Introduction to SQL Databases

Introduction

While the role of a data scientist does involve fun stuff like building and improving machine learning models, there are many other necessary skills that must be mastered.

Data Cleaning/Preparation is a major task which nearly always must be done before getting to the work of analysis or modeling.

Data Wrangling/Munging or the process of transforming data from the raw form to a usable form or just in moving it from one place to another is another major task. Quite often, the data you will be working with will be stored in a relational database, and in order to access it, you will need to use Structured Query Language (SQL)

What is a Database?

- A systematic collection of data
- Supports **storage** and **manipulation** of data
- Access to a database is usually done through a Database Management System (DBMS) (eg. SQL Server, MySQL, Oracle)

Relational Databases / SQL Database / Transactional Database

- Most commonly used structure in enterprise scenarios
- Data is organized into one or more tables of columns and rows, with a unique key identifying each row.
- Tables are linked together through the use of keys, which ensure **referential integrity**
- Relational databases almost exclusively use SQL (Structured Query Language) to perform transactions.

Structured Query Language (SQL)

SQL is a programming language designed to manage data held in a relational database management system.

SQL comes in different flavors with slight variations on syntax and features depending on the type of RDBM you are working with:

- PostgreSQL
- Transact SQL (T-SQL) for Microsoft SQL Server
- SQLite
- MySQL
- PL-SQL for Oracle

Keys and Referential Integrity

Relationships between tables are encoded through the use of **keys**:

- **Primary key:** One or more columns which uniquely identifies each row. A table will have only one primary key.
- **Foreign key:** One or more columns in another table which refer to the primary key in another table. A table can contain more than one foreign key.

Referential integrity means that when a foreign key value is used, it must reference a valid, existing primary key in the parent table. A breakdown in referential integrity can have undesirable side effects:

- Incomplete data being returned, usually with no indication of an error/"lost" records
- Strange results appearing in reports (such as products without an associated company).

Entity Relationship Diagram (ERD)

Displays the tables and how those tables relate/are connected together.

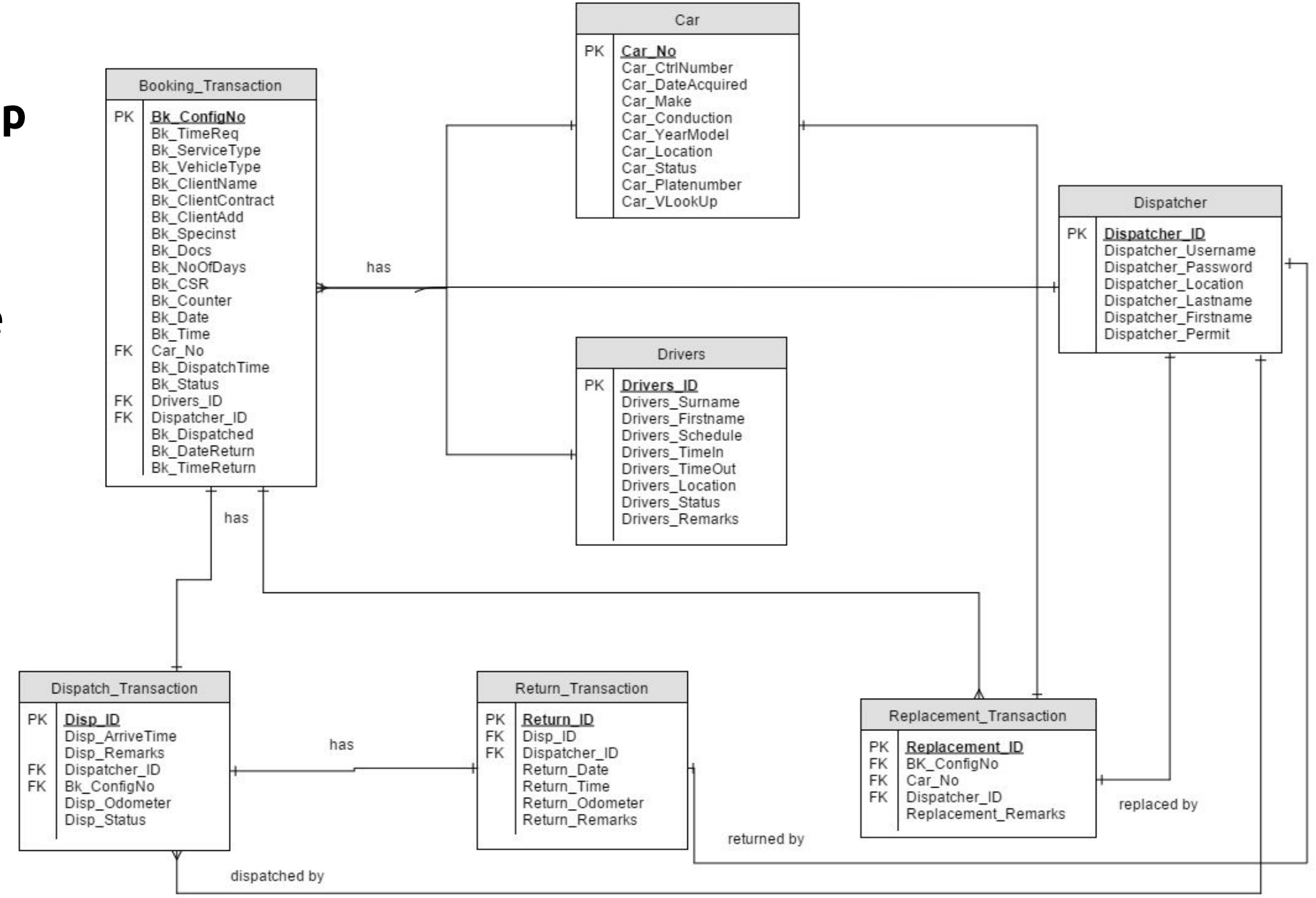

Image: https://sites.google.com/site/dracvbms/entity-relationship-diagram

ACID Properties

These define the key characteristics that SQL databases use to ensure database modifications are saved in a consistent, safe, and robust manner.

- **A**tomic:
  - A database transaction either succeeds or fails.
  - A transaction cannot be completed only partially.
- **C**onsistent:
  - Use of rules and constraints so state is always valid.
  - The data saved can’t violate any of the database’s integrity.
- **I**solation:
  - Transactions happen in isolation; No “mid-air collisions.”
- **D**urability:
  - Once committed a transaction is permanent, regardless of a subsequent system failure.