SQL

COMPUTER SCIENCE MENTORS 61A

December 2 – December 6, 2024

1 SQL

SQL (Structured Query Language) is a declarative programming language that allows us to store, access, and manipulate data stored in databases. Each database contains tables, which are rectangular collections of data with rows and columns. This section gives a brief overview of the small subset of SQL used by CS 61A; the full language has many more features.

1.1 Creating Tables

1.1.1 SELECT

SELECT statements are used to create tables. The following creates a table with a single row and two columns:

```
sqlite> SELECT "Adit" AS first, "Balasubramanian" AS last;
Adit|Balasubramanian
```

AS is an "aliasing" operation that names the columns of the table. Note that built-in keywords such as **AS** and **SELECT** are capitalized by convention in SQL. However, SQL is case insensitive, so we could just as easily write **as** and **select**. Also, each SQL query must end with a semicolon.

1.1.2 UNION

UNION joins together two tables with the same number of columns by "stacking them on top of each other". The column names of the first table are kept.

```
sqlite> SELECT "Adit" AS first, "Balasubramanian" AS last UNION
...> SELECT "Gabe", "Classon";
Adit|Balasubramanian
Gabe|Classon
```

1.1.3 CREATE TABLE

To create a named table (so that we can use it again), the **CREATE TABLE** command is used:

```
CREATE TABLE scms AS
    SELECT "Adit" AS first, "Balasubramanian" AS last UNION
    SELECT "Gabe", "Classon";
```

Created by Jim Fang, Bryan Adolfo Ventura Benitez. Worksheet feedback: links.jimfang.dev/feedback

The remaining examples will use the following `team` table:

```
CREATE TABLE team AS
    SELECT "Gabe" AS name, "cat" AS pet, 11 AS birth_month UNION
    SELECT "Adit",         "none",       10 UNION
    SELECT "Alyssa",       "dog",         4 UNION
    SELECT "Esther",       "dog",         6 UNION
    SELECT "Maya",         "dog",         3 UNION
    SELECT "Manas",        "none",       11;
```

1.2 Manipulating other tables

We can also write **SELECT** statements to create new tables from other tables. We write the columns we want after the **SELECT** command and use a **FROM** clause to designate the source table. For example, the following will create a new table containing only the `name` and `birth_month` columns of `team`:

```
sqlite> SELECT name, birth_month FROM team;
Adit|10
...
Maya|3
```

Note that the order in which rows are returned is undefined.

An asterisk `*` selects for all columns of the table:

```
sqlite> SELECT * FROM team;
Adit|none|10
...
Maya|dog|3
```

This is a convenient way to view all of the content of a table.

We may also manipulate the table columns and use **AS** to provide a (new) name to the columns of the resulting table. The following query creates a table with each teammate's name and the number of months between their birth month and June:

```
sqlite> SELECT name, ABS(birth_month - 6) AS june_dist FROM team;
Adit|4
...
Maya|3
```

1.2.1 WHERE

WHERE allows us to filter rows based on certain criteria. The **WHERE** clause contains a boolean expression; only rows where that expression evaluates to true will be kept.

```
sqlite> SELECT name FROM team WHERE pet = "dog";
Alyssa
Esther
Maya
```

Note that = in SQL is used for equality checking, not assignment.

1.2.2 ORDER BY

ORDER BY specifies a value by which to order the rows of the new table. **ORDER BY** ... may be followed by **ASC** or **DESC** to specify whether they should be ordered in ascending or descending order. **ASC** is default. For strings, ascending order is alphabetical order.

```
sqlite> SELECT name FROM team WHERE pet = "dog" ORDER BY name DESC;
Maya
Esther
Alyssa
```

1.3 Joins

Sometimes, you need to compare values across two tables—or across two rows of the same table. Our current tools do not allow for this because they can only consider rows one-by-one. A way of solving this problem is to create a table where the rows consist of every possible combination of rows from the two tables; this is called an **inner join**. Then, we can filter through the combined rows to reveal relationships between rows. It sounds bizarre, but it works.

An inner join is created by specifying multiple source tables in a **WHERE** clause. For example, `SELECT * FROM team AS a, team AS b;` will create a table with 36 rows and 6 columns. The table has 36 rows because each row represents one of 36 possible ways to select two rows from `team` (where order matters). The table has 6 columns because the joined tables have 3 columns each. We use **AS** to give the two source tables different names, since we are joining `team` to itself. The columns of the resulting table are named `a.name, a.pet, a.birth_month, b.name, b.pet, b.birth_month`.

For example, to determine all pairs of people with the same birth month, we can use an inner join:

```
sqlite> SELECT a.name, b.name FROM team AS a, team AS b WHERE a.name < b.name
   AND a.birth_month = b.birth_month;
Gabe|Manas
```

1.4 Aggregation

Aggregation uses information from multiple rows in our table to create a single row. Using an aggregation function such as **MAX**, **MIN**, **COUNT**, and **SUM** will automatically aggregate the table data into a single row. For example, the following will collapse the entire table into one row containing the name of the person with the latest birth month:

```
sqlite> SELECT name, MAX(birth_month) FROM team;
Manas|11
```

Note that there are multiple rows with the largest birth month. When this happens, SQL arbitrarily chooses one of the rows to use.

The **COUNT** aggregation function collapses the table into one row containing the number of rows in the table:

```
sqlite> SELECT COUNT(*) FROM team;
6
```

1.4.1 GROUP BY

GROUP BY groups together all rows with the same value for a particular column. Aggregation is performed on each group instead of on the entire table. There is then *exactly one row* in the resulting table for each group. As before, type of aggregation performed is determined by the choice of aggregation function. The following gives, for each type of pet, the information of the person with the earliest birth month who has that pet:

```
sqlite> SELECT name, pet, MIN(birth_month) FROM team GROUP BY pet;
Gabe|cat|11
Maya|dog|3
Adit|none|10
```

1.4.2 HAVING

Just as **WHERE** filters out rows, **HAVING** filters out groups. For example, the following selects for all types of pets owned by more than one teammate:

```
sqlite> SELECT pet FROM team GROUP BY pet HAVING COUNT(*) > 1;
dog
none
```

1.5 Syntax

The clauses of a **SELECT** statement are written in the following order:

```
SELECT ... FROM ... WHERE ... GROUP BY ... HAVING ... ORDER BY ...;
```

However, the actual processing steps differ slightly. Notably, all row-by-row filtering (via the **WHERE** clause) occurs *before* aggregation (via the **GROUP BY** clause), ensuring that only filtered rows are aggregated.

1. CSM 61A Content Team wants to put together an end of the year party for all its mentors (and some special guests, too!) themed around everyone's favorite things! Examine the table, `mentors`, depicted below, and answer the following questions.

Name	Food	Color	Editor	Language	Animal
Alyssa	Pork Bulgogi	Navy Blue	Vim	Java	Sea Otter
Vibha	Pasta	Teal	VSCode	Java	Naked Mole Rat
Adit	Protein Bar	Black	Vim	Python	Gorilla
Esther	Goldfish	Pastel Pink	VSCode	Python	Bunny
Aiko	Fries	Sky Blue	VSCode	Java	Cat
Aurelia	Dumpling	Pastel Yellow	VSCode	Python	Panda
Kaelyn	Popcorn	Blue	VSCode	Java	Panda

(a) Create a new table `mentors` that contains all the information above. (You only have to write out the first two rows.)

(b) Write a query that has the same data, but alphabetizes the rows by name. (Hint: Use **`order by`**.)

```
Adit|Protein Bar|Black|Vim|Python|Gorilla
Aiko|Fries|Sky Blue|VSCode|Java|Cat
Alyssa|Pork Bulgogi|Navy Blue|Vim|Java|Sea Otter
Aurelia|Dumpling|Pastel Yellow|VSCode|Python|Panda
Esther|Goldfish|Pastel Pink|VSCode|Python|Bunny
Kaelyn|Popcorn|Blue|VSCode|Java|Panda
Vibha|Pasta|Teal|VSCode|Java|Naked Mole Rat
```

(c) Write a query that lists the food and the color of every person whose favorite language is *not* Python.

```
Pork Bulgogi|Navy Blue
Pasta|Teal
Fries|Sky Blue
Popcorn|Blue
```

2. After more than 100 years of operation, the Ringling Bros. circus is closing. A victory for animal rights advocates, the circus' closure poses a challenge for the zoologists tasked with moving the circus' animals to more suitable habitats.

   The zoologists must first take the animals in a freight elevator with a weight limit of 2000. In order to speed up the process, the zoologists prefer to take groups of animals of the same species in the elevator, rather than one animal at a time.

   Assume the zoologists will only put all of the animals of a particular species in the elevator, or take animals of that particular species one at a time.

   You have access to the table `animals`, with columns containing the animals' names, heights, weights, and species.

Name	Height	Weight	Species
Wilbur	4.1	150	pig
Tigress	4.4	700	tiger
Phil	3.3	79	pig
Dug	3.5	40	dog
Buddy	4	51	dog
Marty	4.9	300	zebra
Richard Parker	5.2	918	tiger

   (a) Write a query that returns the collective weight and species of animals in a group where there is more than one animal of a particular species in a group, and the collective weight of the animals in the group is less than 2000.

   Your query should yield the following result.

```
91 dog
229 pig
1618 tiger
```

(b) To take the animals to their new habitats, the zoologists load the animals into trucks. The zoologists would like to reduce the number of trips taken by pairing up the animals for the trip (species does not matter). However, the trucks have two restrictions:

1. The maximum height of any animal is 5.0.
2. The total weight of both animals cannot exceed 300.

Your query should yield the following result. To match the table below, make sure that the left column animals' names come earlier in the alphabet than their partners.

```
Buddy    Dug
Buddy    Phil
Buddy    Wilbur
Dug      Phil
Dug      Wilbur
Phil     Wilbur
```

3. CS 61A wants to start a fish hatchery, and we need your help to analyze the data we've collected for the fish populations! Running a hatchery is expensive—we'd like to make some money on the side by selling some seafood (only older fish of course) to make delicious sushi.

The table `fish` contains a subset of the data that has been collected. The SQL column names are listed in brackets.

`fish`

Species [species]	Population [pop]	Breeding Rate [rate]	$/piece [price]	# of pieces per fish [pieces]
Salmon	500	3.3	4	30
Eel	100	1.3	4	15
Yellowtail	700	2.0	3	30
Tuna	600	1.1	3	20

`competitor`

Species [species]	$/piece [price]
Salmon	2
Eel	3.4
Yellowtail	3.2
Tuna	2.6

(a) Write a query to find the three most populated fish species.

(b) Write a query to find the total number of fish in the ocean. Additionally, include the number of species we summed. Your output should have the number of species and the total population.

(c) Profit is good, but more profit is better. Write a query to select the species that yields the greatest number of pieces for each price point. For example, if "Trout" and "Catfish" both cost 100 dollars/-piece, you should output the species that has the greater number of pieces. Your output should include the species, price, and number of pieces per fish.

(d) Write a query that returns, for each species, the difference between our hatchery's revenue versus the competitor's revenue for one whole fish.