MY560 Workshop. Querying large-scale online datasets: SQL and Google BigQuery

Pablo Barberá
London School of Economics
www.pablobarbera.com

Workshop website:

pablobarbera.com/SQL-workshop

Hello!

About me

- Assistant Professor of Computational Social Science in the Methodology Department at LSE
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
  - Social media and politics, comparative electoral behavior, corruption and accountability
  - Social network analysis, Bayesian statistics, text as data methods
  - Author of R packages to analyze data from social media
- Contact:
  - `P.Barbera@lse.ac.uk`
  - `www.pablobarbera.com`
  - Office hours: Mondays 15-16:00 and Wednesdays 11-12:00 in COL.7.10

Today's workshop

Session 1, 10–12:00

- Introduction to SQL databases
- Guided coding session: basics of SQL queries
- Challenges 1 & 2: interacting with an SQL database with Facebook data

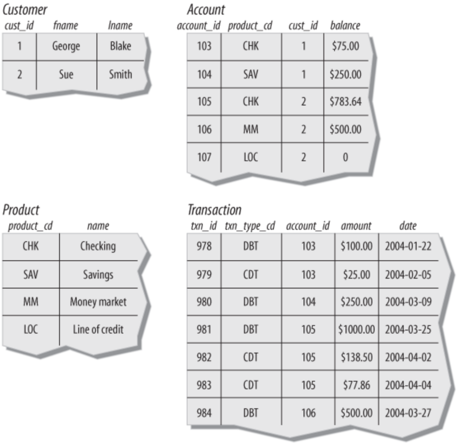
Session 2, 14–16:00

- Guided coding session: introduction to Google BigQuery
- Coding challenge 3: querying a large-scale Twitter database on Google BigQuery
- Guided coding session: advanced Google BigQuery examples
- Coding challenge 4: analyzing a billion-row database with Google BigQuery

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

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