3/4/2024

Sharing and Reuse of Data

Data sharing is the practice of making data used for research publicly available

- Promotes transparency
- Increases confidence in the scientific results
- Helps in replication and catching errors

A lot of scientific is still kept secret

- For privacy reasons
- Because researchers don't want to get "scooped" from someone else using their data

Some research journals now actually require that data used in their publications be made public (to some degree)

Open data is becoming a more desirable practice

Data.gov is an outgrowth of this Open Data Movement

Using sharing can help increase sample sizes, especially for data that can be difficult to collect

Data reused – using data for a purpose other than the one it was collected for

Many companies sell your data to someone else, the common example of data reuse

Downside of data sharing – the Trump administration tried to ban the use of data that was not publicly available

Data sharing is often done through data repositories:

- Have the data described (the meaning of variables, etc.), to make the data easy to reuse by someone else
- Metadata
- Data dictionaries – provide detailed definitions of the variables, how they were collected, methodologies, data types, etc.

SQL

Joins in SQL (specifically in SQLite)

Join types: inner joins (natural join), left (outer) join, cross join

If you don't, the default in SQLite is the inner join

Left joins keep all the data in the "left" table, then adds in data from the second table (right) table that has a match in the left table (discards unmatched things from the right table).

Right joins do the opposite, discards data in the left table that does not match the right table

Full outer joins: matches data in both tables, but discards

Cross join: use this very, very sparingly, especially if you have a lot data. Pairs every row of table 1 with every row of table 2.

{1,2,3,4}x{a,b,c,d}
{(1,a),(1,b),(1,c),(1,d),(2,a),(2,b),(2,c),(2,d), (3,a),(3,b),(3,c),(3,d),(4,a),(4,b),(4,c),(4,d)}

Examples:

```
SELECT * FROM tablename1 INNER JOIN tablename2 ON tablename1.fieldname = tablename2.fieldname;
```

```
SELECT * FROM tablename1 INNER JOIN tablename2 USING(fieldname);
```

```
SELECT * FROM tablename1 NATURAL JOIN tablename2;
```

Left join basically works the same as INNER JOIN

```
SELECT * FROM tablename1 LEFT JOIN tablename2 ON tablename.fieldname = tablename2.fieldname;
```

Or:

LEFT OUTER JOIN

Resources:

1. https://www.nature.com/articles/d41586-019-01506-x
2. https://eudatasharing.eu/what-data-sharing
3. https://mozillascience.github.io/working-open-workshop/data_reuse/
4. https://escholarship.org/uc/item/0jj17309
5. https://www.w3schools.com/sql/
6. https://www.codecademy.com/learn/learn-sql
7. https://www.sqlitetutorial.net/
8. https://www.datacamp.com/tutorial/beginners-guide-to-sqlite
9. https://intellipaat.com/blog/tutorial/sql-tutorial/sql-commands-cheat-sheet/
10. https://www.sqltutorial.org/
11. https://www.guru99.com/sql.html
12. https://www.sqlcourse.com/beginner-course/what-is-sql/

Extended commentary:

Sharing and reuse of data refer to the practice of making data available to others for their use and facilitating the use of data that others have made available. This is an important aspect of data management as it can help to promote collaboration, improve research quality, and reduce the costs and time required to collect and analyze data.

Sharing and reuse of data can be facilitated by making data available in repositories or archives, where they can be accessed by others. This can include data produced by research projects, surveys, or administrative records. Data can be shared and reused by researchers, policymakers, businesses, and the general public, depending on the nature of the data and the permissions granted by the data owners.

There are several benefits to sharing and reusing data, including:

Efficiency: Sharing data can save time and resources by eliminating the need to collect new data for every project.

Reproducibility: Sharing data allows others to replicate research findings, increasing confidence in the results.

Collaboration: Sharing data can promote collaboration between researchers, leading to new insights and discoveries.

Transparency: Sharing data promotes transparency and accountability by allowing others to scrutinize the data and the methods used to collect and analyze it.

However, there are also some challenges associated with sharing and reusing data, including concerns about data privacy and security, intellectual property rights, and the potential for misuse of data. It is important to develop policies and procedures to ensure that data are shared and reused in a responsible and ethical manner.

Overall, sharing and reusing data can be an important aspect of data management, providing opportunities for collaboration, efficiency, and innovation, while promoting transparency and accountability.

Data privacy can be maintained when sharing data through various measures, including:

Anonymization: Data can be anonymized to remove any identifiable information about individuals, such as names, addresses, and social security numbers. This ensures that the privacy of individuals is protected when data is shared.

Aggregation: Data can be aggregated so that it is presented in a summarized form. This can help to protect the privacy of individuals while still allowing data to be shared.

Encryption: Data can be encrypted so that it is only accessible to authorized users. This can help to protect data privacy by preventing unauthorized access to sensitive data.

Access controls: Access controls can be put in place to ensure that only authorized individuals are able to access data. This can help to prevent unauthorized access and ensure that data privacy is maintained.

Data sharing agreements: Data sharing agreements can be put in place to ensure that data is only used for specific purposes and is not shared with unauthorized individuals. These agreements can also include provisions for data privacy and security.

Overall, maintaining data privacy when sharing data requires a combination of technical, organizational, and legal measures to ensure that sensitive data is protected and only accessed by authorized individuals.

In SQL, there are different types of joins that can be used to combine data from multiple tables. The most commonly used types of joins are:

INNER JOIN: The INNER JOIN returns only the rows that have matching values in both tables being joined.

Example:

```
SELECT *
FROM table1
INNER JOIN table2
```

```
ON table1.id = table2.id;
```

LEFT JOIN: The LEFT JOIN returns all the rows from the left table and matching rows from the right table. If there are no matching rows in the right table, the result will contain NULL values.

Example:

```
SELECT *
FROM table1
LEFT JOIN table2
ON table1.id = table2.id;
```

RIGHT JOIN: The RIGHT JOIN returns all the rows from the right table and matching rows from the left table. If there are no matching rows in the left table, the result will contain NULL values.

Example:

```
SELECT *
FROM table1
RIGHT JOIN table2
ON table1.id = table2.id;
```

FULL OUTER JOIN: The FULL OUTER JOIN returns all the rows from both tables and matches the rows that have matching values in either table. If there are no matching rows in either table, the result will contain NULL values.

Example:

```
SELECT *
FROM table1
FULL OUTER JOIN table2
ON table1.id = table2.id;
```

CROSS JOIN: The CROSS JOIN returns the Cartesian product of both tables, i.e., all possible combinations of rows between both tables.

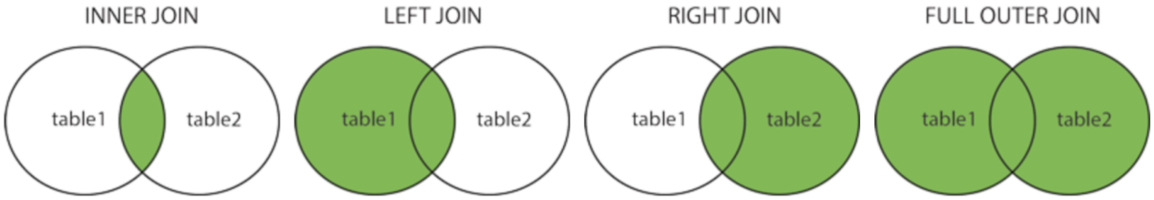
Example:

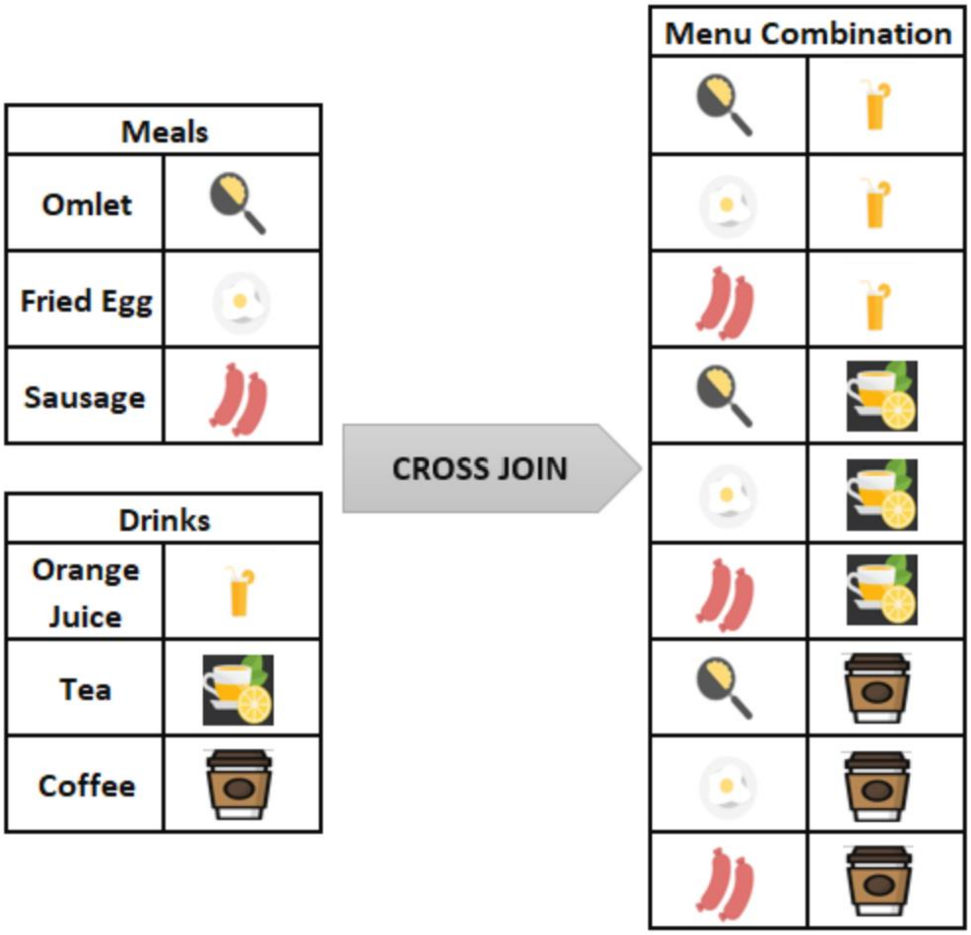
```
SELECT *
FROM table1
CROSS JOIN table2;
```

Different Types of SQL JOINs

Here are the different types of the JOINs in SQL:

- `(INNER) JOIN`: Returns records that have matching values in both tables
- `LEFT (OUTER) JOIN`: Returns all records from the left table, and the matched records from the right table
- `RIGHT (OUTER) JOIN`: Returns all records from the right table, and the matched records from the left table
- `FULL (OUTER) JOIN`: Returns all records when there is a match in either left or right table


INNER JOIN
table1
table2
LEFT JOIN
table1
table2
RIGHT JOIN
table1
table2
FULL OUTER JOIN
table1
table2


Meals
Omlet
Fried Egg
Sausage
Drinks
Orange Juice
Tea
Coffee
CROSS JOIN
Menu Combination