# Assignment 1 : BigQuery Basics & SQL

Question 1 (Aliases)  
Paste the following query into the BigQuery console.  
SELECT  
port\_name name,  
country,  
port\_latitude lat,  
port\_longitude long  
FROM  
`bigquery-public-data.geo\_international\_ports.world\_port\_index`  
where  
lat < 0  
Why does the query validator give an error? Fix the query, while maintaining the column names  
‘lat’ and ‘long’ in the output.

Provide your answer and your query.

Your Answer:

SELECT

port\_name name,

country,

port\_latitude lat,

port\_longitude long

FROM

`bigquery-public-data.geo\_international\_ports.world\_port\_index`

where

port\_latitude < 0;

We cannot reference alias in where clause. Where clause operates on the columns in the from clause clause. Hence its not possible to reference aliases from the SELECT list in the where clause. We have to be repeating the transformation and in this case, it will be port\_latitude < 0 instead of lat < 0.

Question 2 (Table aliases)  
In BigQuery, we reference tables by specifying three objects in a hierarchy - project, dataset and  
table name. For example,  
`bigquery-public-data.github\_repos.sample\_files`  
Notice that we have to escape this with backticks (`) because the project name contains hyphen  
characters. In general, we can reference the table in a join condition by giving the table name  
without the project and dataset reference. Paste the following query into the BigQuery console.

SELECT  
size,  
content,  
binary,  
repo\_name,  
ref  
FROM  
`bigquery-public-data.github\_repos.sample\_files`  
JOIN  
`bigquery-public-data.github\_repos.sample\_contents`  
ON  
sample\_files.id=sample\_contents.id  
Why does the validator give an error? How can you fix the query while still referencing the  
tables by their aliases rather than full table names? (Hint: this is straight from the reading.)

Provide your answer and your query.

Your Answer:

SELECT

size,

content,

binary,

repo\_name,

ref

FROM

`bigquery-public-data.github\_repos.sample\_files` as sample\_files

JOIN

`bigquery-public-data.github\_repos.sample\_contents` as sample\_contents

ON

sample\_files.id=sample\_contents.id

The validator gives error because the query is missing the alias name . We can fix the error by referencing with alias name and then we can use this for joining if we want to retain the alias name and avoid using the entire table name.

Question 3 (Aliases)  
Write a query to select author name, committer name and subject from  
`bigquery-public-data.github\_repos.sample\_commits`. Use column names ‘author’, ‘committer’  
and ‘subject’ in your output.Provide your query.

Your Answer:

SELECT

author,

committer,

subject

FROM

`bigquery-public-data.github\_repos.sample\_commits`

needed to select author.name as author, committer.name as committer

Question 4 (WHERE clauses)  
Select id and sample\_repo\_name from bigquery-public-data:github\_repos.sample\_contents.  
Select only those rows where size is greater than 2000 or there are at least 5 copies. (This is an  
inclusive OR).

(Select \*, except, replace)  
The general idea: minimize the number of columns that you’re selecting to reduce the amount of  
data scanned. This improves performance and reduces costs. Even though not every database  
charges by bytes scanned, the principle applies to any columnar database - minimizing the  
number of columns you scan will reduce resource consumption and improve performance.

Provide your query.

Your Answer:

SELECT

id,

sample\_repo\_name

FROM

`bigquery-public-data.github\_repos.sample\_contents`

where size > 2000 or copies >= 5

Question 5  
Determine the size of the full GitHub contents table,  
bigquery-public-data:github\_repos.contents. Estimate the cost to query the full table without  
running a query. You can see the full table size by clicking on “details” for the table. (Note that  
the query cost in this case is several dollars.) Provide your answers.

Your Answer:

The table size is Table size 2.38 TB.

Cost to run the query -

Active Storage 0 GiB

Long-term Storage 0 GiB

Queries 2.38 TiB

**USD 6.90**

**Total Estimated Cost: USD 6.90 per 1 month**

you can't assume a free TB, please see the announcements where I discussed this

Question 6  
What is the total size of bigquery-public-data:github\_repos.sample\_contents? Estimate the cost  
to query the full table.Provide your answers

Your Answer:

The table size 24.04 GB

The cost to query full table is

Active Storage 0 GiB

Long-term Storage 24.04 GiB

Queries 24.04 TiB

**USD 115.42**

**Total Estimated Cost: USD 115.42 per 1 month**

not sure where you got 24.04 TiB, the query is only gigabytes

Question 7  
Which column of bigquery-public-data:github\_repos.contents is largest? What is its size and  
what is the cost to query it? Hint: use the query validator to see the estimated data scanned for  
various queries.Provide your answers.

Your Answer:

SELECT

  content

FROM

`bigquery-public-data.github\_repos.contents`

limit 1000;

Size - 2.36 TB

Cost to query -

Active Storage 0 GiB

Long-term Storage 0 GiB

Queries 2.36 TiB

**USD 6.80**

**Total Estimated Cost: USD 6.80 per 1 month**

same logic as question 5

Question 8  
Use the EXCEPT keyword to write a query that returns all columns except the largest one. How  
much data would your query scan? Estimate the query cost.Provide your query and answers.

Your Answer:

SELECT

  \* EXCEPT(content)

FROM

`bigquery-public-data.github\_repos.contents`

Size - 15.11 GB

Cost to query -

Active Storage 0 GiB

Long-term Storage 0 GiB

Queries 0.015 TiB

**USD 0.00**

**Total Estimated Cost: USD 0.00 per 1 month**

its not actually zero, its 0.000075

Question 9  
The sizecolumn in bigquery-public-data:github\_repos.sample\_contents is in bytes. Given that a  
kilobyte is 1024 bytes, use the SELECT \* REPLACE syntax to write a query that returns the  
table with size recalculated in kilobytes. (Run your query for this exercise. The query will  
process 24 GB, but this still only costs 12 cents to run once you’ve used up the monthly free  
terabyte for your account.)Provide your query.

Your Answer:

SELECT

  \* REPLACE( size / 1024 AS size)

FROM

`bigquery-public-data.github\_repos.sample\_contents`

Question 10  
Use your query from question 9 as a subquery in a CTE. From the output of the subquery, select  
out id and size. Rename the size column to size\_kb. “Select \* except” syntax keeps all columns  
and doesn’t allow renaming - it simply allows us to do substitutions. With a CTE, we can rename  
and down select to fewer columns. Also notice that the BigQuery optimizer doesn’t scan the  
unneeded columns in the subquery - the new indicated scan size is 139 MB.

Provide your query.

Your Answer:

WITH cte\_cont AS (

SELECT

 \* REPLACE( size / 1024 AS size)

FROM

`bigquery-public-data.github\_repos.sample\_contents`)

SELECT id,size as size\_kb

FROM

cte\_cont