CS 143

Warren Kim

Contents

0.1 Purpose of a Database

We will be studying (mostly) Relational DataBase Management Systems (RDBMS).

Definition: Database

A **database** abstracts how data is stored, maintained, and processed. It is a system that uses advanced data structures to store and index data.

A database abstracts away the data integrity and file management aspect of CRUD operations. Moreover, a database provides us with a single location for all of the data, even if the database itself is distributed.

0.2 Abstraction Layers

There are three layers of abstraction: physical, logical, and view.

Definition: Physical Abstraction

The **physical abstraction** defines the data and its relationships to other data within the database.

Definition: Logical Abstraction

The **logical abstraction** deals with how we interface with the database.

Definition: View Abstraction

The **view abstraction** refers to specific use cases and filters the data from the logical abstraction.

We start by learning the logical abstraction.

0.3 Instances and Schema

Definition: Schema and Instance

A **schema**[a] is the overall design of a database. It defines the structure of the data as well as how it is organized.

An **instance** of a database is the actual set of data stored in the database at a particular moment in time.

[a]Note: schema can also refer to a relation (table).

0.4 Data Models

Data models define how we design databases and interact with data. We want to answer the following:

(i) How do we define data?

(ii) How do we encode relationships among data?

(iii) How do we impose constraints on data?

Data models are either an Implementation model or a Design mechanism. Implementation models build databases from the ground up while design mechanisms are implemented as features in a database. We discuss five major types (an several niche ones).

0.4.1 Relational

In a relational model, all data is stored as a ***relation***[1]. Rows represent individual n-tuple units (***records***). Columns represent (typed) ***attributes*** common to all records in the relations.

0.4.2 Entity-Relationship (ER)

An entity-relationship model uses a collection of basic objects (***entities***) and define ***relationships*** among them.

0.4.3 Object-Oriented

The object-oriented model is similar to OOP with encapsulation, methods, adn object identity. It was originally an implementation model but is now a design mechanism.

0.4.4 Document (Semi-Structured)

A document model stores records as ***documents***, which do ***not*** have an enforced schema. This allows for more versatility in the type of data stored in the database.

0.4.5 Network/Hierarchical/Graphical

A graph model is analogous to how we think. Records are stored as ***nodes*** and relationships between records as ***edges***.

0.4.6 Vector

A vector model stores records as ***vectors*** in $\mathbb{R}^n$, and are stored in a way that enables efficient retrieval and comparison (e.g. nearest neighbor[s]).

0.4.7 Key-Value

A key-value model stores data as a key-value pair (typically using a hash function). In this model, data typically lives in RAM as opposed to disk.

0.5 Database Languages

There are two main semantic systems when working with databases:

(i) Data Manipulation Language (DML)

(ii) Data Definition Language (DDL)

Note that a relational model typically uses SQL for both DDL and DML.

0.5.1 Data Manipulation Language

DML's can either be procedural or declarative.

Definition: Query

A ***query*** is a written expression to retrieve or manipulate data.

[1]Note: tables are an implementation of relations.

Aside: A Note on SQL

SQL is a declarative language, and as such, it is hard to perform sequential or nontrivial[a] computations in SQL. To remedy this, a common option is to write an ***ETL job*** in another language (pick one). We **E**xtract the data from the database (using a connection driver), **T**ransform the data using another lanuage (pick one!), and **L**oad the data into a new table using the same driver. We can schedule these jobs using something like `cron`.

[a]Nontrivial: Any computation where we have to specify *how* to perform the computation.

0.5.2 Data Definition Language

DDL's specify a schema: a collection of attribute names and data types, consistency constraints, and optionally storage structure and access methods. There are four types of consistency constraints:

(i) Domain constraints define the domain of an attribute (e.g. `tinyint`, `enum`, etc.).

(ii) Assertions are business rules that must hold true (e.g. an enforced prerequisite for a class must be present in your transcript before you can add a class to your study list).

(iii) Authorization determines who can do what (e.g. full CRUD, read-only, etc.).

(iv) Referential integrity ensures that links from one table to another must be defined (Suppose we have two relations R, R'. If there is a link $f : R \to R'$, then f is surjective).

0.6 Data Storage and Querying

Definition: Storage Manager

A ***storage manager*** that abstracts away how the data is laid out on disk.

A storage manager is helpful because reading data from disk to RAM is *slow*, and the storage manager handles swapping[2] and makes retrieval efficient.

Definition: Query Manager

A ***query manager*** takes the DML statements and organize them into a *query plan*[a] that "compiles" a query (using relational algebra) and executes the instruction(s).

[a]Note: The query plan dictates the performance of a query.

0.7 Keys

Aside: A Note on Context and Instance

Based on **context** means that the given data is a subset of the complete dataset.
Based on **instance** means that we treat the given data as the complete dataset.

[2]Swapping: Virtual memory in CS111!

0.7.1 Superkey

Definition: Superkey

A **superkey** is a set of one or more attributes that uniquely identifies a record (tuple) and distinguishes it from all other records in the relation.

Formally, let R be a relation with a set $S = \{a_1, a_2, \ldots, a_n : a \text{ is an attribute of } R\}$. A **superkey** is a subset $s \subseteq S$ such that s uniquely identifies each n-tuple in R.

Note that the superkey $s = S = \{a_1, a_2, \ldots, a_n\} = \bigcup_{i=1}^{n} \{a_i\}$ is called the ***trivial superkey***. Additionally, $\emptyset$ is not a superkey. Further note that for every relation R, there exists at most $2^n - 1$ superkeys where n is the number of attributes.

0.7.2 Candidate Key

Definition: Candidate Key

A **candidate key** is a superkey such that no subset of the candidate key is a superkey; i.e. it is the minimal superkey. A candidate key may be null.

Formally, let R be a relation with a set $S = \{a_1, a_2, \ldots, a_n : a \text{ is an attribute of } R\}$. A **candidate key** is a superkey $s \subseteq S$ such that for every propery subset $t \subsetneq s$, t is not a superkey.

0.7.3 Primary Key

Definition: Primary Key and Composite Key

A **primary key** is a candidate key (chosen by the database designer) to enforce uniqueness for a particular use case. A primary key cannot be null.

A **composite key** is a candidate or primary key that is composed of one or more attributes.

0.7.4 Foreign Key

Definition: Foreign Key

A **foreign key** is a set of attributes that links tuples of two relations.

Formally, let R, R' be relations with sets $S = \{a_1, a_2, \ldots, a_n : a \text{ is an attribute of } R\}, S' = \{a'_1, a'_2, \ldots, a'_n : a' \text{ is an attribute of } R'\}$. A **foreign key** is a key $s \subseteq S$ of R that maps to the primary key $p \subseteq S'$ of R'.

Foreign keys are used to enforce referential integrity constraints; i.e. a foreign keys in a relation R are used to protect data in R from being orphaned and/or inconsistent.