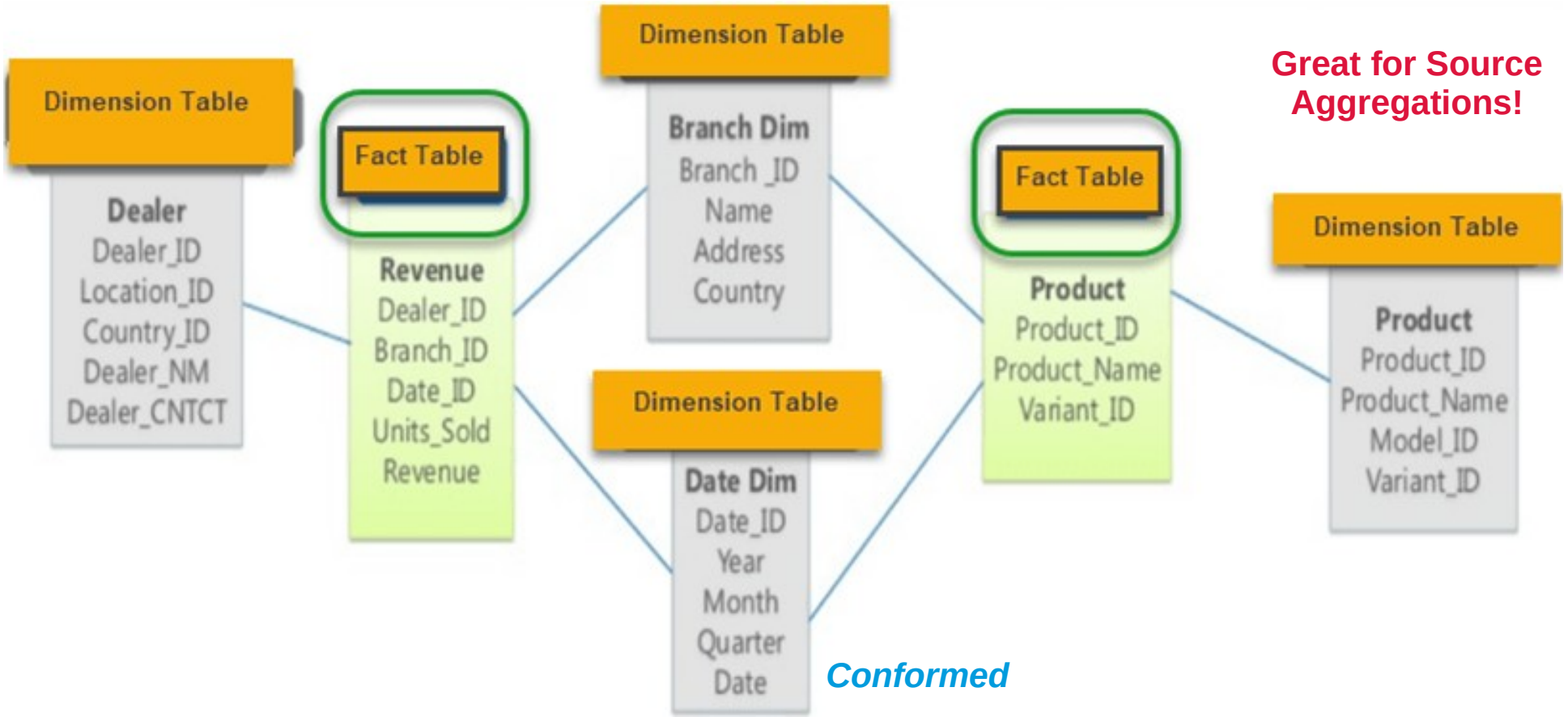
Snowflake Schema



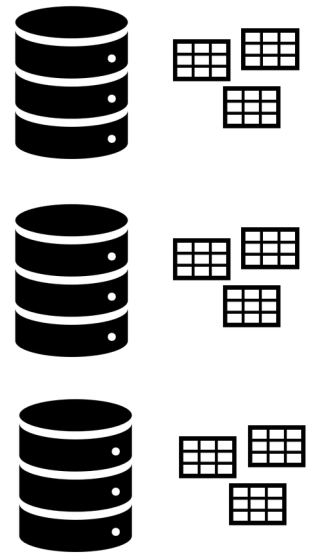
Saves Storage

Expensive Query

Galaxy Schema

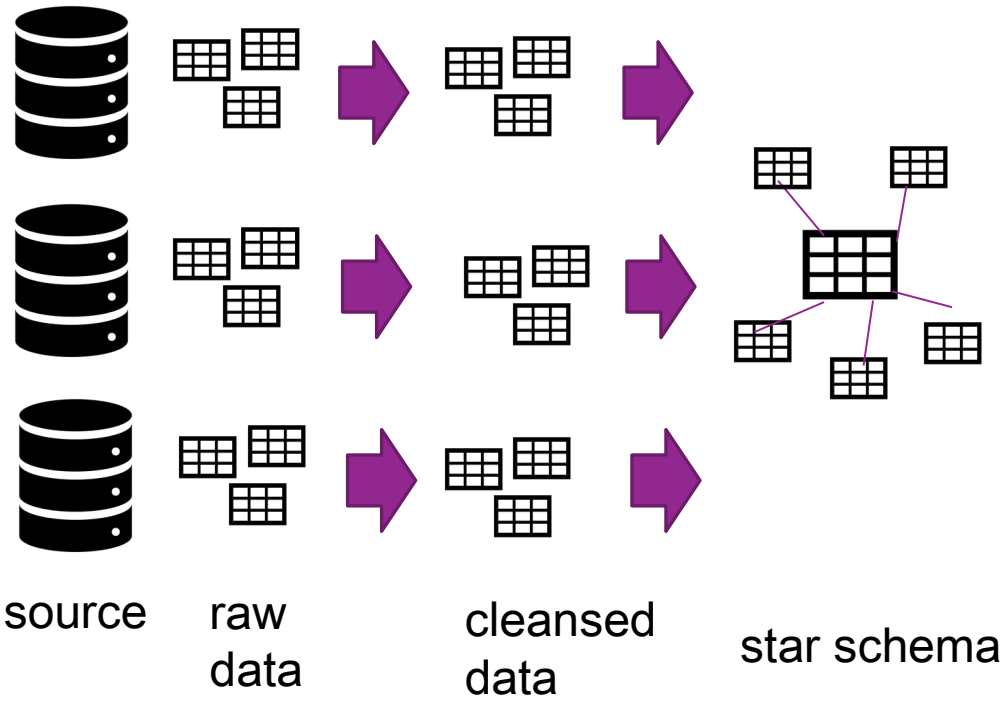


Theory of Databasing: Warehousing



- **Warehousing:** Storing all data for entire enterprise for reporting analytics
- **Consolidation:** Data from many disconnected, irregular sources

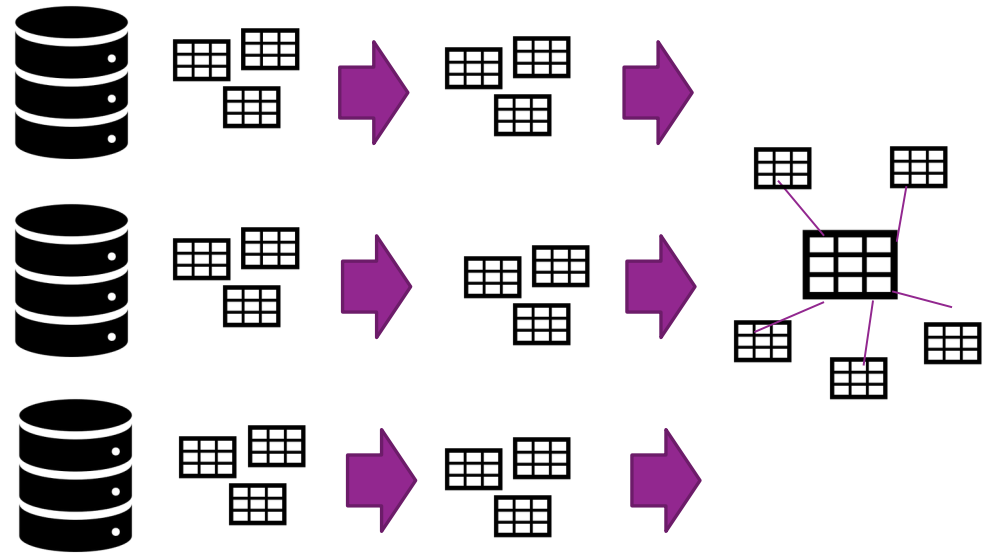
Theory of Databasing: E.T.L.



- **Extract:** Pull data from source systems
- **Transform:** Clean, cast, & calculate data
- **Load:** Insert clean data into schema

Theory of Databasing: Layers

Backend

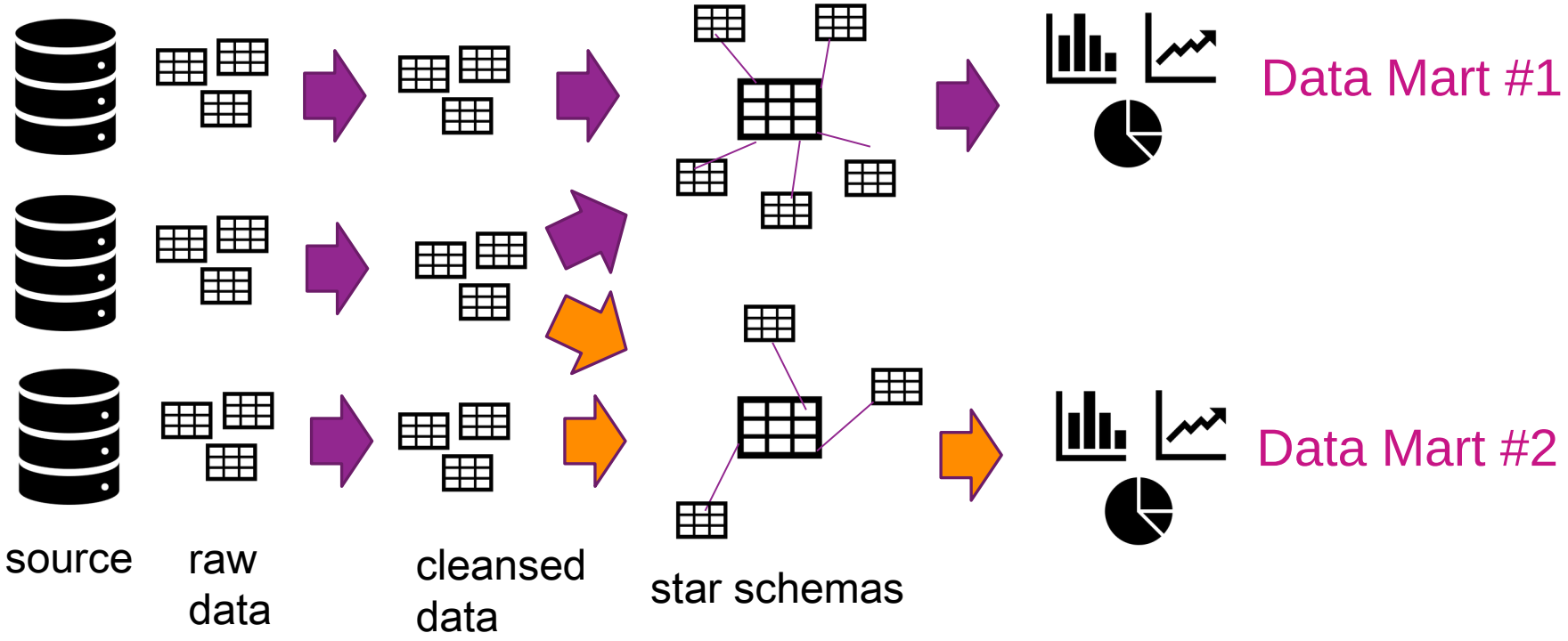


source | raw data | cleansed data | star schema

Frontend

visualisation

Theory of Databasing: Data Marts



The Process of Data Analytics

1. Requirements Elicitation
   a) Establish Stakeholder Stories
   b) Define User Personas
2. Data Collection
   a) Source Discovery & Socketing
   b) Cleansing & Preparation
   c) Brownfield vs. Greenfield
3. Data Modelling
   a) Define dimensions & measures
   b) Define calculations (L.T. vs T.L.)
4. **Model Construction**
   Pythonic DDL & SQL
5. Data Testing
   a) Unit Tests
   b) End-to-End Tests
6. Presentation
   a) Logo Design & CSS
   b) Dashboard (Graphs)
   c) Native vs. Web