База данных — это упорядоченный набор структурированной информации или данных, которые обычно хранятся в электронном виде в компьютерной системе. База данных обычно управляется системой управления базами данных.

Каждая строка олицетворяет уникальный элемент данных, который ею и описывается. Столбцы представляют собой отдельные куски информации, которые известны о данном элементе. Строки обычно называют записями, а столбцы - полями.

Язык структурированных запросов (SQL) – это язык программирования для хранения и обработки информации в реляционной базе данных. Реляционная база данных хранит информацию в табличной форме со строками и столбцами, представляющими различные атрибуты данных и различные связи между значениями данных.

В этом уроке мы разберем одну из самых часто используемых операций SQL — SELECT . С помощью этой команды выполняют выборку данных из таблиц.

Первичный ключ — это поле или набор полей со значениями, которые являются уникальными для всей таблицы. Значения ключа могут использоваться для обозначения всех записей, при этом каждая запись имеет отдельное значение ключа. Каждая таблица может содержать только один первичный ключ.

Индекс (index) – средство ускорения операции поиска записей в таблице, а также выполнения других операций, использующих поиск: извлечение, модификация, сортировка и т. д. Индексный файл (index file) – это файл, в котором хранится информация индекса.

В контексте баз данных транзакцией называют любую операцию, которую база данных обрабатывает как единую единицу работы. Чтобы база данных оставалась согласованной, транзакция должна быть полностью завершена.

Примеры:

**Банковские системы**: Базы данных хранят информацию о счетах клиентов, транзакциях и кредитной истории.

**Интернет-магазины**: Они используют базы данных для управления товарами, заказами и данными клиентов.

 Большие данные — это разнообразные данные, поступающие с более высокой скоростью, объем которых постоянно растет. Таким образом, три основных свойства больших данных — это разнообразие, высокая скорость поступления и большой объем.

Зачем большие компании собирают и анализируют большие данные?

Во-первых, он позволяет компаниям лучше понять своих клиентов и рынок в целом и создавать продукты и услуги, которые нужны потребителям. Во-вторых, анализ данных помогает принять обоснованные бизнес-решения на основе фактов и цифр, что помогает снизить риски и увеличить прибыль.

Netflix анализирует огромные объемы данных о просмотре своих пользователей, чтобы:

1. **Рекомендовать контент**: Система предлагает фильмы и сериалы, основываясь на предыдущих просмотрах и предпочтениях зрителей.
2. **Создавать оригинальный контент**: Анализируя популярные жанры и предпочтения аудитории, Netflix принимает решения о производстве новых фильмов и сериалов.
3. **Оптимизировать качество потоковой передачи**: Компания использует данные о том, как и когда пользователи смотрят контент, для улучшения качества сервиса.

Для реализации этих этапов применяются специальные технологии и инструменты: Hadoop, Spark, Hive и Pig, а также параллельные вычисления, распределенные вычисления, потоковая обработка, облачные вычисления и другие технологии.

Работа с большими данными сопряжена с несколькими сложностями:

1. **Объем**: Большие данные могут достигать терабайтов и петабайтов информации, что требует мощных вычислительных ресурсов для хранения и обработки.
2. **Скорость**: Данные поступают в режиме реального времени, и их нужно быстро обрабатывать, что может создавать сложности в системах обработки.
3. **Разнообразие**: Данные могут быть структурированными (таблицы), полуструктурированными (JSON, XML) и неструктурированными (текст, изображения). Это требует гибких инструментов для их обработки.
4. **Качество данных**: Данные могут содержать ошибки, пропуски или быть неполными. Обеспечение их качества требует дополнительных усилий.
5. **Безопасность и конфиденциальность**: Работа с большими данными требует защиты личной информации и соблюдения законодательства, что может быть сложно.
6. **Аналитика и интерпретация**: Извлечение полезной информации из больших объемов данных требует применения сложных аналитических методов и инструментов.
7. **Инфраструктура**: Необходима надежная и масштабируемая инфраструктура для хранения, обработки и анализа данных, что может быть дорогостоящим.

Большие данные (Big Data) отличаются от обычных данных несколькими ключевыми аспектами:

1. **Объем**: Большие данные представляют собой огромные объемы информации, которые не могут быть эффективно обработаны традиционными методами или инструментами. Обычные данные обычно имеют меньший объем и проще в управлении.
2. **Скорость**: Большие данные часто поступают в режиме реального времени, что требует быстрой обработки и анализа. Обычные данные могут обрабатываться с меньшей срочностью.
3. **Разнообразие**: Большие данные включают различные типы данных — структурированные, полуструктурированные и неструктурированные. Обычные данные чаще всего имеют однородный формат, например, таблицы.
4. **Сложность анализа**: Анализ больших данных требует более сложных методов и инструментов, таких как машинное обучение и аналитика, в то время как обычные данные могут анализироваться с использованием простых статистических методов.
5. **Актуальность**: Большие данные часто включают данные из множества источников (социальные сети, датчики, интернет вещей и т.д.), что делает их более актуальными для принятия решений. Обычные данные обычно ограничиваются одним или несколькими источниками.
6. **Инфраструктура**: Для работы с большими данными необходима специализированная инфраструктура (например, распределенные системы хранения и обработки), тогда как для обычных данных можно использовать стандартные базы данных и программное обеспечение.

Большие данные нужны в маркетинге, перевозках, автомобилестроении, здравоохранении, науке, сельском хозяйстве и других сферах, в которых можно собрать и обработать нужные массивы информации.

Обработка больших данных может значительно улучшить услуги компаний несколькими способами:

1. **Персонализация**: Анализ данных о предпочтениях и поведении клиентов позволяет создавать индивидуализированные предложения и рекомендации, что повышает удовлетворенность пользователей.
2. **Оптимизация процессов**: Большие данные помогают выявлять узкие места в бизнес-процессах, позволяя компаниям оптимизировать операции и снижать затраты.
3. **Прогнозирование спроса**: Анализ исторических данных позволяет компаниям предсказывать будущие тенденции и спрос, что помогает лучше планировать запасы и ресурсы.
4. **Улучшение качества обслуживания**: Мониторинг отзывов и взаимодействий с клиентами позволяет компаниям быстро реагировать на проблемы и улучшать качество обслуживания.
5. **Анализ конкурентной среды**: Сравнительный анализ данных о конкурентам помогает выявить сильные и слабые стороны, а также адаптировать стратегии.
6. **Разработка новых продуктов**: Анализ потребительских трендов и предпочтений помогает компаниям создавать продукты, которые более точно соответствуют запросам рынка.
7. **Управление рисками**: Обработка больших данных позволяет компаниям выявлять потенциальные риски и принимать проактивные меры для их минимизации.