Project Roadmap

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

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

<table>
<thead>
<tr>
<th>player_id</th>
<th>player_name</th>
<th>country</th>
<th>position</th>
<th>goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cristiano Ronaldo</td>
<td>Portugal</td>
<td>Forward</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Pepe</td>
<td>Portugal</td>
<td>Defender</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Neymar</td>
<td>Brazil</td>
<td>Forward</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Messi</td>
<td>Argentina</td>
<td>Forward</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Kylian Mbappe</td>
<td>France</td>
<td>Forward</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Diego Costa</td>
<td>Spain</td>
<td>Striker</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>Eden Hazard</td>
<td>Germany</td>
<td>Midfielder</td>
<td>3</td>
</tr>
</tbody>
</table>

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;

<table>
<thead>
<tr>
<th>player_id</th>
<th>player_name</th>
<th>country</th>
<th>position</th>
<th>goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cristiano Ronaldo</td>
<td>Portugal</td>
<td>Forward</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Pepe</td>
<td>Portugal</td>
<td>Defender</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Neymar</td>
<td>Brazil</td>
<td>Forward</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Messi</td>
<td>Argentina</td>
<td>Forward</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Kylian Mbappe</td>
<td>France</td>
<td>Forward</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Diego Costa</td>
<td>Spain</td>
<td>Striker</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>Eden Hazard</td>
<td>Germany</td>
<td>Midfielder</td>
<td>3</td>
</tr>
</tbody>
</table>

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;**

<table>
<thead>
<tr>
<th>player_id</th>
<th>player_name</th>
<th>country</th>
<th>position</th>
<th>goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cristiano Ronaldo</td>
<td>Portugal</td>
<td>Forward</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Pepe</td>
<td>Portugal</td>
<td>Defender</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Neymar</td>
<td>Brazil</td>
<td>Forward</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Messi</td>
<td>Argentina</td>
<td>Forward</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Kylian Mbappe</td>
<td>France</td>
<td>Forward</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Diego Costa</td>
<td>Spain</td>
<td>Striker</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>Eden Hazard</td>
<td>Germany</td>
<td>Midfielder</td>
<td>3</td>
</tr>
</tbody>
</table>

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;

<table>
<thead>
<tr>
<th>player_id</th>
<th>player_name</th>
<th>country</th>
<th>position</th>
<th>goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cristiano Ronaldo</td>
<td>Portugal</td>
<td>Forward</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Pepe</td>
<td>Portugal</td>
<td>Defender</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Neymar</td>
<td>Brazil</td>
<td>Forward</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Messi</td>
<td>Argentina</td>
<td>Forward</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Kylian Mbappe</td>
<td>France</td>
<td>Forward</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Diego Costa</td>
<td>Spain</td>
<td>Striker</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>Eden Hazard</td>
<td>Germany</td>
<td>Midfielder</td>
<td>3</td>
</tr>
</tbody>
</table>

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
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?

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)
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?

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)
Who takes exactly 3 classes?

Show the answer as a sorted list of sids.

Does this query contain an aggregate function in the SELECT clause?
A. Yes  B. No
BigQuery Demo
Milestone 4