Gabelli School of Business

Fordham University

ISGBBYGB 7990 – Big Data Analytics – Fall 2016

**Assignment 1: Google BigQuery**

Due date: upload assignment by 11-59 pm, 9/26/16

Penalty per day delay: 10% each day up to 3 days after which assignment is unacceptable.

***What to submit***?

1. In your executive summary report, have a cover page with your details.
2. Please have page numbers in top right corner. Double-space the summary.
3. Each page of your report must have a title.

Organize all components of each part separately (don’t put all summaries together followed by all printouts).

1. Add your summary and all the result csv files to a zipped file, name it as [your name Assgn 1]. Upload it to blackboard.

**Part A**

1. Complete the Workshop 1 Google BigQuery handout.
2. For each example (5 examples), in your one page, double-spaced, executive summary briefly describe the following:
   1. Problem statement.
   2. What was done including analysis based on the printouts.
   3. Managerial conclusion free of technical jargon.
   4. Relevant screenshots of your query output.

**Part B**

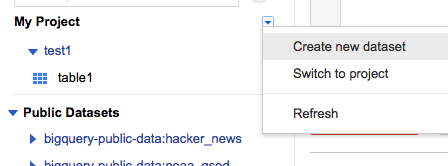
1. For the exercise below, in your one page, double-spaced, executive summary briefly describe the following:
   1. Problem statement.
   2. What was done including analysis based on printouts.
   3. Managerial conclusion free of technical jargon.
   4. Relevant screenshots of your query output.

***What to do?***

1. Upload the sales data file (sales.csv) to Google BigQuery.
2. Perform SQL query analysis.
3. Download the result.

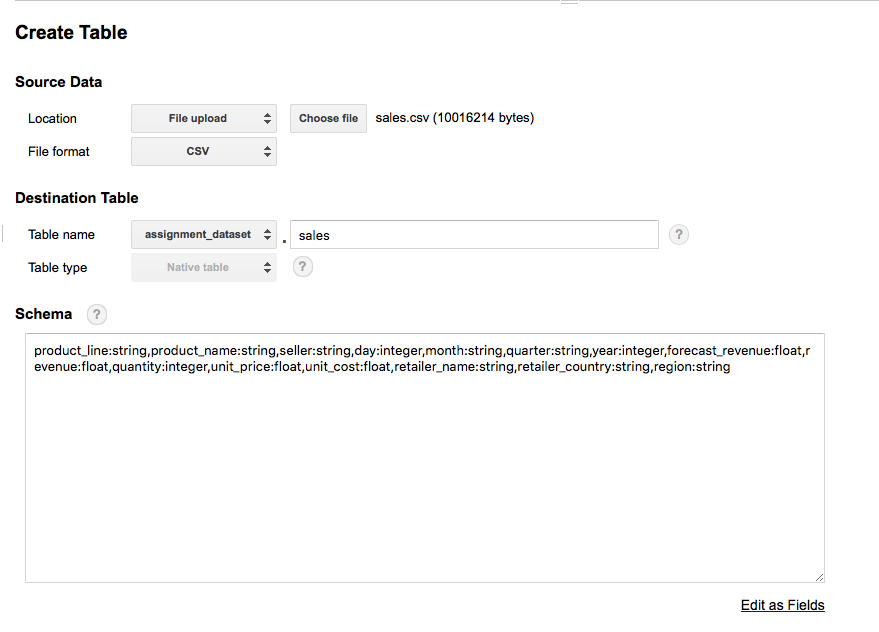
Sales.csv recorded the sales information of a multi-national company. It recorded each sales order in the past 2 years. Data includes: retailers, cost, price, order date, product name and type. Study the data.

1. Under the current project, create a new dataset, name it “**assignment\_dataset**”.



Note: All the names used are case sensitive.

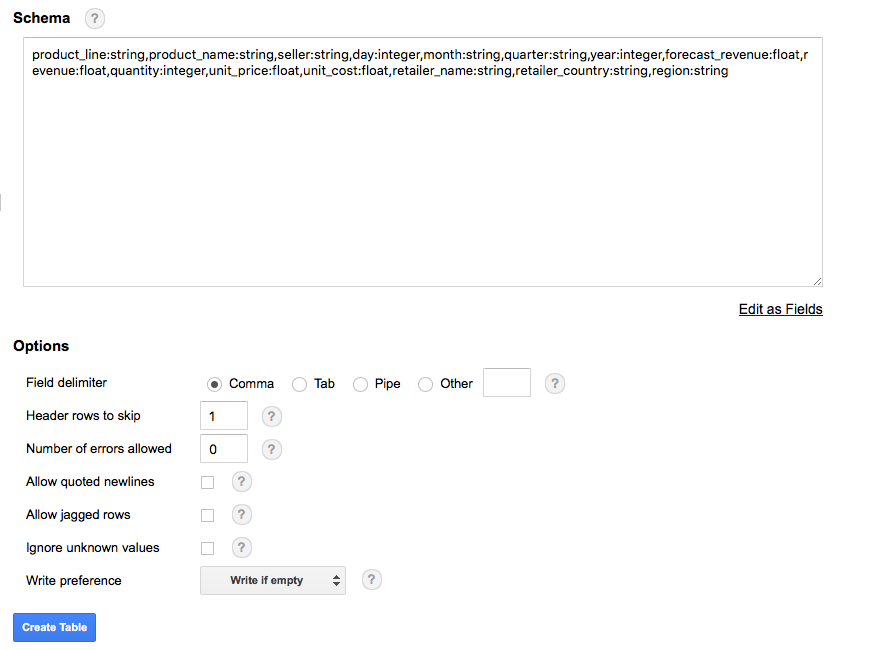
1. In this dataset, create a new table.



1. Define the schema, typing the following schema into the box:

**product\_line:string,product\_name:string,seller:string,day:integer,month:string,quarter:string,year:integer,forecast\_revenue:float,revenue:float,quantity:integer,unit\_price:float,unit\_cost:float,retailer\_name:string,retailer\_country:string,region:string**

1. In advanced options set first row as header row and put **“Header rows to skip : 1 :”**



1. **Query 1**: List the average revenue and average forecasted revenue for all product line in all regions in the past.

The average revenue equals to the sum of revenue in each order divided by the number of orders. In SQL, it can be represented as following formula: sum(revenue)/count(\*)

Therefore the complete query is:

**SELECT product\_line, region, year, sum(revenue)/count(\*) as avg\_revenue,**

**sum(forecast\_revenue)/count(\*) as avg\_forecast,**

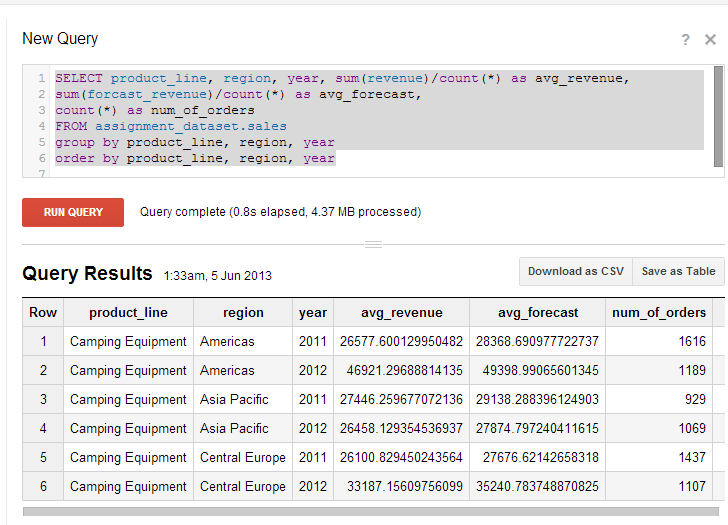
**count(\*) as num\_of\_orders**

**FROM assignment\_dataset.sales**

**group by product\_line, region, year**

**order by product\_line, region, year;**

Type in and run this query to check the result.



1. **Query 2:** Find the most profitable product line in the Americas in 2012.

We can run a nested query to calculate the most profitable product line based on the result of last query.

A complete query is:

**select product\_line,**

**sum(avg\_revenue\*num\_of\_orders) as total\_revenue from**

**(SELECT product\_line, region, year, sum(revenue)/count(\*) as avg\_revenue,**

**sum(forecast\_revenue)/count(\*) as avg\_forecast,**

**count(\*) as num\_of\_orders**

**FROM assignment\_dataset.sales**

**group by product\_line, region, year**

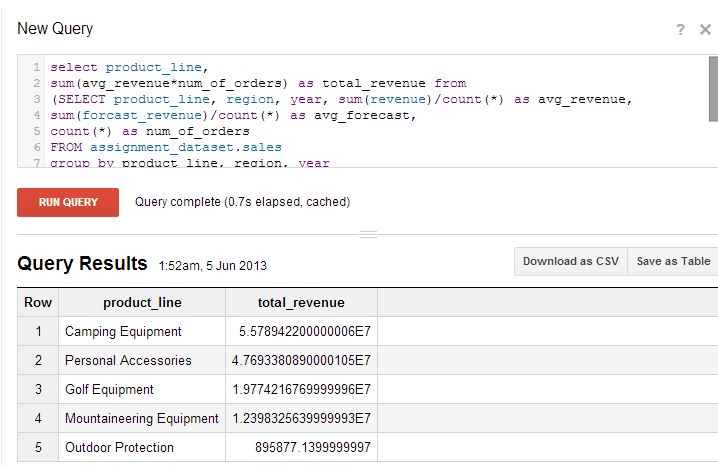
**order by product\_line, region, year)**

**where year =2012 and region='Americas'**

**group by product\_line**

**order by total\_revenue desc;**

Run the Query and check the result.



(Notice: If you are familiar with the data schema and SQL, other different queries can be used in order to solve this problem; for e.g.,

**SELECT product\_line,sum(revenue)as total**

**FROM assignment\_dataset.sales where year=2012 and region='Americas'**

**group by product\_line**

**order by total desc**;

Save your result as “sales result.csv”