University of California, Berkeley
Master of Information and Data Science (MIDS)
W205 – Fundamentals of Data Engineering

Week 2 – SQL Refresher

Agenda for Today's Class

- Attendance and Participation
- Announcements
- Schedule and Due Dates
- Work / Life / School Balance
- Asynch High Level Review in a Nutshell
- Project 1
- Breakouts
- Summary

Attendance and Participation

Please record your attendance and participation for today's class:

GitHub => ucb_mids_w205_repo => README.md => Attendance and Participation

Announcements

- Upcoming holidays and/or breaks
- Makeup classes for holidays
- Upcoming events
- Student evaluations
- Etc.

Schedule and Due Dates

Take a quick look at the next couple of weeks' due dates:

GitHub => ucb_mids_w205_repo => README.md => Schedule and Due Dates

Work / Life / School Balance
Open Discussion

Student feedback

- About 5 minutes
- How are things going related to work / life / school balance?
- How is w205 going? Difficulty? Time?
- Impact of any natural and/or man-made disasters
- Etc.

Asynch High Level Review in a Nutshell

Each week we will spend about 15 minutes reviewing the most important high level concepts from the asynch

Relational Database

- Collection of tables, tables made up of rows and columns
- Parent / child relationships between tables
  - Don’t confuse relation with relationship!
  - Relation = table, result of SQL query or subquery
- SQL – Structured Query Language
  - DDL – Data Definition Language – create, drop, etc.
  - DML – Data Manipulation Language – select, insert, update, delete

Functional Programming

- Procedural
  - if, loops, etc.
  - Python, C/C++, Java, etc.
- Nonprocedural
  - aka Declarative, what to do, not how to do
  - SQL
- Functional Programming
  - Procedural code wrapped around nonprocedural (declarative)
  - Best of both worlds
  - Python calling SQL

Relationships and Keys

- Primary key
  - Uniquely identifies a row in a table
- Relationship
  - Primary key in parent table matches to foreign key in child table
- Data model
  - ERD – entity relationship diagram
  - Columns, primary keys, foreign keys, relationships, etc.

Transactional versus Analytical

- Transactional Database
  - Executes the business
  - 3NF – 3rd Normal Form – no duplication of data
- Analytical Database
  - Evaluates the execution of the business
  - Denormalized – duplication of data to make analytical queries easier

SQL

- SELECT - column list, derived columns, column aliases
- FROM - table or tables
- WHERE - filters pre-aggregation
- GROUP BY – columns to aggregate
- HAVING – filters post-aggregation
- ORDER BY – sorts results

Set Operations and Joins

- Set Operations – combine rows from separate queries
- Join Operations – combine columns from multiple tables
  - Inner join – only matching rows
  - Left outer join – also includes rows in left table without a match
  - Right outer join – also includes rows in right table without a match
  - Dangerous joins
    - No defined relationship
    - Common when joining secondary dataset to primary dataset

Subqueries and Views

- Type 1 Subquery
  - No relationship between inner and outer query
  - Scales up
- Type 2 Subquery
  - Relationship between inner and outer query
  - Does not scale up
- View
  - Acts like a permanent type 1 subquery

Data Visualization

- Optional module to demonstrate some basic data visualizations from data pulled using SQL
- Pie charts, grids, scatter plots, line plots, bar charts, histograms, box plots, violin plots, etc.

Geographic Data Visualization

- Optional module to demonstrate some basic geographic data visualizations from data pulled using SQL
- Maps, markers, heatmaps, choropleths, driving directions, traffic layers, transit layers, etc.
- Geodesic calculations
  - Latitude and longitude points
  - Distance between 2 points
  - Direction between 2 points
  - Given a point, a direction, a distance, find a new point
  - Pulling points using SQL within a given distance from a point

Project 1

GitHub => ucb_mids_w205_repo => projects => project_1

- Videos going over project 1 are provided, so we won't spend class time going over it today
- Breakouts today and 1 breakout next week will be related to project 1

Breakouts

GitHub => ucb_mids_w205_repo => breakouts

(time permitting, we may not get to all of them)

Summary

Instructor will give a brief (about 2 minute) summary of today’s class.