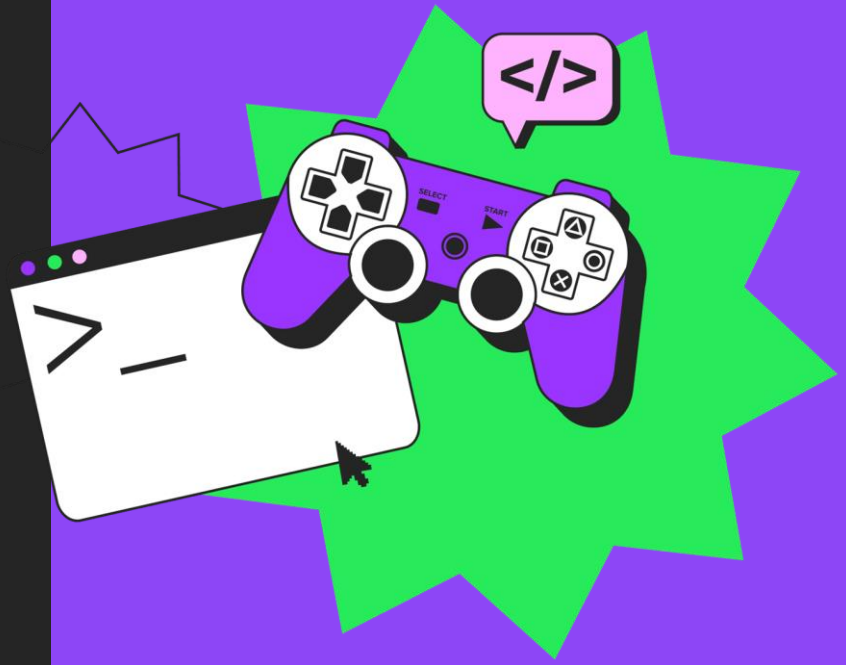
GeekBrains

Введение в подготовку данных для аналитиков. Таблицы фактов и таблицы измерений.

Урок 2



План курса (вертикальный)

1. Модели данных и нормализация таблиц. Схема "звезда".
   Прошедший урок
2. Введение в подготовку данных для аналитиков. Таблицы фактов и таблицы измерений.
   Сегодняшний урок
3. Получение денормализованных таблиц из нормализованных.
   Будущий урок
4. Партицирование данных.
   Будущий урок
5. Обзор возможностей Airflow, установка и настройка.
   Будущий урок
6. Операторы в Airflow и их применение для ETL.
   Будущий урок
7. Построение пайплайнов и визуализация потоков данных в Airflow.
   Будущий урок
8. Специфика применения ETL в различных предметных сферах
   Будущий урок

Что будет на уроке сегодня

- Тенденции в бизнес аналитике
- Понимание бизнеса и данных
- Подготовка и преобразование данных
- Исследование и визуализация данных
- Моделирование
- Оценка
- Развертывание
- Таблицы фактов и измерений

Викторина

Что такое BI?

1. Ключевые показатели эффективности
2. Бизнес аналитика
3. Индекс оценки бизнеса

Что такое BI?

1. Ключевые показатели эффективности
2. Бизнес аналитика
3. Индекс оценки бизнеса

Для чего нужна бизнес-аналитика?

1. Выявлять рыночные тенденции и повышать эффективность бизнеса
2. Установить критерии процессов внутри компании
3. Оба варианта верны

Для чего нужна бизнес-аналитика?

1. Выявлять рыночные тенденции и повышать эффективность бизнеса
2. Установить критерии процессов внутри компании
3. Оба варианта верны

Что входит в понятие анализ данных?

1. Извлечение, трансформация, загрузка
2. Извлечение, подготовка, моделирование

Что входит в понятие анализ данных?

1. Извлечение, трансформация, загрузка
2. Извлечение, подготовка, моделирование

Что такое сглаживание данных?

1. Процесс удаления избыточности
2. Процесс удаления шума из данных
3. Приведение данных к заданому диапазону
4. Все варианты верны

Что такое сглаживание данных?

1. Процесс удаления избыточности
2. Процесс удаления шума из данных
3. Приведение данных к заданому диапазону
4. Все варианты верны

Что такое нормализация данных?

1. Процесс удаления избыточности
2. Процесс удаления шума из данных
3. Приведение данных к заданому диапазону
4. Все варианты верны

Что такое сглаживание данных?

1. Процесс удаления избыточности
2. Процесс удаления шума из данных
3. Приведение данных к заданому диапазону
4. Все варианты верны

В какой таблице хранятся редко изменяеммые данные?

1. Таблица фактов
2. Таблица измерений
3. В обеих

В какой таблице хранятся редко изменяемые данные?

1. Таблица фактов
2. Таблица измерений
3. В обеих

Вопросы?

Вопросы?

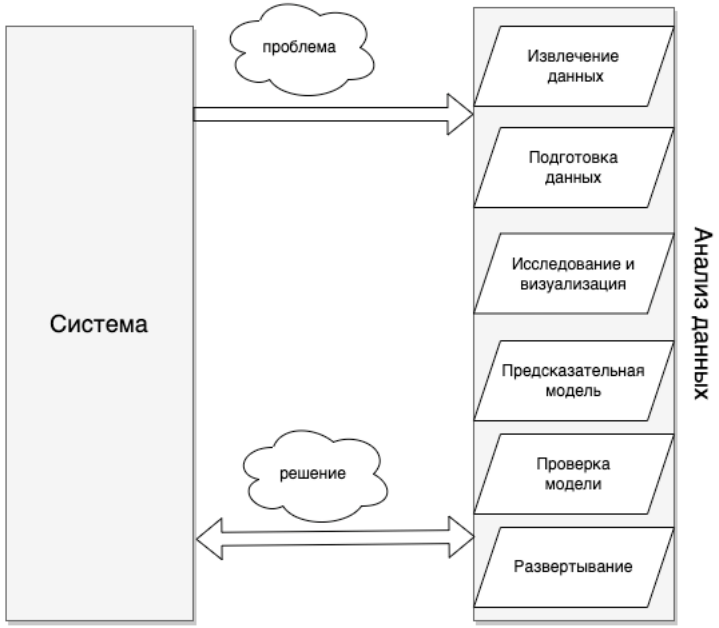
Вопросы?

Практика

Анализ данных

Анализ данных — это всего лишь последовательность шагов, каждый из которых играет ключевую роль для последующих. Этот процесс похож на цепь последовательных, связанных между собой этапов:

- Определение проблемы;
- Извлечение данных;
- Подготовка данных — очистка данных;
- Подготовка данных — преобразование данных;
- Исследование и визуализация данных;
- Моделирование;
- Оценка (проверка) модели;
- Развертывание — визуализация и интерпретация результатов;
- Развертывание — развертывание решения.



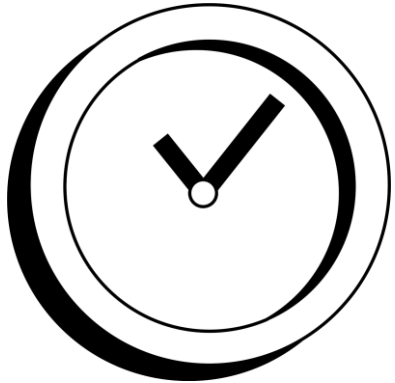
Задание 1

Скачайте датасет. Проанализируйте его на наличие пропусков используя apache spark.

Напишите в чат какие пропущенные значения вы обнаружили и причины их появления.

Задание 1

<<15:00->>



Скачайте датасет. Проанализируйте его на наличие пропусков и аномальных значений используя apach spark.

Напишите в чат какие пропущенные значения вы обнаружили и причины их появления. Привидите датасет к виду ниже. Подсказка: https://stackoverflow.com/questions/70537360/spark-dataframe-get-null-count-for-all-columns

col_name	null_count
children	0
days_employed	2120
dob_years	0
education	0
education_id	0
family_status	0
family_status_id	0
gender	0
income_type	0
debt	0
total_income	2120
purpose	0
purpose_category	6349
total_income2	0

Задание 2

Найдите в датафрейме дубликаты. И удалите их. Значения могут быть одинаковыми но написаны по разному. Например может отличаться размер регистра(заглавные и строчные буквы)

Расскажите о возможных причинах появления дубликатов.

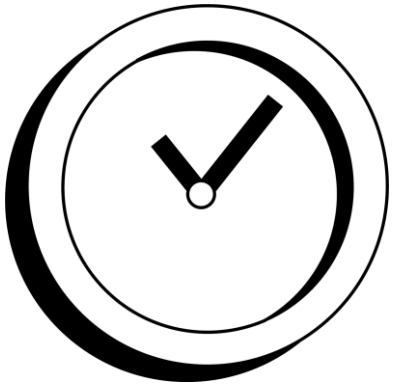
Привидите поля к соотвествующему типу данных.

Задание 2

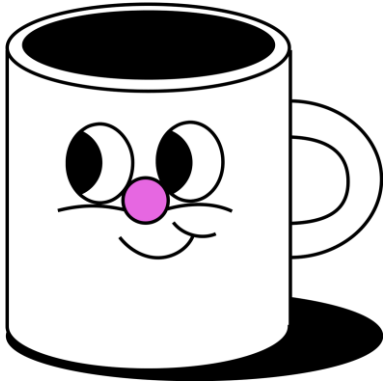
<<15:00->>

Найдите в датафрейме дубликаты. И удалите их. Значения могут быть одинаковыми но написаны по разному. Например может отличаться размер регистра(заглавные и строчные буквы)

Напишите в чат возможные причины появления дубликатов.



Перерыв



<<5:00->>

Задание 3

Сделайте колонку purpose_category в которую войдут следующие категории:

- операции с автомобилем,
- операции с недвижимостью,
- проведение свадьбы,
- получение образования

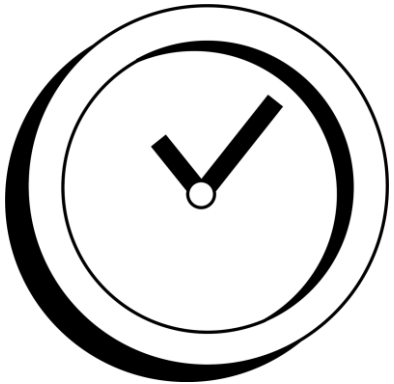
В чат напишите какое количество строк у вас получилось в каждой категории.

Задание 3

<<20:00->>

Сделайте колонку purpose_category в которую войдут следующие категории:

- операции с автомобилем,
- операции с недвижимостью,
- проведение свадьбы,
- получение образования

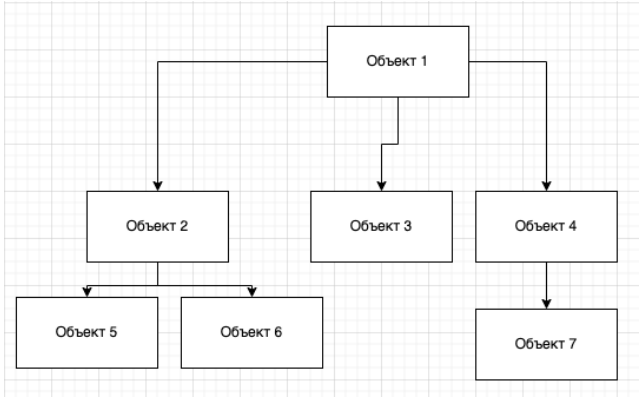


В чат напишите какое количество строк у вас получилось в каждой категории.

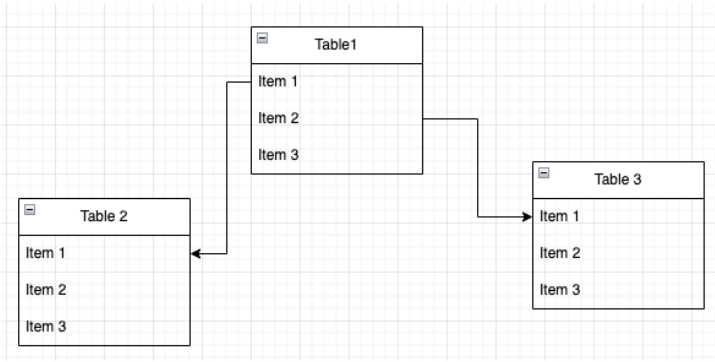
Иерархическая и реляционная модели

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 4

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.~~

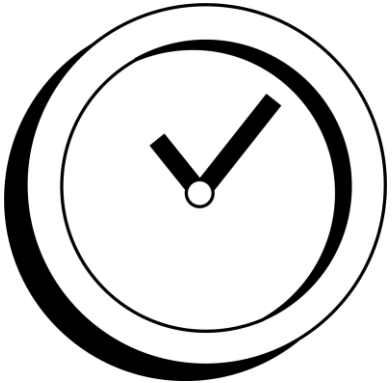
С помощью скрипта scala создайте таблицу с полями Отдел, Начальник, Сотрудник. Напишите функцию проверки существования таблицы в БД. Запишите данные в таблицу с помощью jdbc и apache spark.

Задание 4

<<15:00->>

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.~~



Вопросы?

Вопросы?

Вопросы?

Спасибо за внимание