Задание 1. Определение пола

В таблице Transactions дана информация о транзакциях клиентов.
В таблице Clients дана информация о клиентах с различной датой актуальности данных.

Задания:
1.Получить таблицу с наиболее актуальными данными по всем клиентам с атрибутным составом:
- идентификатор клиента
- ФИО
- пол

2.Получить информацию о том, сколько женщин и мужчин совершало покупки в терминалах за январь 2020 в разрезе MCC кодов.
3. (Дополнительно) Выделить фамилию, имя и отчество из ФИО.

Таблица Transactions (пример данных)

trans_id	client_id	partition_dt	sum_trans	mcc_code
1111	100000	01.01.2020	9520	5611
1112	200000	08.01.2020	1623	5641
1113	300000	09.01.2020	1799	7832
1114	400000	14.01.2020	728	7538
1115	400000	15.01.2020	624	7832
1116	500000	16.01.2020	542	5641
1117	600000	21.01.2020	7982	7832
1118	700000	21.01.2020	886	7538
1119	700000	31.01.2020	432	5641
2000	800000	04.01.2020	680	5611
2001	100000	04.02.2020	3472	5641
2002	400000	05.02.2020	572	5641
2003	500000	06.02.2020	783	7832
2004	600000	07.02.2020	9774	5611
2005	700000	07.02.2020	8717	5641
2006	800000	07.02.2020	502	7832

Таблица Clients (пример данных)

client_id	partition_dt	client_name
100000	01.12.2020	Потапов
200000	01.07.2018	Рыбакова Валерия Львовна
300000	01.01.2021	Трофимов Андрей Ильич
400000	01.08.2008	Балашова Ксения
500000	01.06.2007	Gavrilova Anastasiya Kirillovna
600000	01.02.2021	Быков Семён Семёнович
800000	01.01.2020	Никольская Ульяна Антоновна
800000	01.02.2020	Панфилова Ульяна Антоновна

Решение:

```sql
-- Получаем информацию о том, сколько женщин и мужчин совершало покупки в терминалах за январь 2020
--в разрезе MCC кодов
SELECT
    mcc_code,
    client_gender,
    count(client_id) as num_clients
FROM
    (
    SELECT d.client_id as client_id,
            c.client_name,
            c.client_gender,
            d.partition_dt as partition_dt,
            d.sum_trans,
            d.mcc_code

    FROM
        (
        -- Получаем таблицу с наиболее актуальными данными по всем клиентам с атрибутным составом
        SELECT a.client_id,
            b.client_name,
            splitByChar(' ',client_name) as fio_array,
            arrayElement(fio_array, 1) as client_surname,
            -- (Дополнительно) Выделить фамилию, имя и отчество из ФИО
            --arrayElement(fio_array, 2) as client_name,
            --arrayElement(fio_array, 3) as client_patronymic,
            multiIf(client_surname like '%а', 'Ж', client_surname like '%а', 'Ж',
            client_surname like '%ая', 'Ж', client_surname like '%ая', 'Ж', 'М')
            as client_gender

        FROM
            (
            -- Получаем наиболее актуальные данные
            SELECT DISTINCT client_id,
                max(partition_dt) as partition_dt
            FROM test.Titova_Clients
            GROUP BY client_id
            ORDER BY client_id
            ) as a

        LEFT JOIN test.Titova_Clients as b on a.client_id = b.client_id and a.partition_dt = b.partition_dt
        ) as c

    -- Добавляем данные из таблицы Transactions
    RIGHT JOIN
        (
        SELECT client_id,
            partition_dt,
            sum_trans,
            mcc_code
        FROM test.Titova_Transactions
        WHERE partition_dt between '2020-01-01' and '2020-01-31'
        ) as d on c.client_id = d.client_id
    )

GROUP BY mcc_code, client_gender
ORDER BY mcc_code
```

Вывод:

Table + Add Visualization

mcc_code	client_gender	num_clients
5,611	М	1
5,611	Ж	1
5,641		1
5,641	Ж	2

Задание 2. Сводная таблица по терминалам

В таблице Terminal_transactions дана информация о сумме и количестве транзакций терминалов за 8 месяцев 2021 года.

Задание:
Необходимо написать SQL-запрос, который переводит эту таблицу в pivot и для каждого терминала посчитать долю суммы транзакций каждого месяца от общей суммы в формате:

Terminal_id	Sum_1_2020	Sum_2_2020	Sum_3_2020	…	Prc_1_2020	Prc_2_2020	…

К таблице Terminal_transactions можно обращаться только один раз

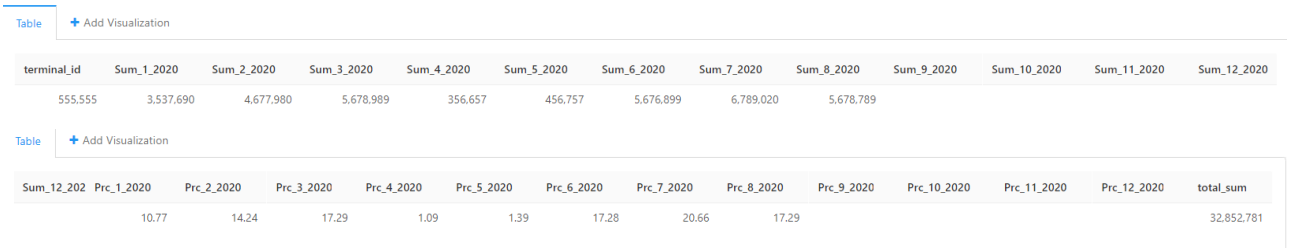
Таблица Terminal_transactions (пример данных)

terminal_id	partition_dt	sum_trans
555555	01.01.2021	3537690
555555	01.02.2021	4677980
555555	01.03.2021	5678989
555555	01.04.2021	356657
555555	01.05.2021	456757
555555	01.06.2021	5676899
555555	01.07.2021	6789020
555555	01.08.2021	5678789

Решение:

```sql
SELECT terminal_id,
    SUM(CASE WHEN partition_mnth = '1' THEN sum_trans END) as Sum_1_2020,
    SUM(CASE WHEN partition_mnth = '2' THEN sum_trans END) as Sum_2_2020,
    SUM(CASE WHEN partition_mnth = '3' THEN sum_trans END) as Sum_3_2020,
    SUM(CASE WHEN partition_mnth = '4' THEN sum_trans END) as Sum_4_2020,
    SUM(CASE WHEN partition_mnth = '5' THEN sum_trans END) as Sum_5_2020,
    SUM(CASE WHEN partition_mnth = '6' THEN sum_trans END) as Sum_6_2020,
    SUM(CASE WHEN partition_mnth = '7' THEN sum_trans END) as Sum_7_2020,
    SUM(CASE WHEN partition_mnth = '8' THEN sum_trans END) as Sum_8_2020,
    SUM(CASE WHEN partition_mnth = '9' THEN sum_trans END) as Sum_9_2020,
    SUM(CASE WHEN partition_mnth = '10' THEN sum_trans END) as Sum_10_2020,
    SUM(CASE WHEN partition_mnth = '11' THEN sum_trans END) as Sum_11_2020,
    SUM(CASE WHEN partition_mnth = '12' THEN sum_trans END) as Sum_12_2020,
    (Sum_1_2020 / total_sum)*100 as Prc_1_2020,
    (Sum_2_2020 / total_sum)*100 as Prc_2_2020,
    (Sum_3_2020 / total_sum)*100 as Prc_3_2020,
    (Sum_4_2020 / total_sum)*100 as Prc_4_2020,
    (Sum_5_2020 / total_sum)*100 as Prc_5_2020,
    (Sum_6_2020 / total_sum)*100 as Prc_6_2020,
    (Sum_7_2020 / total_sum)*100 as Prc_7_2020,
    (Sum_8_2020 / total_sum)*100 as Prc_8_2020,
    (Sum_9_2020 / total_sum)*100 as Prc_9_2020,
    (Sum_10_2020 / total_sum)*100 as Prc_10_2020,
    (Sum_11_2020 / total_sum)*100 as Prc_11_2020,
    (Sum_12_2020 / total_sum)*100 as Prc_12_2020,
    sum(sum_trans) as total_sum
FROM
    (
    SELECT terminal_id,
        EXTRACT(MONTH FROM partition_dt) as partition_mnth,
        sum_trans
    FROM test.Terminal_Transactions
    )
GROUP BY terminal_id
```

Вывод:



Задание 3. Определение брендов магазинов

По названиям терминалов можно определить, к какому бренду относится магазин. Но часто эти названия написаны по-разному для одного и того же бренда и есть вероятность что названия терминалов у разных брендов будут похожи.

Задание:
Необходимо написать регулярные выражения, которые будут правильно определять бренд магазинов "Азбука Вкуса" и "Белорусский трикотаж".

Примеры некоторых наименований терминалов для «Азбуки вкуса» (для белорусского трикотажа необходимо самостоятельно придумать возможные варианты):
AZBUKA VKUSA 82
AZBUKA VKUSA 131 PROSP
AZBUKA VKUSOV
AB DAILY.
OOO AZBUKA VKUSA
AV MARKET
SP_AV AZBUKAVKUSA
"AZBUKA VKUSA" SHOP
AV AZBUKAVKUSA..
AV DAILY..
AZBUKA VKUSA PROSPEKT
AV .AZBUKAVKUSA
MAGAZIN AZBUKA VKUSA
MINIMARKET AZBUKA VKUSA
AV AZBUKAVKUSA.
AZBUKAVKUSA
"AZBUKAVKUSA"
AZBUKA_VKUSA 80
AV.MARKET
"AZBUKA VKUSA" SHOP GO

Примеры наименований других магазинов, которые не должны выбираться с помощью регулярных выражений для азбуки вкуса:
AZBUKA
AVE
VKUSVILL
MAGAZIN
DALY MARKT
ABC

Решение:

```
SELECT terminal_name,
    case when multiMatchAny(replace(replace(terminal_name, '  ', ' '), '  ', ' '),
    ['AZBUKA\s\S', 'AZBUKA_\S', 'AB\s\S', 'AV\s\S', 'AV.\S', 'AV\s', 'AZBUKAVKUSA']) = '1'
    then 'AZBUKA VKUSA'
    when multiMatchAny(replace(replace(terminal_name, '  ', ' '), '  ', ' '),
    ['BELORUSSKIY\s\S', 'BELORUSSKIY_\S', 'BEL\s\S', 'BEL_\S', 'BEL.'])= '1'
    then 'BELORUSSKIY TRIKOTAZH' end as brand_name
FROM test.Terminal_name
ORDER BY brand_name DESC
```

Вывод:

Table + Add Visualization

terminal_name	brand_name
"AZBUKA VKUSA" SHOP	AZBUKA VKUSA
"AZBUKA VKUSA" SHOP GO	AZBUKA VKUSA

< 1 2 >

Задание 4. Определение брендов магазинов

Таблица Merchants является ненормализованной. Она содержит атрибут listofterminal – список терминалов этого магазина с информацией о времени работы.

Задание:
Необходимо привести таблицу к виду магазин – терминал – время работы

Таблица Merchants (пример данных)

merchant	listofterminal
11111	11222222: 12:00-20:00; 11222226: 14:00-17:00
12344	333333:; 444444: 9:00-23:00; 555555: 9:00-23:00
22222	22222221: 9:00-22:00

Решение:

```
SELECT merchant,
    terminal_array2[1] AS terminal,
    replace(concat(terminal_array2[2],':', terminal_array2[3],':',terminal_array2[4]), '::', '')
    as terminal_working_hours

FROM
    (
    -- Разбиваем терминал и время работы на отдельные элементы
    SELECT merchant,
        splitByString(':', terminal_array) as terminal_array2
    FROM
        (
        -- Раскладываем каждое отдельное значение массива в строку
        SELECT merchant, terminal_array
        FROM
            (
            -- Разбиваем listofterminal на массив с элементами
            SELECT merchant,
                splitByString(';', listofterminal) as terminal_array
            FROM test.Merchants
            )
        ARRAY JOIN terminal_array
        )
    )
```

Вывод:

Table + Add Visualization

merchant	terminal	terminal_working_hours
11,111	11222222	12:00-20:00
11,111	11222226	14:00-17:00
12,344	333333	
12,344	444444	9:00-23:00
12,344	555555	9:00-23:00
22,222	22222221	9:00-22:00