Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

30 вопросов по SQL и проектированию баз данных из интервью с ведущими технологическими компаниями

КОНСУЛЬТАЦИИ ПО КАРЬЕРЕ В ОБЛАСТИ РАЗРАБОТКИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Когда вы слышите «ученый данных», вы думаете о моделировании, машинном обучении и других модных словечках. Хотя проектирование баз данных и SQL — не самые привлекательные аспекты работы

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google. Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

специалиста по данным, они являются *очень* важными темами, которые нужно освежить в памяти перед собеседованием по науке о данных.

Итак, вот **20 реальных вопросов по SQL** и **10 реальных вопросов по проектированию баз данных**, заданных ведущими компаниями, такими как Google, Amazon, Facebook, Databricks, Yelp и Robinhood, во время интервью по науке о данных. Мы решили 8 из приведенных ниже задач и поместили ответы на остальные в нашу книгу Ace the Data Science Interview (доступна на Amazon Prime!). Вы также можете БЕСПЛАТНО практиковать эти же вопросы SQL Interview на DataLemur ! Например, ниже приведен реальный вопрос интервью Facebook SQL:

Вопросы для интервью по SQL на DataLemur даже содержат подсказки!

Темы интервью по

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

науке о данных

Обзор вопросов для собеседования по SQL

Первый шаг аналитики для большинства рабочих процессов включает в себя быстрое разделение данных на срезы и нарезку данных в SQL. Вот почему способность эффективно писать базовые запросы является очень важным навыком. Хотя многие могут подумать, что SQL просто включает в себя SELECT и JOIN, есть много других операторов и деталей, связанных с мощными рабочими процессами SQL. Например, использование подзапросов важно и позволяет вам манипулировать подмножествами данных, с которыми могут выполняться последующие операции, в то время как оконные функции позволяют вам вырезать данные без явного объединения строк с помощью GROUP BY. Вопросы, задаваемые в SQL, обычно весьма практичны для компании — такая компания, как Facebook, может задавать

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

различные вопросы об аналитике пользователей или приложений, тогда как компания, такая как Amazon, задает вопросы о продуктах и транзакциях.

Список наиболее важных операторов SQL, которые нужно просмотреть перед собеседованием, можно найти в моем окончательном руководстве по собеседованию по SQL для аналитиков и специалистов по данным .

Руководство на 6000 слов, подробно описывающее все, что я знаю о SQL!

Обзор вопросов по проектированию баз данных

Хотя не обязательно знать внутреннюю работу баз данных (что обычно больше ориентировано на разработку данных), это помогает иметь хорошее понимание основных концепций в базах данных и системах. Базы данных относятся не к конкретным, а к тому, как они работают на высоком уровне и какие проектные решения и компромиссы

принимаются во время построения и запросов. «Системы» — это широкий термин, но он относится к любому набору структур или инструментов, на которые опирается анализ больших объемов данных. Например, распространенной темой интервью является фреймворк MapReduce, который широко используется во многих компаниях для параллельной обработки больших наборов данных.

20 вопросов для собеседования по SQL Data Science

1. **[Robinhood - Easy]** Предположим, вам даны приведенные ниже таблицы для трейдов и пользователей. Напишите запрос, чтобы перечислить 3 города с наибольшим количеством выполненных заказов.

trades

column_name	type
order_id	integer
user_id	integer
symbol	string ("NFLX", "FB", etc.)
price	float
quantity	integer
side	string ("buy", "sell")
status	string ("complete", "cancelled")
timestamp	datetime

users

column_name	type
user_id	integer
city	string
email	string
signup_date	datetime

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

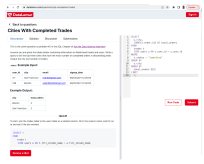
Основатель DataLemur и ранее инженер-программист в Facebook и Google. Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

Вы также можете попрактиковаться в этом точном вопросе интервью Robinhood SQL на DataLemur, если вам нужны несколько подсказок и возможность выполнить запрос решения!



2. **[Facebook — Easy]** Предположим, у вас есть приведенная ниже таблица событий в аналитике приложений. Напишите запрос, чтобы получить рейтинг кликов для приложения в 2019 году.

events

column_name	type
app_id	integer
event_id	string ("impression", "click")
timestamp	datetime

Вот похожая версия этого вопроса Facebook SQL на DataLemur вместе с несколькими подсказками!

3. **[Uber — Easy]** Предположим, вам предоставлена приведенная ниже таблица расходов по типам продуктов. Напишите запрос

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

для расчета совокупных расходов на каждый продукт с течением времени в хронологическом порядке.

total_trans

column_name	type
order_id	integer
user_id	integer
product_id	string
spend	float
date	datetime

4. **[Snapchat — Easy]** Предположим, у вас есть приведенные ниже таблицы сеансов пользователей и таблица пользователей. Напишите запрос, чтобы получить количество активных пользователей ежедневных когорт.

sessions

column_name	type
user_id	integer
session_id	integer
date	datetime

users + Add a view

column_name	type
user_id	integer
email	string
date	datetime

5. **[Facebook — Easy]** Предположим, вам даны приведенные ниже таблицы по пользователям и сообщениям пользователей. Напишите запрос, чтобы получить распределение количества сообщений на одного пользователя.

users

column_name	type
user_id	integer
date	datetime

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

posts

column_name	type
post_id	integer
user_id	string
body	string
date	datetime

6. **[Amazon — Easy]** Предположим, вам предоставлена приведенная ниже таблица покупок пользователей. Напишите запрос, чтобы получить количество людей, которые приобрели хотя бы один продукт в течение нескольких дней.

purchases

column_name	type
purchase_id	integer
user_id	integer
product_id	integer
quantity	integer
price	float
purchase_time	datetime

7. **[Opendoor - Easy]** Предположим, вам дана приведенная ниже таблица цен на жилье для различных почтовых индексов, которые были перечислены. Напишите запрос, чтобы получить 5 лучших почтовых индексов по рыночной доле цен на жилье для любого почтового индекса с не менее чем 10000 домов.

housing

column_name	type
house_id	integer
zip_code	integer
price	float
listing_date	datetime

8. **[Etsy - Easy]** Предположим, вам предоставлена приведенная ниже таблица транзакций пользователей для покупок. Напишите

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

запрос, чтобы получить список клиентов, самая ранняя покупка которых стоила не менее 50 долларов.

user_transactions

column_name	type
transaction_id	integer
product_id	integer
user_id	integer
spend	float
transaction_date	datetime

9. **[Disney - Easy]** Предположим, вам дана приведенная ниже таблица времени просмотра (в минутах) для всех пользователей, где каждый пользователь находится в определенном городе. Напишите запрос, возвращающий все пары городов, общее время просмотра которых отличается друг от друга на 10 000 минут.

watch_activity

column_name	type
user_id	integer
session_id	integer
watch_time	float
city_name	string
date	datetime

10. **[Twitter - Easy]** Предположим, вам предоставлена приведенная ниже таблица твитов каждого пользователя за определенный период времени. Рассчитайте 7-дневное скользящее среднее число твитов от каждого пользователя для каждой даты.

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

tweets + Add a view

column_name	type
tweet_id	integer
msg	string
user_id	integer
tweet_date	datetime

Подсказка: скользящие средние означают, что вам понадобится оконная функция! Если вам нужно освежить знания, ознакомьтесь с моей 30-дневной дорожной картой изучения SQL , в которой есть мой список любимых БЕСПЛАТНЫХ ресурсов SQL и то, как я буду их изучать, чтобы перейти от SQL Zero к SQL HERO!

11. **[Stitch Fix - Easy]** Предположим, вам дана приведенная ниже таблица транзакций от пользователей. Напишите запрос, чтобы получить количество пользователей и общее количество продуктов, купленных на дату последней транзакции, где каждый пользователь группируется по дате последней транзакции.

user_transactions

column_name	type
transaction_id	integer
product_id	integer
user_id	integer
spend	float
transaction_date	datetime

12. **[Amazon — Easy]** Предположим, вам предоставлена таблица ниже с

суммами расходов клиентов на продукты в различных категориях. Подсчитайте три самых покупаемых товара в каждой категории в 2020 году.

product_spend

column_name	type
transaction_id	integer
category_id	integer
product_id	integer
user_id	integer
spend	float
transaction_date	datetime

13. **[DoorDash — Easy]** Предположим, вам предоставлена приведенная ниже таблица транзакций по местам доставки и времени доставки еды — место начала, место окончания и отметка времени для заданного идентификатора еды. Определенные местоположения являются местами агрегации — куда отправляются блюда и куда они затем отправляются в конечный пункт назначения. Рассчитайте время доставки одного блюда в каждый конечный пункт назначения из определенного места агрегации, loc_id = 4.

delivery_times

column_name	type
meal_id	integer
start_loc_id	integer
end_loc_id	integer
cost	float
timestamp	timestamp

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google. Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

14. **[Facebook — Medium]** Предположим, у вас есть приведенные ниже таблицы действий пользователей. Напишите запрос, чтобы получить активное удержание пользователей по месяцам.

user_actions

column_name	type
user_id	integer
event_id	string ("sign-in", "like", "comment)
timestamp	datetime

15. **[Twitter - Medium]** Предположим, вам даны приведенные ниже таблицы для сеансовой активности пользователей. Напишите запрос, чтобы ранжировать пользователей по общей продолжительности сеанса для различных типов сеансов, которые у них были между датой начала (01.01.2020) и датой окончания (01.02.2020).

sessions

column_name	type
session_id	integer
user_id	integer
session_type	string
duration	integer
start_time	datetime

16. **[Snapchat — Medium]** Предположим, вам предоставлены приведенные ниже таблицы о пользователях и их времени, потраченном на отправку и открытие снимков. Напишите запрос,

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

чтобы получить разбивку для каждой возрастной разбивки по проценту времени, затрачиваемого на отправку и открытие снимков.

activities

column_name	type
activity_id	integer
user_id	integer
type	string ('send', 'open')
time_spent	float
activity_date	datetime

age_breakdown

column_name	type
user_id	integer
age_bucket	string

17. **[Google – Medium]** Предположим, вам предоставлена приведенная ниже таблица сеансов пользователей с заданным временем начала и окончания. Сеанс считается одновременным с другим сеансом, если они перекрываются во времени начала и окончания. Напишите запрос для вывода сеанса, который совпадает с наибольшим числом других сеансов.

sessions

column_name	type
session_id	integer
start_time	datetime
end_time	datetime

18. **[Yelp - Medium]** Предположим, вам предоставлена приведенная ниже таблица отзывов пользователей. Определите место с самым высоким рейтингом как компанию, чьи отзывы состоят только из 4 или 5

звезд. Напишите запрос, чтобы получить количество и процент предприятий с самым высоким рейтингом.

reviews

column_name	type
business_id	integer
user_id	integer
review_text	string
review_stars	integer
review_date	datetime

19. **[Google — Medium]** Предположим, вам предоставлена приведенная ниже таблица значений измерений датчика за несколько дней. Каждое измерение может происходить несколько раз в день. Напишите запрос для вывода суммы значений для каждого нечетного измерения и суммы значений для каждого четного измерения по дате.

measurements

column_name	type
measurement_id	integer
measurement_value	float
measurement_time	datetime

Если вам нужны подсказки, а также подробное решение, решите этот вопрос для интервью по Google SQL на DataLemur!

В DataLemur есть вопросы для интервью по SQL с решениями!

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

20. **[Etsy — Medium]** Предположим, вам предоставлена приведенная ниже таблица транзакций из различных результатов поиска продуктов от пользователей на Etsy. Для каждого заданного ключевого слова продукта существует несколько позиций, которые проходят A/B-тестирование, и собираются отзывы пользователей о релевантности результатов (от 1 до 5). Для каждой позиции каждого продукта существует множество отображений, каждое из которых фиксируется идентификатором display_id. Определить высокорелевантный дисплей как тот, при котором соответствующий показатель релевантности равен не менее 4. Напишите запрос, чтобы получить все продукты, имеющие хотя бы одну позицию с > 80% высокорелевантных дисплеев.

product_searches

column_name	type
product	string
position	integer
display_id	integer
relevance	integer
submit_time	datetime

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

10 вопросов на собеседовании по проектированию баз данных и систем

21. **[MongoDB - Easy]** Для каждого свойства ACID дайте описание каждого свойства одним предложением и почему эти свойства важны?
22. **[VMWare - Easy]** Каковы три основных этапа проектирования базы данных? Опишите каждый шаг.
23. **[Microsoft - Easy]** Каковы требования к первичному ключу?
24. **[DataStax — Easy]** Дерево B+ может содержать не более 5 указателей в узле. Какое минимальное количество ключей в листьях?
25. **[Databricks — Easy]** Опишите MapReduce и задействованные операции.
26. **[Microsoft - Easy]** Назовите одно основное сходство и различие между предложением WHERE и предложением HAVING в SQL (это важно знать, если вы хотите ответить на этот вопрос

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

интервью eBay SQL
)

27. **[Facebook — Просто]** Как триггер позволяет встроить бизнес-логику в базу данных?
28. **[DataStax - Easy]** Каковы шесть уровней безопасности базы данных и кратко объясните, что влечет за собой каждый из них?
29. **[Databricks — Easy]** Допустим, у вас есть оператор перетасовки, в котором входными данными является набор данных, а выходными данными — просто случайно упорядоченная версия этого набора данных. Опишите шаги алгоритма на английском языке.
30. **[Rubrik - Medium]** Опишите, что такое кеш и что такое блок. Скажем, у вас есть фиксированный объем общего хранилища данных — каковы некоторые компромиссы при изменении размера блока?

7 настоящих вопросов на собеседовании по Amazon SQL

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

Для дополнительной практики я собрал некоторые идеи о процессе собеседования по Amazon SQL для аналитиков данных и специалистов по данным, а также курировал 7 реальных вопросов для собеседования по Amazon SQL в блоге ниже:

Практикуйте эти настоящие вопросы из интервью Amazon SQL, чтобы получить работу своей мечты в Amazon, где вы будете сокрушить местные малые предприятия по одной аналитической информации, основанной на данных!

SQL и решения для интервью с базами данных

Проблема №4 Решение:

По определению, ежедневные когорты — это активные пользователи с

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

определенного дня. Во-первых, мы можем использовать подзапрос для получения сеансов новых пользователей по дням, используя внутреннее соединение с пользователями. Это необходимо для фильтрации только активных пользователей по определенной дате присоединения к когорте. Затем мы можем получить отдельный счетчик, чтобы вернуть количество активных пользователей:

```
WITH new_users_by_date AS
    SELECT sessions.*
    FROM sessions
  JOIN users on
    sessions.user_id = us
    sessions.date = users
)
SELECT date, COUNT(DISTIN
    FROM new_users_by_dat
    GROUP BY 1 ORDER BY 1
```

Проблема №8 Решение:

Хотя мы могли бы использовать самообъединение на transaction_date = MIN(transaction_date) для каждого пользователя, мы также можем использовать оконную функцию RANK(), чтобы получить порядок покупок по покупателю, а затем использовать этот подзапрос для фильтрации клиентов, у которых первая покупка

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google. Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

(ранг один) составляет не менее 50 долларов. Обратите внимание, что для этого требуется, чтобы подзапрос также включал расходы.

```
WITH purchase_rank AS (
    SELECT user_id, spend
        RANK() OVER
            (PARTITION BY
        FROM user_transac
)

SELECT
    user_id
FROM
    purchase_rank
WHERE rank = 1 AND spend
```

Задача №11 Решение:

Во-первых, нам нужно получить последнюю дату транзакции для каждого пользователя, а также количество продуктов, которые они приобрели. Это можно сделать в подзапросе, где мы GROUP BY user_id и берем COUNT (DISTINCT product_id), чтобы получить количество продуктов, которые они приобрели, и MAX (transaction_date), чтобы получить дату последней транзакции (при приведении к дате). Затем, используя этот подзапрос, мы можем просто выполнить агрегацию по столбцу даты транзакции в предыдущем подзапросе, выполняя COUNT() для

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

количества пользователей и SUM() для количества продуктов:

```
WITH latest_date AS (
    SELECT user_id,
        COUNT(DISTINCT pr
        MAX(transaction_d
    FROM user_transaction
    GROUP BY )

SELECT curr_date,
    COUNT(user_id) AS num
    SUM(num_products) AS
FROM
latest_date
GROUP BY 1
```

Задача №16 Решение:

Мы можем получить разбивку общего времени, затраченного на каждое действие каждым пользователем, отфильтровав значение activity_type и взяв сумму затраченного времени. При этом мы хотим выполнить внешнее соединение с возрастным сегментом, чтобы получить общее время по возрастному сегменту для обоих типов активности. Это приводит к двум нижеприведенным подзапросам. Затем мы можем использовать эти два подзапроса, чтобы суммировать их, объединив соответствующие сегменты возраста и взяв долю времени отправки и долю времени открытия

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

для каждого сегмента возраста:

```
WITH send_timespent AS (
    SELECT age_breakdown.
    FROM age_breakdown
  LEFT JOIN activities on
    WHERE activities.type
    GROUP BY 1
),
open_timespent AS (
    SELECT age_breakdown.
    FROM age_breakdown
  LEFT JOIN activities on
    WHERE activities.type
    GROUP BY 1
),

SELECT a.age_bucket,
    s.send_timespent / (s
    o.open_timespent / (s
FROM age_breakdown a
LEFT JOIN send_timespent
LEFT JOIN open_timespent
GROUP BY 1
```

Задача №18 Решение:

Во-первых, нам нужно получить места, где отзывы все 4 или 5 звезд. Мы можем сделать это, используя предложение HAVING вместо предложения WHERE, поскольку все отзывы должны быть 4 звезды или выше. Для условия HAVING мы можем использовать оператор CASE, который фильтрует 4 или 5 звезд, а затем СУММУ по ним. Затем его можно сравнить с общим количеством строк конкретных отзывов business_id, чтобы убедиться, что количество лучших

отзывов совпадает с общим количеством отзывов. С соответствующими предприятиями мы можем затем выполнить внешнее соединение с исходной таблицей для business_id, чтобы получить COUNT различных совпадений business_id, а затем процент, сравнив COUNT из первых мест с общим COUNT для business_id:

```
WITH top_places AS (
    SELECT business_id
    FROM user_transaction
    GROUP BY 1
    HAVING
        SUM(CASE WHEN rat
    )

SELECT
    COUNT(DISTINCT t.busi
    COUNT(DISTINCT t.busi
FROM reviews r
LEFT JOIN top_places t
    ON r.business_id = t.
```

Задача №21 Решение:

ACID — это набор свойств, который гарантирует, что даже в случае ошибок, отключений электроэнергии и других непредвиденных обстоятельств база данных все равно будет работать. Это важная основа для изучения систем баз данных.

A: Атомарность, означающая, что вся

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

транзакция происходит целиком или не происходит вообще (никаких частичных транзакций). Это предотвращает частичные обновления, которые могут быть проблематичными. Следовательно, транзакции не могут быть «в процессе» для любого пользователя.

C: Непротиворечивость, означающая, что существуют ограничения целостности, так что база данных непротиворечива до и после данной транзакции. По сути, если я ищу то, что находится в строке 3, а затем делаю это снова без каких-либо изменений в базе данных (без удаления или вставки), я должен получить тот же результат. Любая ссылочная целостность обеспечивается соответствующими проверками первичных и внешних ключей.

I: Изоляция, означающая, что транзакции происходят изолированно, и поэтому несколько транзакций могут выполняться независимо друг от друга без помех. Это правильно поддерживает параллелизм.

D: Надежность, означающая, что после

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

завершения транзакции эта информация теперь обновляется в базе данных даже в случае сбоя системы.

Задача №25 Решение:

MapReduce — это фреймворк, который активно используется при обработке больших наборов данных в большом количестве кластеров (на многих машинах). Внутри групп машин есть рабочие узлы (которые выполняют вычисления) и главные узлы (которые делегируют задачи для каждого рабочего узла). Три шага, как правило, следующие.

1. Шаг карты: каждый рабочий узел применяет определенные операции карты к входным данным (которые главный узел гарантирует, что они не будут дублироваться) и записывает выходные данные в буфер памяти.
2. Данные перераспределяютс на основе выходных ключей из функции карты предыдущего шага, так что для любого заданного ключа он находится на одном и том же рабочем узле.
3. Каждый рабочий узел обрабатывает каждый ключ

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

параллельно, используя определенные операции сокращения для получения выходного результата.

Поскольку функции отображения и сокращения могут выполняться параллельно, объем выполняемой обработки ограничивается только объемом доступных вычислений и данных. Обратите внимание, что существуют пограничные случаи при сбоях рабочих узлов — в этих случаях желаемые операции могут быть перепланированы главным узлом.

Задача №27 Решение:

Триггер похож на условие CHECK, но каждый раз, когда происходит обновление базы данных, условие триггера будет проверяться, чтобы увидеть, не было ли оно нарушено. Это позволяет вам реализовать некоторый уровень контроля и гарантии того, что все ваши записи данных соответствуют определенному условию. Например, триггер, указывающий, что все значения ID должны быть > 0, гарантирует, что вы не получите нулевых или отрицательных значений. Когда кто-то попытается

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

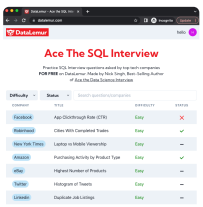
Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

ввести такое значение, запись не пройдет.

При этом есть причины, по которым не следует включать бизнес-логику в триггеры базы данных. Например: 1) введение побочных эффектов, которые приводят к ошибкам или другим непредвиденным последствиям, или 2) проблемы с производительностью, и в этом случае возникает каскадный эффект на триггеры, что приводит к блокировке и другим проблемам.

Как получить больше вопросов для интервью по науке о данных

Хотите больше подобного? Купите нашу 301-страничную книгу для подготовки к интервью по науке о данных на Amazon ! А если вам нужна интерактивная платформа SQL-интервью, DataLemur поможет вам.



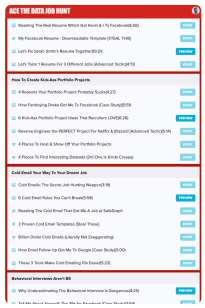
Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

Настоящие вопросы интервью по SQL Data Science на DataLemur!

Мы также выпустили бесплатный 9-дневный ускоренный курс собеседования , в котором представлены краткий обзор каждой главы из нашей книги и моего видеокурса Ace the Data Job Hunt !



25+ видеоуроков по Ace the Data Job Hunt

У нас также есть 40 реальных вопросов о вероятности и статистике, заданных FANG и Wall Street, и 30 вопросов о машинном обучении.

Прочтите 40 вопросов, заданных в интервью для специалистов Facebook, Microsoft, Two Sigma и Bloomberg.

Присоединяйтесь к 30 000+ подписчиков в 38 странах.

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ