Topics

This lecture is for stuff I forgot or didn’t have time to cover so far…

- Miscellaneous SELECT
  - DISTINCT
  - JOIN clause and outer joins
  - SET operations
  - Stuff we don’t cover
- Bulk loading and COPY
- Workflow

DISTINCT

- SELECT queries may return duplicate rows
  - E.g.
    ```sql
    SELECT instructor FROM mines_courses;
    ```
  - DISTINCT keyword lets us remove duplicates:
    ```sql
    SELECT DISTINCT instructor FROM mines_courses;
    ```
  - Also, can use with aggregates (not all DBMSes):
    ```sql
    SELECT COUNT(DISTINCT instructor) FROM mines_courses;
    SELECT AVG(DISTINCT enrollment) FROM mines_courses;
    ```
    (the last is a silly application, just for illustration)

JOINS

We’ve so far been using “WHERE clause” joins:
```sql
SELECT a.x, b.y from a, b WHERE a.z = b.z;
```

We can also use a JOIN expression explicitly:
```sql
SELECT a.x, b.y FROM a INNER JOIN b ON (a.z = b.z);
```
More JOIN

Can use with multiple conditions:
SELECT a.x, b.y
FROM a INNER JOIN b ON (a.z = b.z AND a.zz = b.zz);

Can use with multiple tables:
SELECT a.x, b.y, c.z
FROM a INNER JOIN b ON (a.foo = b.foo)
INNER JOIN c ON (b.bar = c.bar);

INNER JOIN Notes

- The INNER keyword is optional (INNER is the default)
- Performance is same using WHERE clause or INNER JOIN
- Choice between WHERE clause or JOIN clause is completely up to you:
  - Personal preference
  - Readability (JOIN clause can get complicated with many tables)

OUTER JOIN

- Inner joins only return rows that match join condition:
  - What if we want all rows from one or both tables?
  - Outer join lets us get everything, pairing up rows where possible.
  - Example (in class):
    SELECT mc.course_id, mc.section, mc.title, mc.instructor, mcf.email, mcf.office
    FROM mines_courses AS mc
    LEFT OUTER JOIN mines_cs_faculty AS mcf
    ON (mc.instructor = mcf.name);

LEFT and RIGHT

SELECT mc.course_id, mc.section, mc.title, mc.instructor, mcf.email, mcf.office
FROM mines_courses AS mc
LEFT OUTER JOIN mines_cs_faculty AS mcf
ON (mc.instructor = mcf.name);

The above gives all rows from mines_courses (the left-hand table in the join).

RIGHT does the opposite.

Note that OUTER is optional – LEFT and RIGHT imply OUTER.

FULL OUTER JOIN

As the name implies, gives all rows from both tables in join, matching rows where possible.
Same example as above, replacing LEFT with FULL (OUTER is optional again).
To see effects, first have to insert a bogus faculty member into mines_cs_faculty...

Applications of Outer Joins

Great for finding missing data, e.g., data where a foreign key is null, by using IS NULL in WHERE:

SELECT mc.course_id, mc.section, mc.title, mc.instructor, mcf.email, mcf.office
FROM mines_courses AS mc
LEFT OUTER JOIN mines_cs_faculty AS mcf
ON (mc.instructor = mcf.name)
WHERE mcf.name IS NULL;
Oracle Outer Joins

Oracle defines a special operator that allows left and right outer joins to be created using WHERE clause (very handy!):

```
SELECT a.x, b.y
FROM a, b
WHERE a.z = b.z(+); -- right outer join!
```

Set Operations

Set union : UNION
Set intersection : INTERSECT
Set difference: EXCEPT

[ALL]

- By default, set operators imply DISTINCT
- This is because sets, mathematically, contain only distinct units
- To avoid this behavior, add ALL keyword
  ```
  SELECT ...
  UNION ALL
  SELECT ...
  ```

Other Stuff
Stuff We Don’t Cover

- Recursive queries
- WITH clause
- LIMIT clause (limits # of rows)
- Various “syntactic sugar”

BULK LOADING

Bulk Loading

- Term for loading lots of data into database efficiently
- Compare to doing millions of INSERT statements...
- Typical workflow (sometimes known as ETL: Extract, Transform, Load):
  - Acquire data in text, .csv, JSON, or other format
  - Preprocess if needed/desired using non-SQL tools
  - Bulk load data as is (dirty, incomplete, incorrectly formatted) into staging table(s)
  - Postprocess if needed (clean up, format) using SQL
  - Use SQL to load into final table(s)

Bulk Loading in PostgreSQL: COPY

- COPY command in PostgreSQL SQL
- Fast bulk loading from various formats
- Lets DB administrators only load data from server filesystem
- COPY in psql
- Fast bulk loading across network
- Lets non-administrators bulk load from local filesystem
- Other bulk loaders/ETL tools
- Commercial and free tools available
- Google “ETL postgresql”

Dangers of Command Line SQL

- Databases are dynamic storage
  - You can modify data quickly with SQL
  - Everyone sees modified data immediately
- SQL is immediate, irrevocable
  - What if you make a mistake?
    - E.g., production database migration (new schema)
  - Don’t do command line SQL except for exploratory programming on dev DB!

Lessons learned the hard way
Scripts Are Your Friends

- Keep dev, test, production databases (at minimum)
- During development, write & test scripts on dev
- When time for a software release/database migration:
  - Clone a new db from test
  - Apply scripts
  - Fix scripts
  - Repeat as necessary
- Scripts go in your version control system (git)!!
  - In theory, should be able to reproduce prod database schema (not data, necessarily) from scratch by running all scripts from beginning of project.

DBAs/SysAdmins Are Also Friends

- Your production database should be backed up nightly
  - Not for crashes, necessarily, as SQL databases can recover from those on their own
  - Instead, need it to recover from developer mistakes 😊
- However, make sure DBAs/sysadmins test backups regularly (DB backup is tricky!)

Up Next

- Next lecture:
  - Quiz review.
  - Monday, Sept. 24
  - Quiz on SQL (all material so far except history)
  - Wednesday, Sept. 26
  - Start database design and ERD