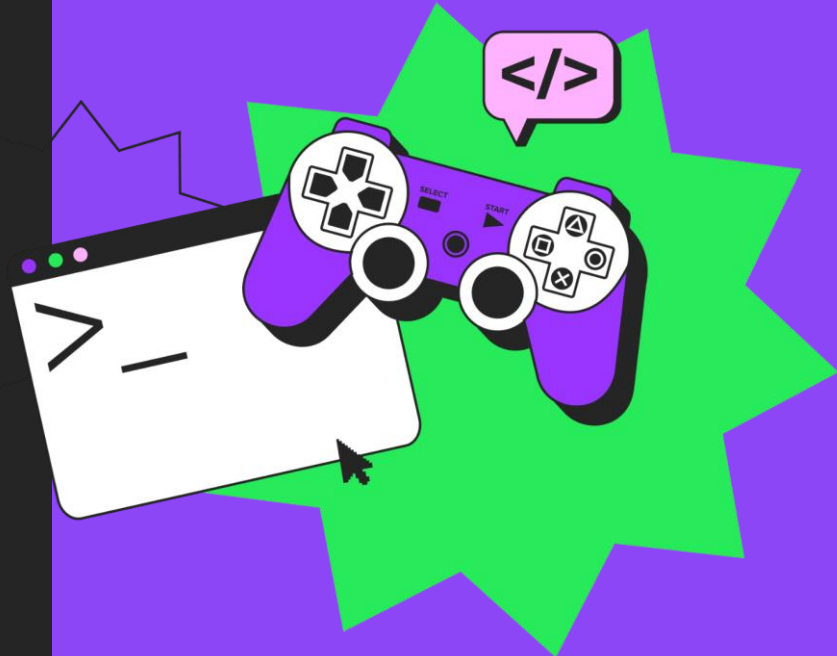


Модели данных и нормализация таблиц. Схема "звезда".

Урок 1





Знакомство и содержание урока



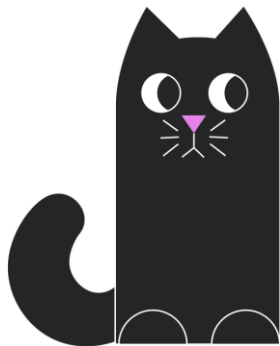
Антон Иоффе

Senior Data Scientist, SAP

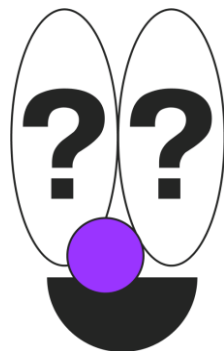
Последние 5 лет работаю в Data Science. Начинал как backend разработчик. Живу в Израиле.

- ✧ Система анонимизации видеофайлов, поиск поддельных чеков и счетов, автоматический аудит командировочных расходов и т.д.
- ✧ Запатентовано 7 различных алгоритмов связанных с машинным обучением

**Ответьте на несколько вопросов
сообщением в чат**



Из какого вы города?



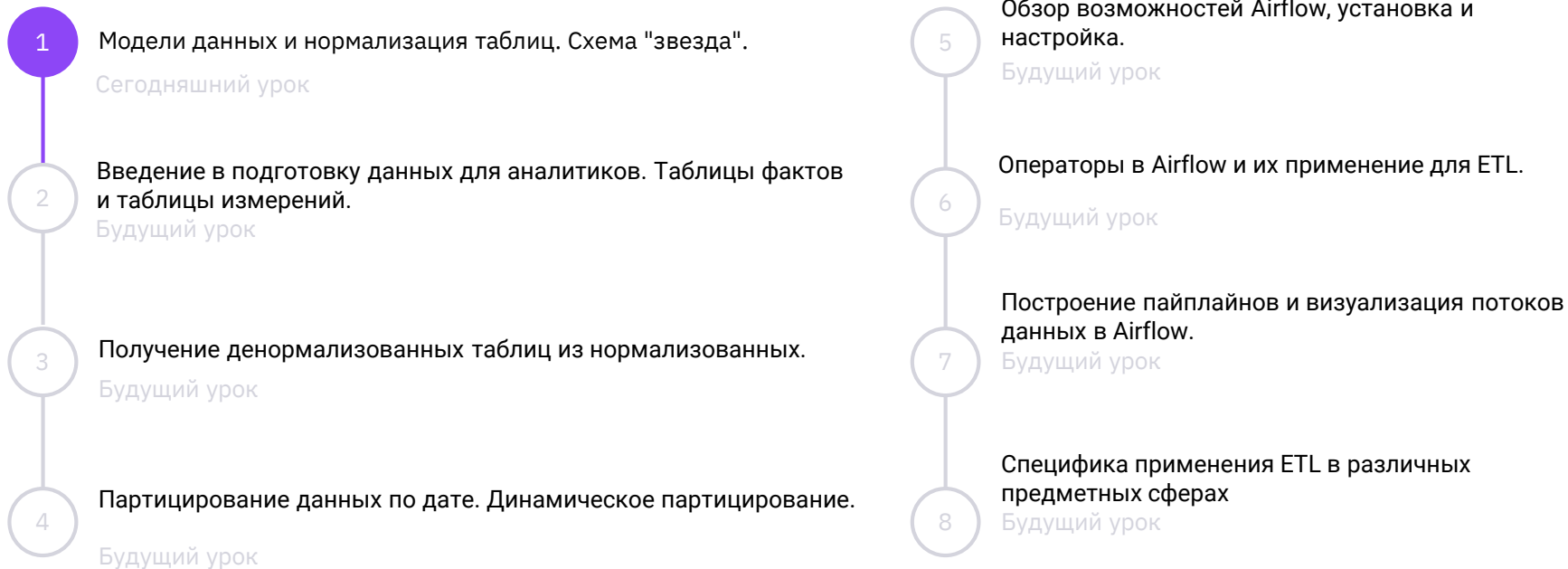
Сколько вам лет?



**Кем вы работаете сейчас?
Как долго?**








План курса (вертикальный)





Что будет на уроке сегодня

-  Основные функции ETL-систем
-  Как работает ETL-система
-  Модели данных
-  Нормализация таблиц
-  Схема звезда



Викторина



Что такое ETL?

1. Термин для процессов, которые происходят, когда данные переносят из нескольких систем в одно хранилище
2. Широко распространенный протокол передачи данных
3. База данных



Что такое ETL?

1. Термин для процессов, которые происходят, когда данные переносят из нескольких систем в одно хранилище
2. Широко распространенный протокол передачи данных
3. База данных



Какие процессы являются основными в ETL?

1. Чтение, добавление, удаление
2. Вставка, Трансформация, Сложение
3. Извлечение, трансформация, загрузка



Какие процессы являются основными в ETL?

1. Чтение, добавление, удаление
2. Вставка, Трансформация, Сложение
3. Извлечение, трансформация, загрузка



Какие модели данных существуют?

1. Структурная, нормальная, древовидная
2. Иерархическая, сетевая, реляционная
3. Статическая и динамическая



Какие модели данных существуют?

1. Структурная, нормальная, древовидная
2. Иерархическая, сетевая, реляционная
3. Статическая и динамическая



Для чего нужна нормализация таблиц?

1. Ликвидация избыточности
2. Ликвидация противоречий
3. Осуществление корректного редактирования и обработки данных
4. Все варианты верны



Для чего нужна нормализация таблиц?

1. Ликвидация избыточности
2. Ликвидация противоречий
3. Осуществление корректного редактирования и обработки данных
4. Все варианты верны



Какая нормальная форма считается достаточной?

1. Третья нормальная форма
2. Шестая нормальная форма
3. Нормальная форма Бойса-Кода



Какая нормальная форма считается достаточной?

1. Третья нормальная форма
2. Шестая нормальная форма
3. Нормальная форма Бойса-Кода



Должны ли быть нормализованы таблицы в схеме звезда?

1. Да должны
2. Нет не должны
3. Таблицы измерений не должны, а таблица фактов должна
4. Таблица фактов не должна, а таблицы измерений должны



Должны ли быть нормализованы таблицы в схеме звезда?

1. Да должны
2. Нет не должны
3. Таблицы измерений не должны, а таблица фактов должна
4. Таблица фактов не должна, а таблицы измерений должны



Вопросы?

Вопросы?



Вопросы?





Практика

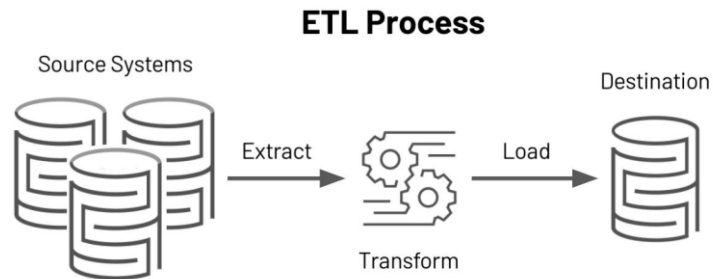


ETL

Как построить структуру ETL процесса?

Напомню что ETL состоит из трех основных этапов:

- **извлечение** данных из различных источников
- **трансформация** и очистка данных для приведения их к единообразию или в соответствие с бизнес-задачами.
- **загрузка** в хранилище данных





Задание 1

Распределите источники или действия по отношению к одному из трех процессов входящих в состав ETL

1. MySQL DB
2. Хранилище данных
3. CSV files
4. Удаление дубликатов
5. Нормализация данных
6. Маппинг
7. Oracle DB

Нарисуйте структурную схему ETL для этих процессов используя app.diagrams.net и поделитесь картинкой в чате



10 минут

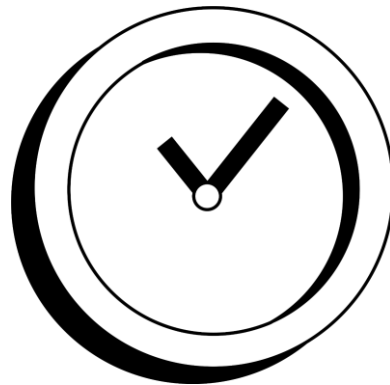


Задание 1

<<10:00->>

Распределите источники или действия по отношению к одному из трех процессов входящих в состав ETL

1. MySQL DB
2. Хранилище данных
3. CSV files
4. Удаление дубликатов
5. Нормализация данных
6. Маппинг
7. Oracle DB



Нарисуйте структурную схему ETL для этих процессов используя app.diagrams.net и поделитесь картинкой в чате



Задание 2

Разработайте процесс ETL для системы магазин-склад. Предполагается что у магазина есть база данных транзакций(продаж, возвратов). У склада есть база данных товаров (поступивших, отгруженных). Конечное хранилище должно давать возможность аналитического анализа популярных товаров, и товаров которые нужно заказать на склад.

Нарисуйте структурную схему ETL используя app.diagrams.net и поделитесь картинкой в чате. Внизу каждого этапа ETL сделайте короткое описание, что будет происходить на этом этапе.



15 минут

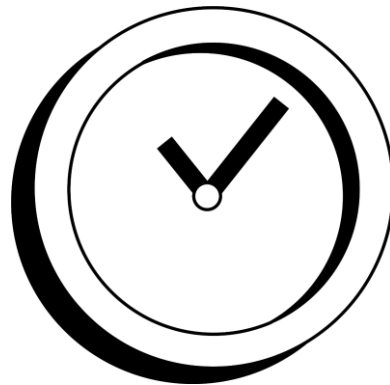


Задание 2

<<15:00->>

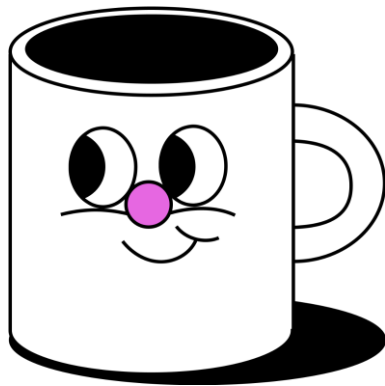
Разработайте процесс ETL для системы магазин-склад. Предполагается что у магазина есть база данных транзакций(продаж, возвратов). У склада есть база данных товаров (поступивших, отгруженных). Конечное хранилище должно давать возможность аналитического анализа популярных товаров, и товаров которые нужно заказать на склад.

Нарисуйте структурную схему ETL используя app.diagrams.net и поделитесь картинкой в чате. Внизу каждого этапа ETL сделайте короткое описание, что будет происходить на этом этапе.





Перерыв



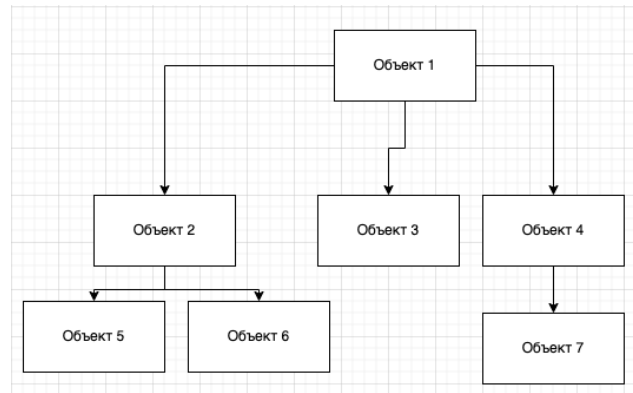
<<5:00->>



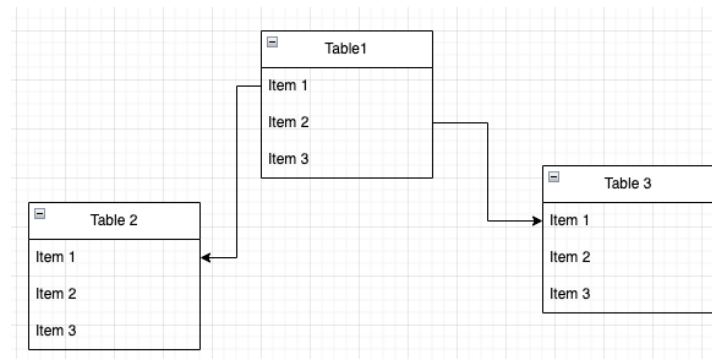
Иерархическая и реляционная модели

Иерархическая модель представляет собой совокупность элементов, расположенных в порядке их подчинения от общего к частному и образующих перевернутое по структуре дерево (граф).

Реляционная модель данных объекты и связи между ними представляет в виде таблиц, при этом связи тоже рассматриваются как объекты. Все строки, составляющие таблицу в реляционной базе данных, должны иметь первичный ключ. Все современные средства СУБД поддерживают реляционную модель данных.



Пример иерархической модели



Пример реляционной модели



Задание 3

Постройте иерархическую и реляционную модели описывающие структуру предприятия состоящие из объектов
Отдел, Начальник, Сотрудник

Нарисуйте схему моделей используя app.diagrams.net и поделитесь картинкой в чате.



10 минут

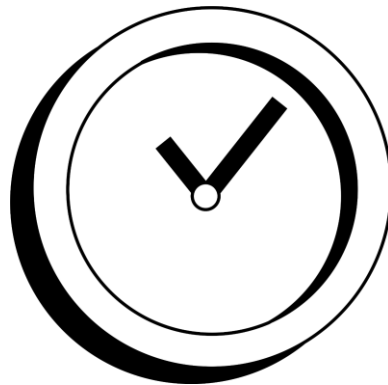


Задание 3

Постройте иерархическую и реляционную модели описывающие структуру предприятия состоящие из объектов
Отдел, Начальник, Сотрудник

Нарисуйте схему моделей используя app.diagrams.net и поделитесь картинкой в чате.

<<10:00->>





Нормализация таблиц

Нормализация представляет собой процесс реорганизации данных путем ликвидации избыточности данных и иных противоречий с целью приведения таблиц к виду, позволяющему осуществлять непротиворечивое и корректное редактирование данных.

Существует 7 нормальных форм:

- 1 НФ
- 2 НФ
- 3 НФ
- НФ Бойса-Кода
- 4 НФ
- 5 НФ
- 6 НФ

Каждая нормальная форма (за исключением первой) подразумевает, что к данным уже была применена предыдущая нормальная форма. База данных считается нормализованной, если к ней применяется третья нормальная форма и выше.





Задание 4

Код предмета	Предмет	Учитель	Код студента	Фамилия студента	Имя студента
П01	Проектировани БД	Моисеев	С01	Рогов	Василий
			С02	Бахмутов	Павел
			С03	Васильев	Лев
П02	Машинное обучение	Щербань	С02	Бахмутов	Павел
			С03	Васильев	Лев

Приведите таблицу в 1 НФ. Результатом поделитесь в чате

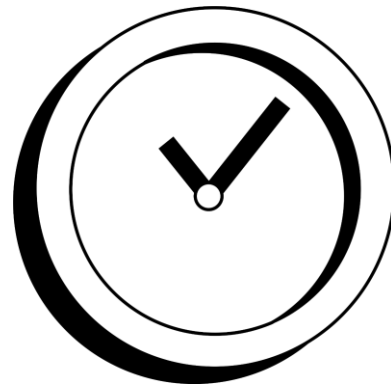
5 минут



Задание 4

<<05:00->>

Код предмета	Предмет	Учитель	Код студента	Фамилия студента	Имя студента
П01	Проектировани БД	Моисеев	С01	Рогов	Василий
			С02	Бахмутов	Павел
			С03	Васильев	Лев
П02	Машинное обучение	Щербань	С02	Бахмутов	Павел
			С03	Васильев	Лев



Приведите таблицу в 1 НФ. Результатом поделитесь в чате



Задание 5

Код предмета	Предмет	Учитель	Код студента	Фамилия студента	Имя студента
П01	Проектировани БД	Моисеев	C01	Рогов	Василий
			C02	Бахмутов	Павел
			C03	Васильев	Лев
П02	Машинное обучение	Щербань	C02	Бахмутов	Павел
			C03	Васильев	Лев

Приведите таблицу в 2 НФ. Используя таблицу полученную в предыдущем задании. Результатом поделитесь в чате



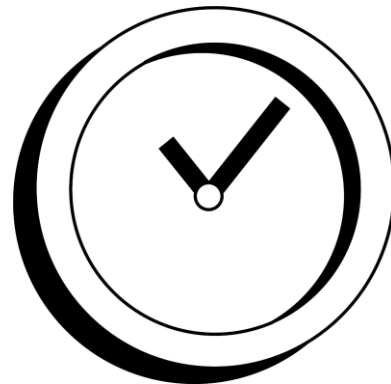
5 минут



<<05:00->>

Задание 5

Код предмета	Предмет	Учитель	Код студента	Фамилия студента	Имя студента
П01	Проектировани БД	Моисеев	С01	Рогов	Василий
			С02	Бахмутов	Павел
			С03	Васильев	Лев
П02	Машинное обучение	Щербань	С02	Бахмутов	Павел
			С03	Васильев	Лев



Приведите таблицу в 2 НФ. Используя таблицу полученную в предыдущем задании. Результатом поделитесь в чате



Задание 6

Код студента	Фамилия студента	Имя студента	Дата рождения	Возраст	Последнее обновление
C01	Рогов	Василий	11.09.1990	32	20.11.2022
C02	Бахмутов	Павел	08.07.1995	27	03.08.2022
C03	Васильев	Лев	22.12.1997	24	01.04.2022

Приведите таблицу в 3 НФ. Результатом поделитесь в чате

5 минут

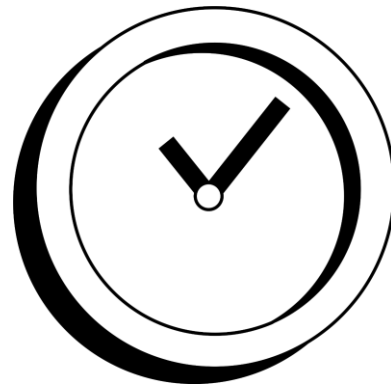


<<05:00->>

Задание 6

Код студента	Фамилия студента	Имя студента	Дата рождения	Возраст	Последнее обновление
C01	Рогов	Василий	11.09.1990	32	20.11.2022
C02	Бахмутов	Павел	08.07.1995	27	03.08.2022
C03	Васильев	Лев	22.12.1997	24	01.04.2022

Приведите таблицу в 3 НФ. Результатом поделитесь в чате





Задание 7

Постройте архитектуру хранилища данных для магазина, используя схему звезда. В хранилище должны быть представлены данные о покупатели, товаре, времени продажи. Нарисуйте структурную схему используя app.diagrams.net и поделитесь картинкой в чате



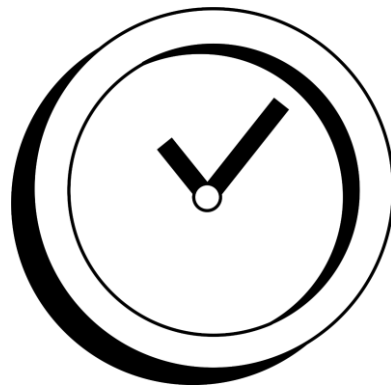
15 минут



Задание 7

Постройте архитектуру хранилища данных для магазина, используя схему звезда. В хранилище должны быть представлены данные о покупателе, товаре, времени продажи. Нарисуйте структурную схему используя app.diagrams.net и поделитесь картинкой в чате

<<15:00->>





Вопросы?

Вопросы?



Вопросы?





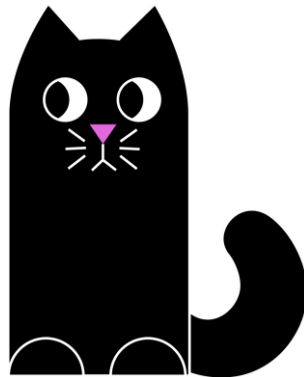
Домашнее задание



Домашнее задание

1. Нарисуйте архитектуру ETL процесса для сбора и анализа данных компанией которая хочет провести маркетинговую кампанию, используя app.diagrams.net. Сделайте описание почему вы считаете что архитектура должна выглядеть именно так.
2. Постройте реляционную и иерархическую модели данных для магазина который продает телефоны.
3. Определите в какой нормальной форме данная таблица, приведите ее ко 2 и 3 нормальной формам последовательно.

Employee_ID	Name	Job_Code	Job	City_code	Home_city
E001	Alice	J01	Chef	26	Moscow
E001	Alice	J02	Waiter	26	Moscow
E002	Bob	J02	Waiter	56	Perm
E002	Bob	J03	Bartender	56	Perm
E003	Alice	J01	Chef	56	Perm





Спасибо за внимание

