



**Академия
Аналитиков
Авито**

AAA SQL+DB

7. Операции со множествами

Группировки с подытогом



Повторим пройденное

- ▶ Чем отличаются вложенные от связанных подзапросов
- ▶ Как применять операторы ALL и ANY
- ▶ Как применять операторы EXISTS и NOT EXISTS
- ▶ Как получить 10-ую строку в таблице, по выбранному нами порядку

- ▶ Что можно сделать с помощью CASE WHEN ?
- ▶ Как получить квартал из колонки с датой?

ЧТО БУДЕМ ДЕЛАТЬ СЕГОДНЯ

- ▶ Работа со множествами
 - Объединение
 - Разность
 - Пересечение
- ▶ Группировки с подытогом
 - Итог по выборочным группам
 - Итог по всем разрезам
 - Неаддитивные метрики



**Академия
Аналитиков
Авито**

Работа со множествами

Объединение, пересечение, разность



Зачем это нужно?



Обсуждение



5 минут

Зачем это нужно?

Сценарии из нашего опыта

- Нужно построить аналитику на 2+ наборах данных (не загружая их в одну таблицу)
- Нужно взять только уникальные строки из нескольких таблиц
- Нужно понять одинаковые ли выборки или есть расхождения в каких-то колонках
- Нужно понять есть ли пересечение в 2х или больше таблицах, без сложных join-ов
- Нужно создать тестовое отношение, не материализуя его в таблицу, чтобы проверить какую-то функцию



Обсуждение



UNION ALL



Вопросы по прериду

- как работает **UNION ALL** ?
- какие требования к отношениям, которые мы хотим объединить?
- что будет если убрать **ALL** ?
- какие названия будут у колонок получившегося отношения?



Обсуждение



3 минуты

Задача 1: Простое объединение

Условие:

Даны 2 простые таблицы с целыми числами

d7_set1 (n int);

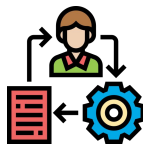
d7_set2 (n int);

Вопрос:

Выведите все различные числа из табличек d7_set1, d7_set2.

Решите задачу только с использованием union all.

	123 n ↑↓
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13
14	14
15	15
16	16
17	17
18	18



Практика



3 минуты

Задача 2: Фамилии новых пользователей

Условие:

Покупатели d7_buyer(id int, surname text, ...);

Продавцы d7_seller(id int, surname text, ...);

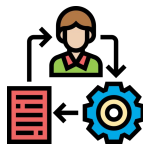
Менеджеры d7_manager(id int, surname text, ...);

Регистрации d7_user(id serial, registration_date date, ...);

Вопрос:

Выведите все фамилии пользователей без учета роли при условии регистрации после 2020-11-01.

Сравнение дат работает с неявным приведением типа
`registration_date >= '2020-11-01'`



Практика

	T surname 🏆🏆
1	Игнатьев
2	Ковров
3	Кузикин
4	Копориков
5	Потёмкин
6	Алехин
7	Винокуров
8	Савасин
9	Дорогов
10	Игнатенков
11	Евлентьев
12	Кантонистов



5 минут

Задача 3: Количество активных пользователей

Условие:

Покупатели d7_buyer(id int, surname text, last_action_date date);

Продавцы d7_seller(id int, surname text, last_action_date date);

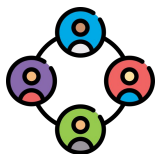
Менеджеры d7_manager(id int, surname text, last_action_date date);

Регистрации d7_user(id serial, registration_date date, ...);

	123 count ↑↓
1	54

Вопрос:

Посчитайте количество пользователей, совершавших действия после 2020-06-01



Работаем вместе

Решите задачу с указанием константы с датой только 1 раз



12 мин



Академия
Аналитиков
Авито

Пересечение и разность

INTERSECT ALL
EXCEPT ALL



Вопросы по прериду

- как работает **EXCEPT ALL** ? какие вы знаете аналоги?
- как работает **INTERSECT ALL** ?
- какие требования к отношениям, которые участвуют в операции?
- какие названия будут у колонок получившегося отношения?
- что будет если убрать **ALL** ?



Обсуждение



2 минут

Задача 4: Простая разность

Условие:

Даны 2 простые таблицы с целыми числами

d7_set1 (n int);

d7_set2 (n int);

Вопрос:

1. Найдите записи из d7_set1, которых нет в d7_set2.
2. Реализуйте то же самое через join, без использования ключевых слов except или minus



Практика

	123 n 1 ?
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9



5 минут

Задача 5: Простое пересечение

Условие:

Даны 2 простые таблицы с целыми числами

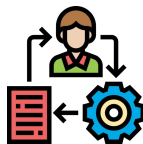
d7_set1 (n int);

d7_set2 (n int);

	123 n ?
1	11
2	10
3	14
4	13
5	12

Вопрос:

1. Найдите записи, которые и в d7_set1 и в d7_set2.
2. Реализуйте то же самое через join, без использования INTERSECT



Практика



5 минут

Задача 6: Исключительные продавцы

Условие:

Покупатели d7_buyer(id int, surname text, ...);

Продавцы d7_seller(id int, surname text, ...);

Менеджеры d7_manager(id int, surname text, ...);

Регистрации d7_user(id serial, registration_date date, ...);

	123 count	↑↓
1	21	

Вопрос:

Найдите количество продавцов, которые не совершали действие как покупатель после 2020-06-01



Практика



5 минут

Задача 7: Первый постоянный байер



Условие:

Покупатели d7_buyer(id int, surname text, ...);

Продавцы d7_seller(id int, surname text, ...);

Менеджеры d7_manager(id int, surname text, ...);

Регистрации d7_user(id serial, registration_date date, ...);

	 min 
1	2020-01-20

Вопрос:

Найдите минимальную дату регистрации покупателя, который не совершал действий, как продавец



Работаем вместе



12 мин

Задача 8: Только нужные месяцы

Условие:

Покупатели d7_buyer(id int, surname text, ...);

Продавцы d7_seller(id int, surname text, ...);

Менеджеры d7_manager(id int, surname text, ...);

Регистрации d7_user(id serial, registration_date date, ...);

Вопрос:

Найдите месяцы регистрации, в которые не было пользователей, одновременно являющихся продавцами и покупателям. Колонку с месяцем приведите к дате.



Работаем вместе

Используйте except + intersect (без join и group by)

registration_month
2020-01-01
2020-02-01
2020-03-01
2020-05-01
2020-06-01
2020-07-01
2020-08-01
2020-09-01
2020-12-01



17 мин

Summary по блоку 1



Теория



2 минуты

Summary по блоку 1

- ▶ **UNION, EXCEPT, INTERSECT** - количество и типы колонок должны совпадать
- ▶ Если в запросе несколько **UNION, EXCEPT, INTERSECT** - нужны скобки
- ▶ Если убрать **ALL** - выполнится distinct



Теория



Перерыв

10 минут



grouping sets, rollup, cube



Зачем это нужно?



Обсуждение



5 минут

Зачем это нужно?

Сценарии из нашего опыта, примеры разрезов

- Нужно смотреть изменение метрики по каждому городу, по региону, по стране, желательно на одном отчете.
- Нужно смотреть количество новых пользователей в каждой подкатегории и в каждой категории
- Нужно смотреть данные по дням, по неделям и по месяцам.



Обсуждение

Вопросы по прериду

- Что делает GROUPING SETS, где его писать ?
- Как добавить агрегацию по всей таблице?
- Как понять, к какому из разрезов относится строчка?
- Какие комбинации дает ROLLUP (col1, col2)
- Какие комбинации дает CUBE (col1, col2), назовите те, которых нет в предыдущем



Обсуждение

Что такое неаддитивная метрика? Можно пример.



8 минут

Вопросы по прериду

- Что делает GROUPING SETS, где его писать ?
 - **GROUP BY GROUPING SETS ((c1), (c2, c3))**
 - Как добавить агрегацию по всей таблице?
 - GROUP BY GROUPING SETS ((c1), (c2, c3), ())
 - Как понять, к какому из разрезов относится строчка?
 - **SELECT GROUPING(c1, c2, c3)::bit(3) - битовая маска**
можно понять при желании
 - Какие комбинации дает ROLLUP (col1, col2)
 - **((col1, col2), (col1), ())**
 - Какие комбинации дает CUBE (col1, col2), назовите те, которых нет в предыдущем
 - **((col1, col2), (col1), (col2), ())**
- Что такое неаддитивная метрика? Можно пример.
- **COUNT(distinct col1)**



Обсуждение



2 минут

Пример: Неаддитивная метрика (с “решением”)

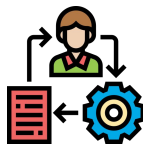
event_date ÷	room_id ÷	floor_id ÷	user_id ÷	paid ÷
2021-01-01	A	1	1	1000
2021-01-02	B	1	1	1000
2021-01-02	A	2	2	1000

Условие:

Лог отеля d7_hotel (event_date, room_id, floor_id, user_id, paid);

Вопрос:

Подсчитать *количество* уникальных посетителей отеля и уплаченную сумму в разрезе этажей и комнат.



Практика

Подвести подытог по этажам.



8 минут

Пример: Неаддитивная метрика

“Решение”
(найдите ошибку):

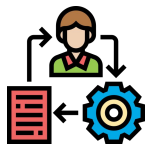
event_date ÷	room_id ÷	floor_id ÷	user_id ÷	paid ÷
2021-01-01	A	1	1	1000
2021-01-02	B	1	1	1000
2021-01-02	A	2	2	1000

```
select room_id, floor_id,  
       sum(paid) paid, count(distinct user_id) count  
from d7_hotel  
group by 1, 2;
```

room_id ÷	floor_id ÷	sum ÷	count ÷
A	1	1000	1
A	2	1000	1
B	1	1000	1

```
select floor_id, sum("paid") paid, sum("count") count  
from (...) sq  
group by 1
```

floor_id ÷	paid ÷	count ÷
2	1000	1
1	2000	2



Практика

Пример: Неаддитивная метрика

“Решение”
(найдите ошибку):

event_date ÷	room_id ÷	floor_id ÷	user_id ÷	paid ÷
2021-01-01	A	1	1	1000
2021-01-02	B	1	1	1000
2021-01-02	A	2	2	1000

```
select room_id, floor_id,  
       sum(paid) paid, count(distinct user_id) count  
from d7_hotel  
group by 1, 2;
```

room_id ÷	floor_id ÷	sum ÷	count ÷
A	1	1000	1
A	2	1000	1
B	1	1000	1

```
select floor_id, sum("paid") paid, sum("count") count  
from (...) sq  
group by 1
```

floor_id ÷	paid ÷	count ÷
2	1000	1
1	2000	2



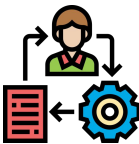
Практика

Пример: Неаддитивная метрика

Правильное решение

event_date	room_id	floor_id	user_id	paid
2021-01-01	A	1	1	1000
2021-01-02	B	1	1	1000
2021-01-02	A	2	2	1000

```
select coalesce(floor_id::varchar, 'Any') floor,
       coalesce(room_id::varchar, 'Any') room,
       sum(paid),
       count(distinct user_id)
from d7_hotel
group by rollup
       (floor_id, room_id)
```



Практика

floor_id	paid	count
2	1000	1
1	2000	2

floor	room	sum	count
1	A	1000	1
1	B	1000	1
1	Any	2000	1
2	A	1000	1
2	Any	1000	1
Any	Any	3000	2

Задача 9: Подытог с UNION ALL

Условие:

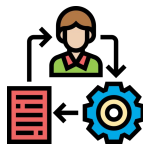
Лог отеля d7_hotel
(event_date, room_id,
floor_id, user_id, paid);

floor	room	sum	count
1	A	1000	1
1	B	1000	1
1	Any	2000	1
2	A	1000	1
2	Any	1000	1
Any	Any	3000	2

Вопрос:

Подсчитать количество уникальных посетителей отеля и уплаченную сумму в разрезе этажей и комнат.

Подвести подытог по этажам и общий



Практика



7 минут

Задача 10: Количество пользователей в разрезах

Условие:

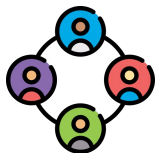
Регистрации

d7_user(id serial, registration_date date, **role** varchar(1));

	_q	_m	role	count
16	2020-07-01	2020-08-01	<null>	6
17	2020-07-01	2020-09-01	<null>	5
18	2020-07-01	<null>	b	10
19	2020-07-01	<null>	m	1
20	2020-07-01	<null>	s	7
21	2020-07-01	<null>	<null>	18
22	2020-10-01	2020-10-01	<null>	5
23	2020-10-01	2020-11-01	<null>	7
24	2020-10-01	2020-12-01	<null>	5
25	2020-10-01	<null>	b	8
26	2020-10-01	<null>	m	4
27	2020-10-01	<null>	s	7
28	2020-10-01	<null>	<null>	17

Вопрос:

Посчитайте **количество** уникальных пользователей (d7_user) в разрезе *квартала* регистрации и *месяца*.
Подведите итог по каждому разрезу.



Работаем вместе

Добавьте разбиение по роли пользователей только в кварталную статистику.



12 мин

В чем отличие, если бы мы использовали cube?

Задача 11*: Rollup через джойн на битовую маску

Условие:

Регистрации

d7_user(id serial, registration_date date, **role** varchar(1));

Вопрос:

Посчитайте **количество** уникальных пользователей в разрезе квартала регистрации и месяца.



Работаем вместе

Подведите подытог по каждому разрезу и общий подытог

Используйте join, а не rollup

	quarter_	month_	count
1	2020-01-01	2020-01-01	4
2	2020-01-01	2020-02-01	1
3	2020-01-01	2020-03-01	5
4	2020-01-01	<null>	10
5	2020-04-01	2020-04-01	8
6	2020-04-01	2020-05-01	4
7	2020-04-01	2020-06-01	6
8	2020-04-01	<null>	18
9	2020-07-01	2020-07-01	7
10	2020-07-01	2020-08-01	6
11	2020-07-01	2020-09-01	5
12	2020-07-01	<null>	18
13	2020-10-01	2020-10-01	5
14	2020-10-01	2020-11-01	7
15	2020-10-01	2020-12-01	5
16	2020-10-01	<null>	17
17	<null>	<null>	63



18 мин

Задача 11*: Rollup через джойн (ааа, сложна!!1)

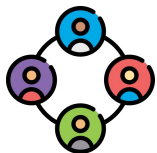
Условие:

Регистрации

d7_user(id serial, registration_date date, **role** varchar(1));

Подсказка: нужно сделать join на этот подзапрос

```
select *  
from (select 1 as q, 1 as m  
      union all  
      select 1 as q, 0 as m  
      union all  
      select 0 as q, 0 as m) _
```



Работаем вместе

Summary по блоку 3



Теория

Summary по блоку 3

- ▶ Ключевые слова - **Grouping sets, Rollup, Cube**
- ▶ Подытог по аддитивным метриками можно посчитать в несколько этапов
- ▶ Подытог можно посчитать через **UNION ALL**
- ▶ Если поле не участвует в разрезе - по нему проставляется **NULL**

- ▶ С помощью **UNION ALL** можно создать тестовое или вспомогательное отношение. Чтобы проверить или размножить запрос.



Теория

Вопросы

Что осталось непонятым?

Фидбек

Что нового вы узнали на этом занятии?

Что показалось самым важным?

Что будете применять и в каких ситуациях?

Что хочется изучить подробнее?

Ссылка на Обратную связь.

Что делаем в следующий раз?

- ▶ Порядковый номер строки в нужном разрезе.
- ▶ Находить следующее значение в том же разрезе по заданному порядку
- ▶ Нарастающий итог без **Self Join**
- ▶ Считать агрегаты без **Group by**

Обязательно ознакомьтесь с преридом =)

Домашнее задание

Мягкий дедлайн: 10:00 BC

Жесткий дедлайн: 10:00 BT

ОБСУЖДЕНИЕ ДЗ

01.

Напишите запрос, который найдет все числа (вывести только уникальные), которые появляются в таблице последовательно (при сортировке по id) как минимум три раза подряд.

```
select distinct t1.num  
from sequence t1  
join sequence t2  
on t1.id = t2.id - 1  
    and t1.num = t2.num  
join sequence t3  
on t2.id = t3.id - 1  
    and t2.num = t3.num;
```

ОБСУЖДЕНИЕ ДЗ

02. Найти студентов, которые не решили правильно ни одной задачи по sql сложности 1

```
select s.student_id
from students_day4 s
where NOT EXISTS (
  select 1
  from registry_day4 rd
  join task_book_day4 tb
    on rd.subject = tb.subject and rd.task_no = tb.task_no and tb.difficulty = 1
  where s.student_id = rd.student_id
  and tb.subject = 'sql'
  and accepted
);
```

ОБСУЖДЕНИЕ ДЗ

03. Нужно найти наибольшую сумму транзакций, которые сделал юзер за 10 суток (max_sum_10_day).

```
select user_id, max(sum_10_day) max_sum_10_day
from (
    select tr.user_id,
           tr.dtime,
           (select sum(amount)
            from transactions_day6 prev
            where prev.user_id = tr.user_id
                  and prev.dtime > tr.dtime - interval 10 day and prev.dtime <= tr.dtime) sum_10_day
    from transactions_day6 tr
) t
group by 1;
```

ОБСУЖДЕНИЕ ДЗ

04. Найти сотрудников, которые находятся на 3-ем месте по размеру зарплаты.

```
select name, salary from employee
where salary = (
    select distinct salary from employee
    order by coalesce(salary, 0) desc
    limit 1 offset 2
);
```