### **SQL Question - Calculating Retention Metrics**

#### **Question Prompt**

We capture data on how users are using our product, with the schemas below. Using this data we would like to report on monthly “engaged” retention rates. **Monthly “engaged” retention** is defined here as the % of users from each registration cohort that continued to use the product as an “engaged” user having met the threshold of >= 30 minutes per month. We are looking for the retention metric calculated for within 1-3 calendar months post registration.

**Write a SQL query** whose output format is in the following, using the input tables described below:

**Results Table: retention**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| registration\_month | total\_users | m1\_retention | m2\_retention | m3\_retention |
| Jan, 2019 | 3 | 67% | 33% | 33% |
| Feb, 2019 | 1 | 100% | 0% | 0% |

Note that user *aaa* used the product 2 times within the first month (2019-01-03, 2019-02-01), 0 times in the 2nd month, and 1 time in the 3rd month (2019-03-04), post the user aaa’s initial registration (2019-01-03). User *bbb* used the product once in 1st month (2019-01-03) and once in 2nd month (2019-02-04) post registration (2019-01-02), but the 1st month usage is <30 minutes so the user doesn’t count in the m1\_retention metric. Note that we want to calculate this usage metric as across all geographies.

**Input Table: users**

|  |  |
| --- | --- |
| user\_id | registration\_date |
| aaa | 2019-01-03 |
| bbb | 2019-01-02 |
| ccc | 2019-01-15 |
| ddd | 2019-02-07 |

To get this exercise done, I have created database “mydb”

**create database mydb;**

I have created dummy table named “registration\_data”

**-- Create the 'registration\_data' table**

**CREATE TABLE registration\_data (**

**user\_id VARCHAR(3),**

**registration\_date DATE**

**);**

I have inserted the values in this exercise as an example

**-- Insert data into the 'registration\_data' table**

**INSERT INTO registration\_data (user\_id, registration\_date)**

**VALUES**

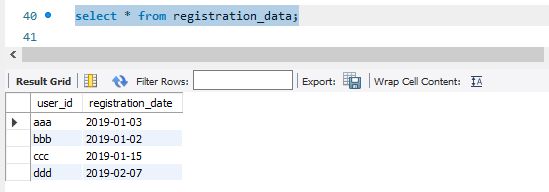
**('aaa', '2019-01-03'),**

**('bbb', '2019-01-02'),**

**('ccc', '2019-01-15'),**

**('ddd', '2019-02-07');**

**select \* from registration\_data;**



**Input Table: usage**

|  |  |  |  |
| --- | --- | --- | --- |
| user\_id | usage\_date | usage\_location | time\_spent |
| aaa | 2019-01-03 | US | 38 |
| aaa | 2019-02-01 | US | 12 |
| aaa | 2019-03-04 | US | 30 |
| bbb | 2019-01-03 | US | 20 |
| bbb | 2019-02-04 | Canada | 31 |
| ccc | 2019-01-16 | US | 40 |
| ddd | 2019-02-08 | US | 45 |

#### **Input Table Schemas**

The schemas for the input tables are described below.

**Input Table Schema: users**

|  |  |
| --- | --- |
| Column Name | Type |
| user\_id | string |
| registration\_date | date |

Id is the unique identifier for our users and is the primary key column for this table. Registration\_date reflects the date that the user first registered for an account to use the product. Each row of the table is unique and represents the record for a user who signed up for using our product.

**Input Table Schema: usage**

|  |  |
| --- | --- |
| Column Name | Type |
| user\_id | string |
| usage\_date | date |
| usage\_location | string |
| time\_spent | int |

Id is the unique identifier for our users. Usage\_date reflects the date that the user is using the product. Usage\_location shows where the usage is taking place at the moment of use. Time\_spent represents how long the user is spending on the product in minutes. Each row of the table represents a daily record for the total amount of time that a specific user is spending on the product.

I have created another table; named u\_usage

**-- Create the 'usage' table**

**CREATE TABLE u\_usage (**

**user\_id VARCHAR(3),**

**usage\_date DATE,**

**usage\_location VARCHAR(10),**

**time\_spent INT**

**);**

Insert the data

**-- Insert data into the 'usage' table**

**INSERT INTO u\_usage (user\_id, usage\_date, usage\_location, time\_spent)**

**VALUES**

**('aaa', '2019-01-03', 'US', 38),**

**('aaa', '2019-02-01', 'US', 12),**

**('aaa', '2019-03-04', 'US', 30),**

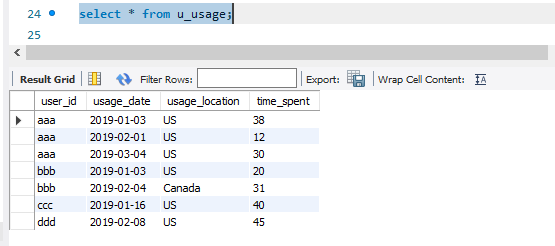
**('bbb', '2019-01-03', 'US', 20),**

**('bbb', '2019-02-04', 'Canada', 31),**

**('ccc', '2019-01-16', 'US', 40),**

**('ddd', '2019-02-08', 'US', 45);**

Show records of table



Now lets write the query to get following outcome

**Results Table: retention**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| registration\_month | total\_users | m1\_retention | m2\_retention | m3\_retention |
| Jan, 2019 | 3 | 67% | 33% | 33% |
| Feb, 2019 | 1 | 100% | 0% | 0% |

**WITH MonthlyUsage AS (**

**SELECT**

**u.user\_id,**

**r.registration\_date AS reg\_date,**

**u.usage\_date,**

**TIMESTAMPDIFF(MONTH, r.registration\_date, u.usage\_date) AS month\_diff,**

**u.time\_spent**

**FROM**

**u\_usage u**

**JOIN**

**registration\_data r ON u.user\_id = r.user\_id**

**)**

**, RetentionMetrics AS (**

**SELECT**

**DATE\_FORMAT(reg\_date, '%b, %Y') AS registration\_month,**

**COUNT(DISTINCT user\_id) AS total\_users,**

**COUNT(DISTINCT CASE WHEN month\_diff = 0 THEN user\_id END) AS m1\_retention,**

**COUNT(DISTINCT CASE WHEN month\_diff = 1 THEN user\_id END) AS m2\_retention,**

**COUNT(DISTINCT CASE WHEN month\_diff = 2 THEN user\_id END) AS m3\_retention**

**FROM**

**MonthlyUsage**

**WHERE**

**month\_diff <= 2 AND time\_spent >= 30**

**GROUP BY**

**registration\_month**

**)**

**SELECT**

**registration\_month,**

**total\_users,**

**CONCAT(ROUND((m1\_retention / total\_users) \* 100, 2), '%') AS m1\_retention,**

**CONCAT(ROUND((m2\_retention / total\_users) \* 100, 2), '%') AS m2\_retention,**

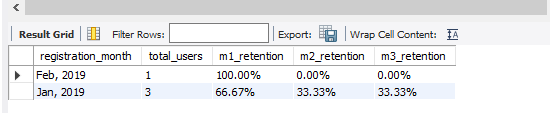
**CONCAT(ROUND((m3\_retention / total\_users) \* 100, 2), '%') AS m3\_retention**

**FROM**

**RetentionMetrics**

**ORDER BY**

**registration\_month;**

****