CS 327E Class 4

October 1, 2018

Project Roadmap

*Dataset3 is required only if Dataset2 contains a single table.

Milestone 1: Setup on GCP and BQ.

Milestone 2: Data loads of Dataset1, Dataset2 (and Dataset3*).

Milestone 3: Joins on Dataset1.

Milestone 4: Aggregations on Dataset1.

Milestone 5: Subqueries on Dataset1.

Milestone 6: Data modeling on Dataset1.

Milestone 7: Setup on Beam and Dataflow.

Milestone 8: Reformat Dataset1 and/or Dataset2 (and Dataset3*).

Milestone 9: Integrate Dataset1 and Dataset2 (and Dataset3*).

A Few Examples

	Dataset1	Dataset2
Transportation	Airline on-time performance (source: BTS)	Storm events (source: NOAA)
Housing	Short-term rentals in various cities (source: Airbnb)	Long-term rentals nationwide (source: Zillow)
Political Campaigns	Federal campaign finance (source: Federal Election Commission)	State campaign finance (source: TX Ethics Commission)
Movies	Hollywood movies, directors, actors (source: IMDB)	Bollywood movies, actors and songs (source: Cinemalytics)
Music	Artists and songs (source: MusicBrainz)	Artists, labels, recordings on vinyl and other formats (source: Discog)

1) Which is not an aggregate function?

A. SUM()
B. COUNT(*)
C. AVG()
D. MIN()
E. None of the above

2) Consider the `World_Cup_Players_2018` table shown below. What is the output from Q1 when run on this table?

```
Q1: SELECT COUNT(*) FROM World_Cup_Players_2018;
```

World_Cup_Players_2018

player_id	player_name	country	position	goals
1	Cristiano Ronaldo	Portugal	Forward	4
2	Pepe	Portugal	Defender	1
3	Neymar	Brazil	Forward	2
4	Messi	Argentina	Forward	1
5	Kylian Mbappe	France	Forward	4
6	Diego Costa	Spain	Striker	3
7	Eden Hazard	Germany	Midfielder	3

A. 7
B. 4
C. 3
D. 0
E. NULL

3) Consider the `World_Cup_Players_2018` table shown below. What is the output from Q2 when run on this table?

```
Q2: SELECT MIN(goals) FROM World_Cup_Players_2018;
```

World_Cup_Players_2018

player_id	player_name	country	position	goals
1	Cristiano Ronaldo	Portugal	Forward	4
2	Pepe	Portugal	Defender	1
3	Neymar	Brazil	Forward	2
4	Messi	Argentina	Forward	1
5	Kylian Mbappe	France	Forward	4
6	Diego Costa	Spain	Striker	3
7	Eden Hazard	Germany	Midfielder	3

A. 0
B. 1
C. 2
D. 3
E. NULL

4) Consider the `World_Cup_Players_2018` table shown below. What is the output from Q3 when run on this table?

```
Q3: SELECT MAX(goals) FROM World_Cup_Players_2018;
```

World_Cup_Players_2018

player_id	player_name	country	position	goals
1	Cristiano Ronaldo	Portugal	Forward	4
2	Pepe	Portugal	Defender	1
3	Neymar	Brazil	Forward	2
4	Messi	Argentina	Forward	1
5	Kylian Mbappe	France	Forward	4
6	Diego Costa	Spain	Striker	3
7	Eden Hazard	Germany	Midfielder	3

A. 3
B. 4
C. 5
D. 6
E. 7

5) Consider the `World_Cup_Players_2018` table shown below. What is the output from Q4 when run on this table?

```
Q4: SELECT SUM(goals) FROM World_Cup_Players_2018;
```

World_Cup_Players_2018

player_id	player_name	country	position	goals
1	Cristiano Ronaldo	Portugal	Forward	4
2	Pepe	Portugal	Defender	1
3	Neymar	Brazil	Forward	2
4	Messi	Argentina	Forward	1
5	Kylian Mbappe	France	Forward	4
6	Diego Costa	Spain	Striker	3
7	Eden Hazard	Germany	Midfielder	3

A. < 7
B. 7 - 17
C. 18
D. > 18
E. NULL

Syntax of Global Aggregate Queries

```
SELECT <list of aggregate functions>
FROM <single table>
JOIN <single table> ON <common fields>
WHERE <boolean conditions>
```

Syntax of Aggregate Queries with Groups

```
SELECT <unaggregated fields>, <aggregate functions>
FROM <single table>
JOIN <single table> ON <common fields>
WHERE <boolean conditions>
GROUP BY <unaggregated fields>
ORDER BY <list of fields to sort on>
```

Syntax of Aggregate Queries with Groups

```
SELECT <unaggregated fields>, <aggregate functions>
FROM <single table>
JOIN <single table> ON <common fields>
WHERE <boolean conditions>
GROUP BY <unaggregated fields>
HAVING <boolean conditions>
ORDER BY <list of fields to sort on>
```

Flavors of `COUNT()`

1) `COUNT(*)`
2) `COUNT( <some_field> )`
3) `COUNT(DISTINCT <some_field> )`

Flavors of `COUNT()`

1) `SELECT COUNT(*)`
   `FROM employee`
   Count = 6

2) `SELECT COUNT(emp_dep)`
   `FROM employee`
   Count = 5

3) `SELECT COUNT(DISTINCT emp_dep)`
   `FROM employee`
   Count = 3

Table Details: employee

Schema | Details | Preview

Row	empid	emp_name	emp_dep
1	3	Sarah	null
2	2	Mike	1
3	6	Sunil	1
4	23	Dave	2
5	5	Jim	4
6	37	Morgan	4

First Question

How many students are taking each class?

Student(sid, fname, lname, dob)

Class(cno, cname, credits)

Teacher(tid, fname, lname, dept)

Takes(sid, cno, grade)

Teaches(tid, cno)

Second Question

For each class with at least 2 students, how many students are taking such a class?

Student(sid, fname, lname, dob)

Class(cno, cname, credits)

Teacher(tid, fname, lname, dept)

Takes(sid, cno, grade)

Teaches(tid, cno)

iClicker Question

For each class with at least 2 students, how many students are taking such a class?

Student(sid, fname, lname, dob)

Class(cno, cname, credits)

Teacher(tid, fname, lname, dept)

Takes(sid, cno, grade)

Teaches(tid, cno)

Does this query require a `HAVING` clause?

A. Yes
B. No

Third Question

For each student who is 19-years old or above and is earning at least 3 class credits, how many total class credits are such students earning?

Student(sid, fname, lname, dob)

Class(cno, cname, credits)

Teacher(tid, fname, lname, dept)

Takes(sid, cno, grade)

Teaches(tid, cno)

iClicker Question

For each student who is 19-years old or above and is earning at least 3 class credits, how many total class credits are such students earning?

Student(<u>sid</u>, fname, lname, dob)
Class(<u>cno</u>, cname, credits)
Teacher(<u>tid</u>, fname, lname, dept)
Takes(<u>sid</u>, <u>cno</u>, grade)
Teaches(<u>tid</u>, <u>cno</u>)

Does this query require a `WHERE` clause?

A. Yes B. No

Fourth Question

Who takes exactly 3 classes?

Show the answer as a sorted list of sids.

Student(sid, fname, lname, dob)

Class(cno, cname, credits)

Teacher(tid, fname, lname, dept)

Takes(sid, cno, grade)

Teaches(tid, cno)

iClicker Question

Who takes exactly 3 classes?

Show the answer as a sorted list of sids.

Student(<u>sid</u>, fname, lname, dob)
Class(<u>cno</u>, cname, credits)
Teacher(<u>tid</u>, fname, lname, dept)
Takes(<u>sid</u>, <u>cno</u>, grade)
Teaches(<u>tid</u>, <u>cno</u>)

Does this query contain an aggregate function in the `SELECT` clause?
A. Yes B. No

BigQuery Demo

Milestone 4

http://www.cs.utexas.edu/~scohen/milestones/Milestone4.pdf