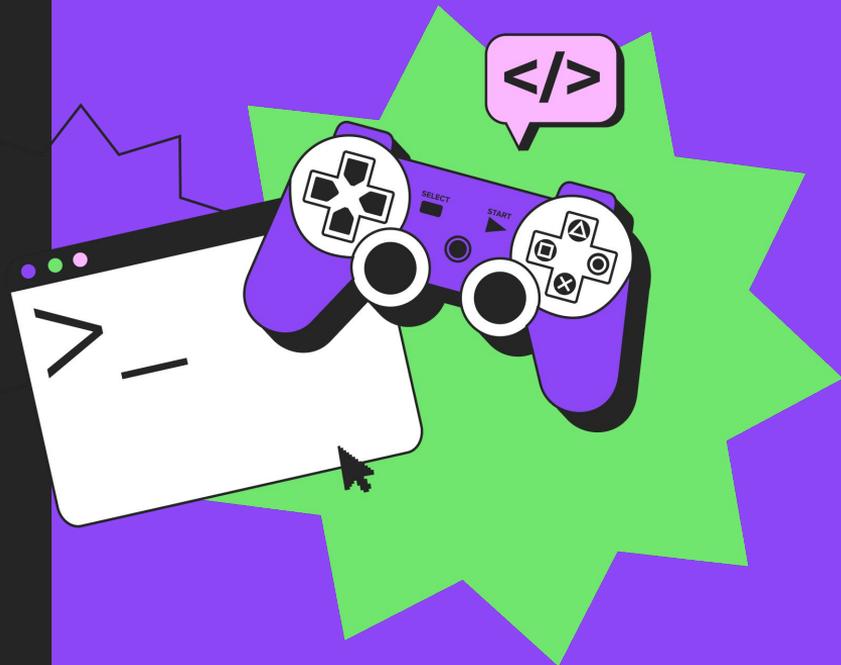


# Партицирование данных

Урок 4





## 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.  
Прошедший урок
- 3** Получение денормализованных таблиц из нормализованных.  
Прошедший урок
- 4** Партиционирование данных.  
Сегодняшний 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

-  Зачем нужно партицирование данных
-  Виды партицирования
-  Горизонтальное партицирование
-  Когда НЕ разбивать таблицу
-  Вертикальное партицирование
-  Функциональное партицирование
-  Преимущества партицирования
-  Недостатки партицирования



# Викторина



## Что такое ВІ?

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 заданному диапазону
4. Все варианты верны



## Что такое сглаживание данных?

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



## Что такое нормализация данных?

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



## Что такое сглаживание данных?

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



## **В какой таблице хранятся редко изменяемые 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?



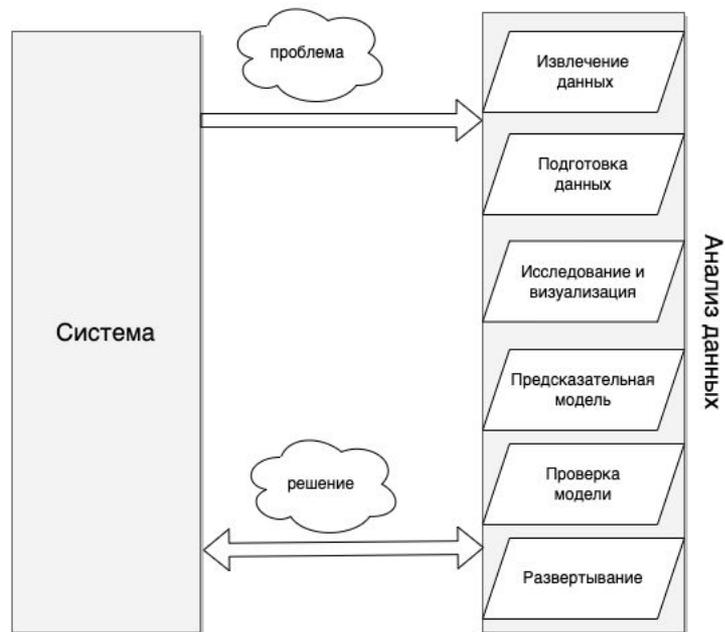


# Практика

## 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

1. Установить `pyspark` этой командой `cd\ & cd C:\Users\Alex\AppData\Local\Programs\Python\Python38 & python -m pip install pyspark==3.2.4`
2. Разобрать работу скрипта `s4.py`
3. Используя `pyspark` считать файл `s4.xlsx`. Сделать выборку по `"title" == "news"`. Добавить столбец с текущей меткой данных. Записать датасет в `mysql`.



**15 минут**



## Задание 2

1. Посмотреть структуру файла s2.xlsx
2. С помощью пандаса выполнить данный запрос:
3. Считать спарком файл с графиком платежей, с помощью оконных функций добавить поля с накопленных итогов по выплатам процентов и основного долга.
4. С помощью библиотеки matplotlib.pyplot построить графики по выплатам процентов и основного долга.

```
CREATE TABLE if not exists spark.`tasketl4b` (  
  `№` INT(10) NULL DEFAULT NULL,  
  `Месяц` DATE NULL DEFAULT NULL,  
  `Сумма платежа` FLOAT NULL DEFAULT NULL,  
  `Платеж по основному долгу` FLOAT NULL DEFAULT NULL,  
  `Платеж по процентам` FLOAT NULL DEFAULT NULL,  
  `Остаток долга` FLOAT NULL DEFAULT NULL,  
  `проценты` FLOAT NULL DEFAULT NULL,  
  `долг` FLOAT NULL DEFAULT NULL  
)  
COLLATE='utf8mb4_0900_ai_ci'  
ENGINE=InnoDB
```



## Задание 1

Создайте в Postgress таблицу news с полями id, category\_id, rate, title, author

Сделайте таблицы для партицирования по category\_id (возможные значения 1, 2, 3) которые будут наследоваться от основной таблицы

Создайте правила для добавления в эти таблицы

Добавьте несколько записей в каждую таблицу

Добавьте запись с category\_id = 4

Сделайте выборку из всех таблиц



**15 минут**



## Задание 1

Создайте в Postgress таблицу news с полями id, category\_id, rate, title, author

Сделайте таблицы для партицирования по category\_id (возможные значения 1, 2, 3) которые будут наследоваться от основной таблицы

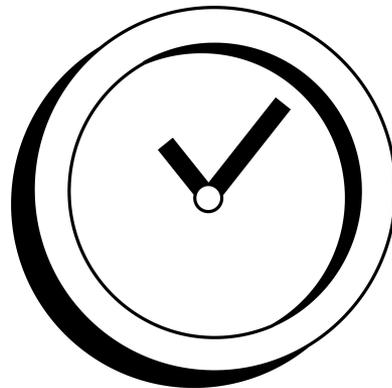
Создайте правила для добавления в эти таблицы

Добавьте несколько записей в каждую таблицу

Добавьте запись с category\_id = 4

Сделайте выборку из всех таблиц

<<15:00->>





## Задание 2

Сделайте таблицы для партицирования новостей по rate (возможные значения до 100, от 100 до 200, больше 200) которые будут наследоваться от основной таблицы

Создайте правила для добавления в эти таблицы

Добавьте несколько записей в каждую таблицу

Сделайте выборку из всех таблиц



**15 минут**



## Задание 2

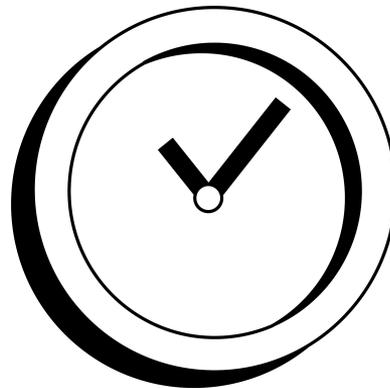
Сделайте таблицы для партицирования новостей по rate (возможные значения до 100, от 100 до 200, больше 200) которые будут наследоваться от основной таблицы

Создайте правила для добавления в эти таблицы

Добавьте несколько записей в каждую таблицу

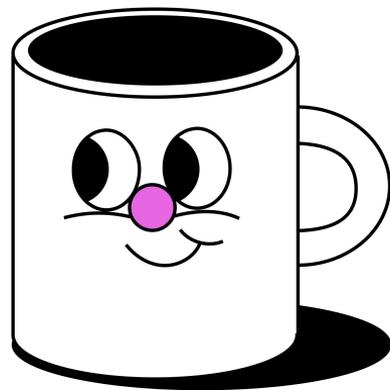
Сделайте выборку из всех таблиц

<<15:00->>





## Перерыв



<<5:00->>



## Задание 3

1. Откройте консоль Postgress
2. Создайте таблицу vehicles с полями vehicle\_type, plate\_number, year\_of\_issue, weight, owner
3. Сделайте таблицы для горизонтального партицирования по весу машины(от 1 тонны до 2.5 тонн, от 2.5 до 4 тонн, больше 4 тонн)
4. Сделайте таблицы для горизонтального партицирования по году выпуска машины (до 2000, с 2000 до 2019, после 2019)
5. Создайте правила добавления данных для каждой таблицы
6. Добавьте транспортные средства чтобы в каждой созданной таблице было не менее трех транспортных средств
7. Добавьте несколько мотоциклов весом меньше одной тонны
8. Сделайте выбор из всех таблиц в том числе и из основной
9. Сделайте выбор только из основной таблицы



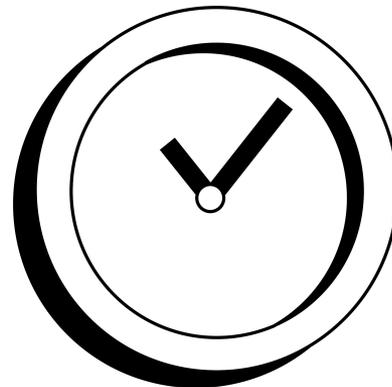
40 минут



## Задание 3

<<40:00->>

1. Откройте консоль Postgress
2. Создайте таблицу vehicles с полями vehicle\_type, plate\_number, date\_of\_issue (в формате DD-MM-YYYY), weight, owner
3. Сделайте таблицы для горизонтального партицирования по весу машины (от 1 тонны до 2.5 тонн, от 2.5 до 4 тонн, больше 4 тонн)
4. Сделайте таблицы для горизонтального партицирования по году выпуска машины (до 2000, с 2000 до 2019, после 2019)
5. Создайте правила добавления данных для каждой таблицы
6. Добавьте транспортные средства чтобы в каждой созданной таблице было не менее трех транспортных средств
7. Добавьте несколько мотоциклов весом меньше одной тонны
8. Сделайте выбор из всех таблиц в том числе и из основной
9. Сделайте выбор только из основной таблицы





## Задание 4

1. Загрузите из Excel файла график ипотечных платежей через Spark.
2. При необходимости напишите запросы на создание и удаление таблицы в mysql.
3. Через Spark добавьте поля по накопленному итогу по процентам и долгу.
4. Конвертируйте spark df в pandas df и с помощью matplotlib постройте графики с кумулятивными выплатами долга и процентов.



**40 минут**



Вопросы?



Вопросы?



Вопросы?





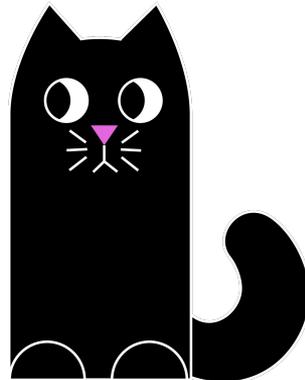
# Домашнее задание



## Домашнее задание

На основе сайта [yandex.ru](https://yandex.ru):

- Определите, на каком протоколе работает сайт.
- Проанализируйте структуру страницы сайта
- Внесите не менее 10 изменений на страницу с помощью инструмента разработчика и представьте скриншоты было/стало.
- Создайте прототип низкой детализации (дополнительное задание, если на семинаре дошли до задания №8)





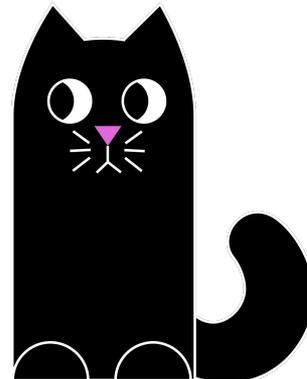
## Домашнее задание

За основу возьмите Задание 4 решенное на семинаре.

В файле s4\_2 параметры кредита: Займ 9400000, срок 30 лет, ставка 10.6%.

Через <https://calcus.ru/kreditnyj-kalkulyator-s-dosrochnym-pogasheniem> добавьте два листа в Excel с постоянным платежом 120 или 150 тыс. руб.

Добавьте графики с досрочным погашением по этим параметрам. Т.е. линии по выплатам основного долга и процентов если платеж будет 120 или 150 тыс. руб. В результате должно получиться 6 линий. Используйте разные цвета.





Спасибо за внимание

