MY472 – Week 8: APIs

Pablo Barberá & Akitaka Matsuo

MY 472: Data for Data Scientists

November 20, 2018

Course website: lse-my472.github.io

Course outline

1. Introduction to data
2. The shape of data
3. Cloud computing
4. Basics of HTML and CSS
5. Using data from the internet
6. (Reading week)
7. Working with APIs
8. Creating and managing databases
9. Interacting with online databases
10. Exploratory data analysis
11. Parallel computing

Seminar schedule

7. APIs
   - 3rd marked assignment (in groups)
   - Deadline: November 27
8. SQL
9. Online databases
   - 4th marked assignment (in groups)
   - Deadline: December 7th
10. Exploratory data analysis
11. Course wrap-up
    - 5th marked assignment (individual)
    - Deadline: December 21st

Take-home exam due January 18

Plan for today

- SQL
  - Relational databases
  - SQL language
  - Components of an SQL query
  - SQL in the cloud
  - Examples: querying public Facebook data

Introduction to SQL

Databases

- Database systems: computerized mechanisms to store and retrieve data.
- Relational databases: data is represented as tables linked based on common keys (to avoid redundancy).

Customer

cust_id	fname	lname
1	George	Blake
2	Sue	Smith

Account

account_id	product_cd	cust_id	balance
103	CHK	1	$75.00
104	SAV	1	$250.00
105	CHK	2	$783.64
106	MM	2	$500.00
107	LOC	2	0

Product

product_cd	name
CHK	Checking
SAV	Savings
MM	Money market
LOC	Line of credit

Transaction

txn_id	txn_type_cd	account_id	amount	date
978	DBT	103	$100.00	2004-01-22
979	CDT	103	$25.00	2004-02-05
980	DBT	104	$250.00	2004-03-09
981	DBT	105	$1000.00	2004-03-25
982	CDT	105	$138.50	2004-04-02
983	CDT	105	$77.86	2004-04-04
984	DBT	106	$500.00	2004-03-27

SQL

- SQL (pronounced S-Q-L or SEQUEL) is a language designed to query relational databases
- Used by most financial and commercial companies
- The result of an SQL query is always a table
- It's a nonprocedural language: define inputs and outputs; how the statement is executed is left to the *optimizer*
- How long SQL queries depends on optimization that is opaque to user (which is great!)
- SQL is a language that works with many commercial products:
  - Oracle Database, SQL Server (MS), MySQL, PostgreSQL, SQLite (all three open-source), Google BigQuery, Amazon Redshift...
  - Performance will vary, but generally faster than standard data frame manipulation in R (and much more scalable)

Components of a SQL query

- SELECT columns
- FROM a table in a database
- WHERE rows meet a condition
- GROUP BY values of a column
- ORDER BY values of a column when displaying results
- LIMIT to only X number of rows in resulting table

- Always required: SELECT and FROM. Rest are optional.
- SELECT can be combined with operators such as SUM, COUNT, AVG...
- To merge multiple tables, you can use JOIN

SQL at scale

Google BigQuery

- One of many commercial SQL databases available (Amazon RedShift, Microsoft Azure, Oracle Live SQL...)
- Used by many financial and commercial companies
- **Advantages:**
  - Integration with other Google data storage solutions (Google Drive, Google Cloud Storage)
  - Scalable: same SQL syntax for datasets of *any* size
  - Easy to collaborate and export results
  - Affordable pricing and cost control
  - API access allows integration with R or python
  - Excellent documentation

Example: using SQL to query public Facebook data

see `01-sql-intro.Rmd`

see `02-sql-advanced.Rmd`