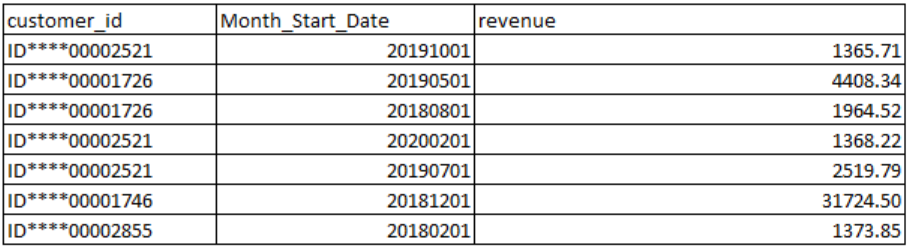
**AWS GC BI – SQL Test**

**Situation**

There is a table (Table 1 as below) where it keeps the historical spending(revenue) of each customer in the past few years by month, the table has a few million records.

**Table 1:**



**Month\_Start\_Date:** calendar month in the format (YYYYMMDD (20190101), “DD” part is always “01” to represent the given month.

**Revenue:** Spending (revenue) of each customer in the given month.

As the cloud utility business model, assuming each customer will continue having spending in the months to come once get started (the spending could be up and down month over month). **So** **in order to identify critical movements, here is the SQL in AMAZON Redshift (Redshift, simply saying, it is an analytical database)**

With revenue\_base as

(

SELECT customer\_id

, month\_start\_date

, datediff(month,c.start\_point\_date,month\_start\_date)+1 as month\_sequence

, revenue

FROM gc\_adhoc.fact\_fake\_account\_revenue as f

INNER JOIN (

SELECT

MIN(month\_start\_date) AS start\_point\_date

FROM gc\_adhoc.fact\_fake\_account\_revenue

) as c

on 1=1

)

SELECT

a.customer\_id

,sum(CASE WHEN month\_sequence= 1 then revenue else 0 end) as revenue\_2018\_jan

,count(distinct month\_start\_date) as A

,(sum((Month\_Sequence-u\_x)\*(revenue-u\_y)))/sum (power((Month\_Sequence-u\_x),2)) AS B

,MAX(u\_y)-(sum((Month\_Sequence-u\_x)\*(revenue-u\_y)))/sum(power((Month\_Sequence-u\_x),2))\*MAX(u\_x) AS C

FROM revenue\_base AS a

INNER JOIN

(

select

customer\_id

,avg(Month\_Sequence\*1.0) AS u\_x

,avg(revenue) AS u\_y

from revenue\_base

GROUP BY 1

) AS a\_g ON a.customer\_id=a\_g.customer\_id

GROUP BY a.customer\_id

HAVING sum (power((Month\_Sequence-u\_x),2))>0

**Questions need your answers**

1. **At the very beginning, we don’t have the table created yet in the database, and there are hundreds of csv files with exact the same column as table 1 in a consolidated folder, and the folder are updated every month with new files added, please descript what would you do in order to create a solution to refresh the data monthly.**

*[As an example, you can start with*

*step 1: I am going to create the table in Redshift (or any other database you are familiar with)]*

*….*

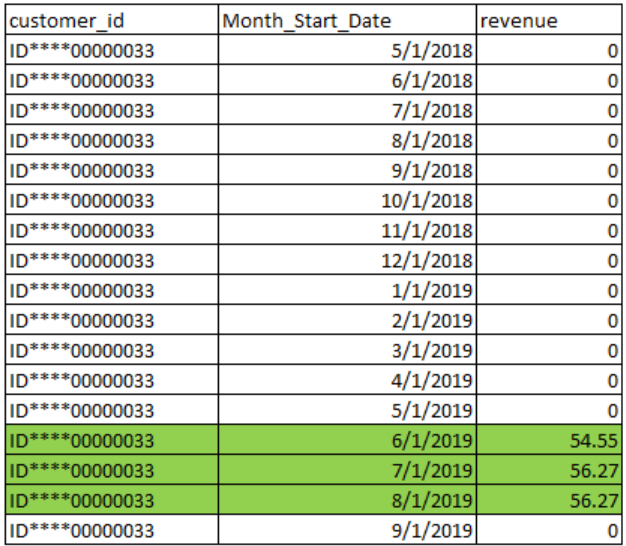
1. **What is the final output of the SQL against the table 1? Please list all of the output columns with the explanation of each and also please be specific what kind of customers are not captured by the SQL**

*[For example]*

|  |  |
| --- | --- |
| Output Column Name | Detail Explation |
| customer\_id | the id of customer to represent one customer |
| revenue\_2018\_jan | total revenue of given customer in Jan 2018 |
| …. |  |

*[The query excludes those customers with……]*

1. **Study the logic of the SQL query, description the purpose of the SQL query and explain where could it be used in what scenario per your best knowledge.**
2. **It’s been noticed that some of records are not captured in the query and the data by customer are not always consistent to have every month, please modify the SQL to**
3. **To capture the missing customers**
4. **To be based on the start month and end month for each customer, for example below, the start month should be 20180501 and the end month should be 20190801 for customer of ID\*\*\*\*00000033**



*Please put your revised SQL here:*