CENG315

Information Managment Lecture Notes

Ege Özkan

October 22, 2020

Chapter 1

Introduction - October 15, 2020

1.1 Databases and Database Systems

Databases hold data. Database systems are software systems that manages the records in a database. There are five fundemental requirements for a database system.

- Database systems must be persistent, data must be storable and remain for the future.
- Databases must be able to handle getting large.
- Databases should be sharable, multiple users should be able to reach it at the same time.
- Databases must be kept accurate.
- Databases must be usable.

1.1.1 Record Storage

Databases can be made persistent in different ways.

Storing database records in text files

- Simplest approach.
- One file per record type.
- Each record could be a line of text, with its values seperated by tabs.

1	joe	2020
2	amy	2013
3	lee	2000

Its advantages are the database system has to do very little, and a use could easily examine and modify the files with a text, but it is slow. *(?)*

1.1.2 Data Models and Schema

Data models are different ways to express connections between records while Schemas are the implementations of these methdos for a specific database.

File-system v. Relational

In the file system modal, each record type has a file, with one record per line, programs that read and write to the file is responsible for understanding this. In the relational data model, each record type has its **table** and each record has **fields** for each value. User access to the database happens via this record and field model and records that fit certain conditions can be quarried.

These models are at different levels of abstraction, relational model is a **conceptual**

model, since there is no need to know **how** schemas are specified and implemented, the conceptual schema describes what the data *is*. Whereas the file-system is called a **physical model**, physical schemas say how the data is *implemented*.

Physical Data Independence

A conceptial schema is certainly nicer to use than a physical scheme. Operations on a conceptual schema is implemented by the database schema. Database system has a **database catalog** that contains descriptions of the physical and conceptual schemas. Givven an SQL querry, the database system translates the conceptual abstraction to the physical one and interract with it on the users behalf. If the user does not have to deal with the physical level, this is called the Physical Data Independence.

It is easy to use, quaries are optimized automatically and it is isolated from changes to the physical schema.

Logical Data Independence

The set of tables personalized for a particular user is called the user's **external schema**. If users can be given their own external schema in a database system, it is told that this Database System supports Logical Data Independence.

It has three benefits:

- Each users gets a customized external schema, they see only the information they need.
- The user is isolated from changes to conceptual schema.
- It is safer.

STUDENT(SId, SName, GradYear, MajorId)
DEPT(DId, DName)
COURSE(CId, Title, DeptID)
SECTION(SectId, CourseId, Prof, **Year**)
DEPARTMENT(DId, Name)

Figure 1.1: An example schema

1.2 Relational Databases

The relational modal is a conceptual model since its schemas do not depend on the pyhsical level.

1.2.1 Tables

The database is organized into **tables**, which contain zero or more **records** (ie: table rows), and at least one **fields** (ie: the columns of the table.) Each record has a value for each field, and all fields has a specific **type**. Often, when discussing tables, the type information ignored.

Null Values

A `null` value denotes a value that *does not exist* or is *unknown*. It occur if the data collection is incomplete or if data has not arrived yet.

1.2.2 Superkeys and Keys

In the relational model, the access to data is not handled by indices. Instead, a record must be referenced by specifying field values. Since not all values are guranteed to be unique for all users, a unique identifier field is called a **superkey** to distinguish it. Adding a field to a superkey, will generate another superket. A **key** is a superkey with the property that no subset of its fields is a super key.

Primary Keys

In the Schema at Figure 1.1's, `STUDENT` table `SId` is a key. Whereas in `SECTION` there may be multiple keys if each professor teaches only one class. Therefore, since a table may have multiple keys, a key is chosen as a **Primary Key**, whose values *should never be null*, and who is used to refeer to each record.

For instance, in Figure 1.1, `STUDENT` table, `SId` can be the primary key. This is no coincidance, IDs are most times fit to be primary keys.

Foreign Keys

The information in a database is split among tables, these are not isolated from each other, a **foreign key** is a field (or fields) of one table which corresponds to the primary key of another table. For instance, in Schema at Figure 1.1, `CourseId` of the `SECTION` table is a foreign key.

Foreign Keys can be used to create logical connections between different types of records. In the Schema at Figure 1.1, `CourseId` of the `SECTION` table creates a logical connection between the `SECTION` table and `COURSE` table, since the objects these represent in real life, Sections and Courses are bound by a logical connection as well. (Each section is a section of a course).

Foreign Keys and Referential Integrity

The specification of a foreign key asserts **referential integrity**. Which requires each non-null foreign key value to be the key value of some record. Database system must ensure that if the primary keys of a table is modified in some ways, the foreign keys in other tables refeering to primary keys must also be updated accordingly, or set to `null` in worst case scenerio.

1.2.3 Constraints

A **constraint** describes the allowable states that fields can have in a table. There are four important kinds of constraints. **Null Value Constraints** limit fields to not have `null` values. **Key constraints** specify that two records cannot have the same value. **Referential integrity constraints** specify referential integrity, finally **integrity constraints**.

Integrity constraints

These constraints encodes *business rules*. They can detect bad data entry and can enfore the *rules* of the organization. They may apply to tables, individual records or the entire database.

1.2.4 Tavle Specification in SQL

Listing 1.1: the SQL specification of the STUDENT table

```
create table STUDENT (
        SId int not null,
        SName varchar(10) not null,
        MajorId int,
        GradYear int,

        primary key (SId),
        foreign key (MajorId) references DEPT
                on update cascade
                on delete set null,
        check (SId > 0),
        check (GradYear >= 1863)
)
```

In Listing 1.1 we can see constraints and fields. The action specified with the `on delete` and `on update` keywords can be one of the following:

Cascade causes the same query to apply to each foreign key record.

Set null causes the foreign key values to be set to null.

Set default causes the foreign key values to be set to their default value.

No action causes query to be rejectd if there exists and effected value with the foreign key.

Chapter 2

Relational Algebra - October 22, 2020

ID	Name	Dept. Name	Salary
22222	Einstein	Physics	95000
12212	Tesla	Physics	4354

Table 2.1: Instructors.

2.1 Structure of Relational Databases

Databases are structured with atrributes and values as tuples corresponding to those attributes.

2.1.1 Attributes

The domain of teh attribute is a set of allowed values. Attribute values are normally required to be **atomic**.

The **null** value is a special value that signifes that the value is unknown, or does not exist, it is a member of every domain. However, it causes complications.

2.1.2 Schema vs Instance

A database schema is the logical structure of the database. `instructor(ID, name, dept_name, salary)`. A database instance is the snapshot of the database in a given time.

Using common attributes in relation schemas is one of *(?)*. There is also need for a *(?)*.

2.1.3 Keys

A **superkey** is a set of one or more attributes that allow us to identify uniquely a tuple in relation. Let $L \subset R$, superkey K is a **candidate key** if K is minimal. One of the candidate keys is selected to be **primary key**, they should be chosen such that its attribute values are never or very rarely changed.

Foreign key constraint states that value in one relation must appear in another. **Referencing relation** is the relation that refeers to another and **Referenced relation** is the reference that is being referenced.

2.2 Relational Query Languages

A **query language** is a language in which a user requests infromation from the database. **Relational algebra** provides a set of operations that take one or more relations as input and return a relation as an output.

2.3 Operations of Relational Algebra

Relational algebra provides operations that take relations as input and returns relations as output.

2.3.1 Select Operation

Select operator selects $\sigma_p(r)$ (or `select(r, p)` to denote the selection of rows (horizontal selection) to denote selection on relation r with respect to predicate p.

For instance, $\sigma_{A=B \wedge D>5}(r)$ would select tuples of relation r, such that its A and B attributes are equal and values of D attribute is greater than 5

On the Table 2.1, $\sigma_{\text{dept\_name="Physics"(instructor)}}$ would return a tuple of instructors whose department is Physics.

Selection predicate can take comparasions using $=, \neq, >, \geq, <, \leq$ and multiple predicates can be combined using **connectives.** $\wedge, \vee$ and $\neg$.

For instance on the department table with schema `department(dept_name, building, budget)`, $\sigma_{\text{dept\_name=}\texttt{building}}(\text{department})$ would return departments whose names equal to their building's name.

2.3.2 Project Operation

An unary operation that returns its argument relation with certain attributes left out. $\Pi_{A_1,A_2,A_3,\ldots,A_k}(r)$ or `project(r,` $A_1, A_2, A_3, \ldots, A_k$`)` where A_n are attribute names and r is a relation.

In essance project operation returns tuples with only the values whose attributes are listed in the operation.

2.3.3 Composition of Relational Operations

Since the result of a relational operation are itself a relations, operations can be given as input to other operations, ie: they can be composed together into a **relational-algebra expression**, finding the names ofa ll instructors in the physics department can be done by:

$$\Pi_{\texttt{name}}(\sigma_{\texttt{dept\_name="Physics"}}(\texttt{instructor})) \tag{2.1}$$

2.3.4 Cartesian Product Operation

Composes two relations together to a single product, `instructor` $\times$ `teaches` relation, where `instructor(id, name, dept_name, salary)` and `teaches(id, course_id, year` results in the relation `instructor`$\times$`teaches(instructor.id name, teaches.id, course_id, year)`

However, as one can see, common attributes are not joined, therefore the cartessian product may not (and most likely will not) result in logical results.

When to attribute names are the same, they can be distinguished by attaching the name of the relation prior to the attribute name.

2.3.5 Join Operation

To avoid the mistake of illogical results, one can write:

$$\sigma_{\texttt{instructor.id=teaches.id}}(\texttt{instructor} \times \texttt{teaches}). \tag{2.2}$$

The join operator is the equivalent of this expression. **Natural join** operation is denoted by $\bowtie$ Outputs of the rows from the two input relations that have the same value on all atributes that have the same name is joined.

Consider relations r and s, let θ be a predicate on attributes in the schema $r \cup s$. The join operation $r \bowtie_\theta s$ is defined as $r \bowtie_\theta s = \sigma_\theta(r \times s)$

Such as $\texttt{teaches} \bowtie_{\texttt{teaches.id=instructor.id}} (\texttt{instructor})$ is equivalent to $\sigma_{\texttt{instructor.id=teaches.id}}(\texttt{instructor} \times \texttt{teaches})$

2.3.6 Union Operation

The union operation $r \cup s$ combines two relations as long as they have the same **arity** (number of attributes) and the attribute domains are compatible. (Same indexed attributes have the same domain.)

The expression $\pi_{\texttt{course\_id}}(\sigma_{\texttt{semester="Fall"} \wedge \texttt{year}=2017}(section) \cup \pi_{\texttt{course\_id}}(\sigma_{\texttt{semester="Spring"} \wedge \texttt{year}=2018}(section)$ on the relation `section` with schema `section(course_id, sec_id, semester, year, building, room, number, time_slot_id)` wil select `course_id` row of the course that are though on Fall 2017 **or** Fall 2018.

2.3.7 Set Intersection Operation

Set intersection $s \cap r$ works exactly the same (and have the same assumptions.), but instead of working like *or*, it works like **and**.

2.3.8 Set Difference Operation

Set differnce $s - r$ works similar to intersection and union, but it selects those tuples that are on the first relation and **not** on the second relation.

2.3.9 Rename Operation

Given the relational algebra expression E, the expression $\rho_x(E)$ returns the expression E under the name X.

It can also return an output whose attribute names are changed when they are listed $\rho_{x\{A_1,A_2,\ldots,A_n\}}(r)$.

2.3.10 Assignment Operation

The assignment operation $\leftarrow$ works like assignment in a programming language, relation algebra expressions can be assigned to temporary relation variables.

$$\begin{aligned} \texttt{Physics} &\leftarrow \sigma_{\texttt{dept\_name="Physics"}}(\texttt{instructor}) \\ \texttt{Musics} &\leftarrow \sigma_{\texttt{dept\_name="Musics"}}(\texttt{instructor}) \\ &\texttt{Musics} \cup \texttt{Physics} \end{aligned}$$

2.3.11 Equivalent Queries

Since there is more than one way to write a query in relational algebra, queries that are not identical may be **equivalent**, they give the same result on any database.

Alternative Notation

On a related note, queries can be written with the alternative notation shown. For instance, `select(p, r)` instead of $\sigma_p(r)$