Week 4: Monday

Data Literacy

Datastores and Databases

What is a Datastore?

- What do you think?
- Any way to store data.
- A CSV, a JSON file, a database, etc.

Examples of Simple Datastores

```
Title,Author,Year
*The Great Gatsby*,F. Scott Fitzgerald,1925
*To Kill a Mockingbird*,Harper Lee,1960
*The Catcher in the Rye*,J.D. Salinger,1951
*The Grapes of Wrath*,John Steinbeck,1939
```

Examples of Simple Data stores, cont.

Title	Author	Year
The Great Gatsby	F. Scott Fitzgerald	1925
To Kill a Mockingbird	Harper Lee	1960
The Catcher in the Rye	J.D. Salinger	1951
The Grapes of Wrath	John Steinbeck	1939

Examples of Simple Data stores, cont.

```
[
  {
    "title": "The Great Gatsby",
    "author": "F. Scott Fitzgerald",
    "year": 1925
  },
  {
    "title": "To Kill a Mockingbird",
    "author": "Harper Lee",
    "year": 1960,
  },
  {
    "title": "The Catcher in the Rye",
    "author": "J.D. Salinger",
    "year": 1951
  },
  {
    "title": "The Grapes of Wrath",
    "author": "John Steinbeck",
    "year": 1939
  }
]
```

Problems with simple Data stores

- Great thing about simple data stores are they are simple files.
- But they are not great for:
  - Searching
  - Updating
  - etc.
- TLDR: They don't disaggregate the data.
- We cannot combine complicated data.

Problems with simple Data stores, cont.

- They show up in many places where they shouldn't be.
- Renault One F1 team used Excel to manage their parts & build data store: 77,000 lines of it.

- DO NOT DO THIS!
- Too many accidents waiting to happen.
- No way to efficiently connect parts to builds without eliminating data.

Enter Databases

- Oracle (MySql): "A database is an organized collection of structured data, typically stored electronically in a computer system."
- For this reason, some consider excel as a possible database.
- Think of data as individual records of structured information.
- Database == Collection of individual records.
- CRUD: Create, Read, Update, Delete.

Database vs Excel Spreadsheet

- Spreadsheet is a file.
- CRUD operations update the file not individual records.
- Database: CRUD operations only impact individual records.
- Excel spreadsheets filter to show only certain records.
- Databases retrieve individual records.

The Filing System—A Paper-Based Database

Types of Databases

- Relationsal Databases (SQL)
- NoSQL Databases
  - Document-oriented databases
  - Graph Databases
  - Others (Key-Value, column-family, etc.) ← We won't discuss these.

Relational Databases

- SQL: Structured Query Language.
- Tables: Rows and Columns.

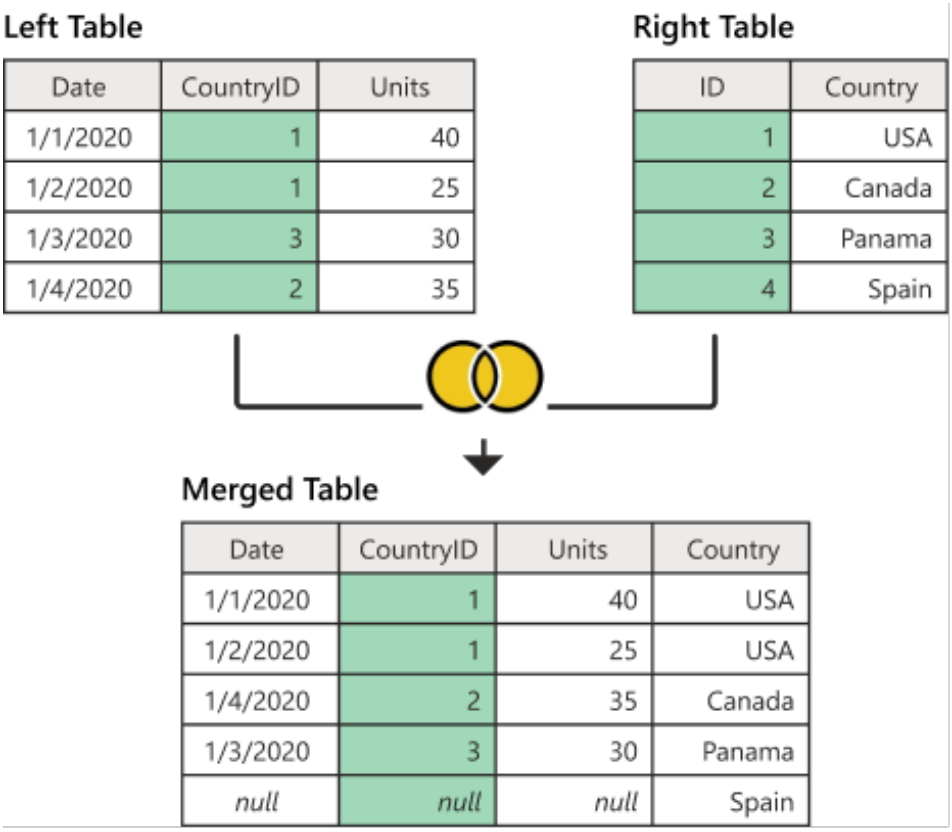
- Rows: Records.
- Columns: Fields.
- Joins: Combining tables.
- Relational database == Collection of tables == Collection of records.
- Fundamental rule: do not duplicate data (normalization).

Relational Databases, cont.

- Represented similarly as a spreadsheet.
- Table of parts:

id	part	company	cost
1	Engine	Renault	1000
2	Tires	Pirelli	500
3	Wings	Red Bull	2000

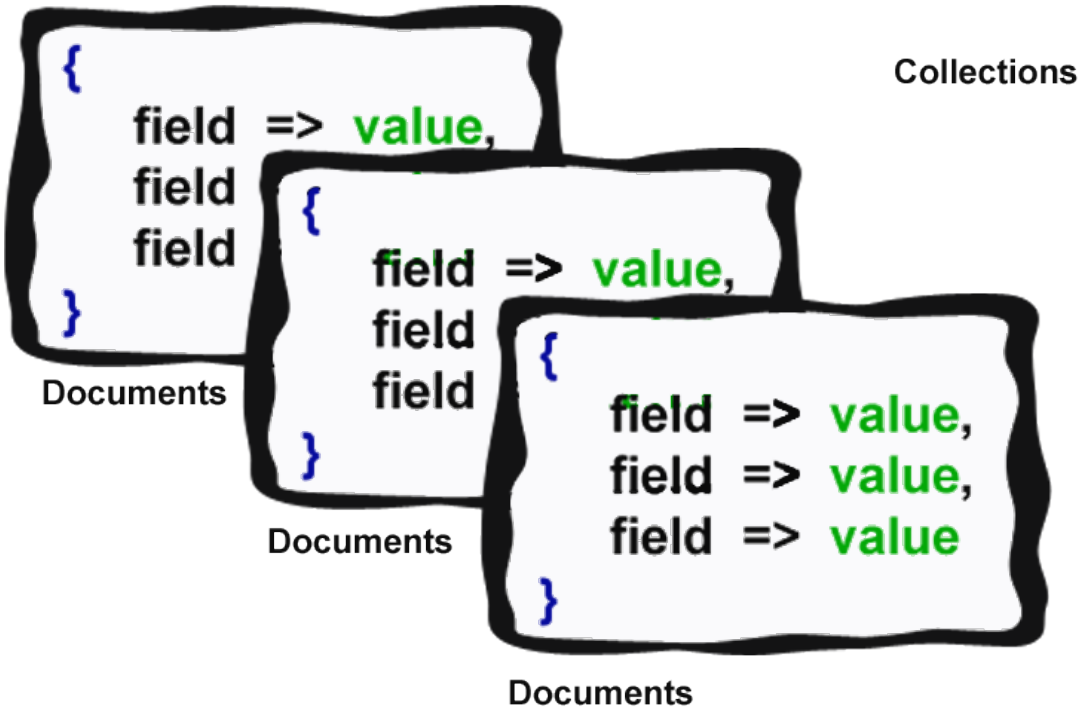
Relational Databases, Joins tables



MongoDB

- NoSQL database.
- Document-oriented database.
- JSON-like documents (BSON).
- No schema.
- No joins.
- No normalization.
- Collections of BSON documents.
- Fundamental rule: If it's queried together, store it together.

MongoDB, cont.



MongoDB, cont.

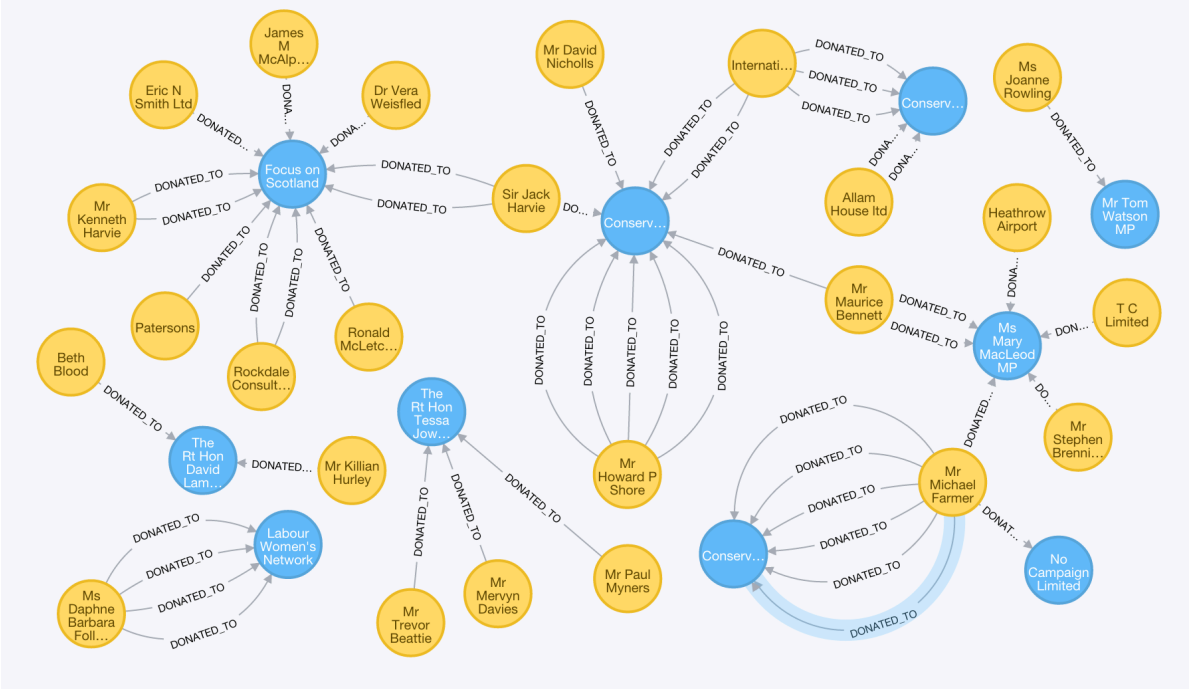
- Collection of Parts Documents:

```
{
  "part": "Engine",
  "company": "Renault",
  "cost": 1000
},
{
  "part": "Tires",
  "company": "Pirelli",
  "cost": 500
},
{
  "part": "Wings",
  "company": "Red Bull",
  "cost": 2000
}
```

Graph Databases (Neo4j)

- Organizes based on nodes and edges.
- Nodes: Labeled Entities (stores information).
- Edges: Labeled Relationships (stores information).
- No schema.
- No joins.
- No normalization.
- Fundamental rule: (People) - [TRAVEL_TO] -> (Places).

Graph Databases, cont.



Graph Databases, cont.

- Great at recursive queries
- Great at traversing relationships

Recursive query, explained

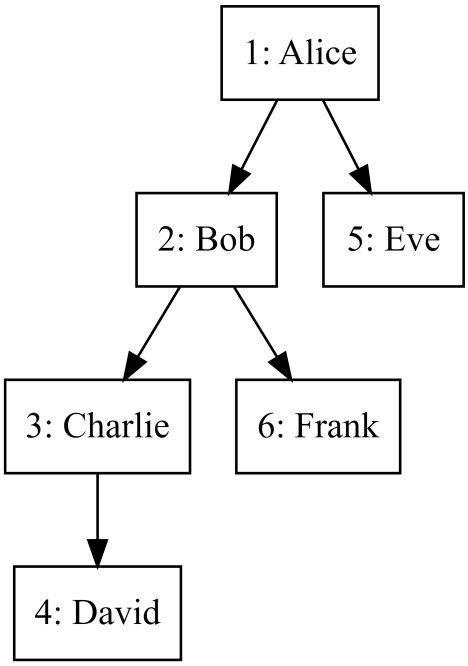
- Employees table

id	name	manager_id
1	Alice	null
2	Bob	1
3	Charlie	2
4	David	3
5	Eve	1
6	Frank	2

- Find all employees who are a part of Bob's team.

Recursive query, cont.

- Graphs databases find this easily



When to use Which?

Type	Use Case
Relational	Relational data (such as parts that go to a car)
MongoDB	Unstructured data, flexibility, non-relational
Neo4j	Relationships, traversing relationships, etc.

Connecting Databases

Using Multiple Databases

- Frequently we mix databases and use them to cross reference one another.
- For example, we might cross-reference census data with crime data.
- Open Data

Public Databases

- Census data is frequently cross-referenced with other data.
- States keep official databases of the geography of the state (roads, borders, county lines etc.)
- Counties keep databases of property ownership (Think Zoom).

Private Partnerships

- Credit reporting agencies.
- Whenever you agree to share your data with third-party partners.