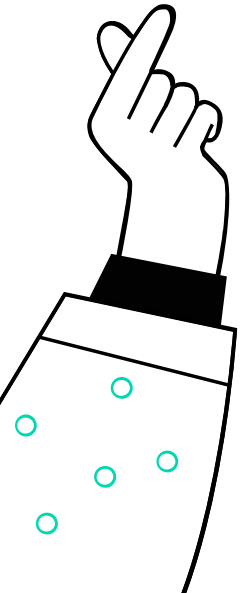
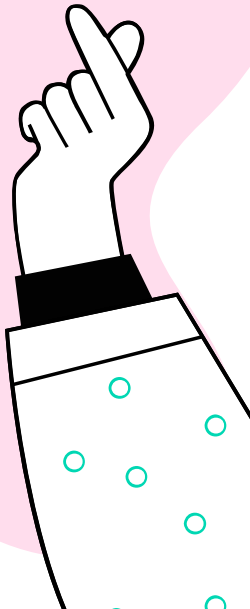


Собираем витрину, как в бигтехе:

от описания до результата



Aigul Sibgatullina
tg: @data_engineerette

План вебинара

01
ЗНАКОМСТВО

02
ПРЕДМЕТНАЯ ОБЛАСТЬ

03
ТЕХНИЧЕСКИЕ ШТУКИ

04
ПРАКТИКА

05
ИТОГИ

06
ПОЛЕЗНОСТИ

01

ЗНАКОМСТВО

Кто я

- 5+ лет опыта
- 3 года в дата-аналитике
- 2 года в дата-инженерии
- аудит, антифрод, маркетинг
- веду тг-канал
- попала на стажировку по Java и бесплатно слетала в Адлер на 2 недели
- выиграла лонгборд от CTC Love

Кто вы :)

Напишите в чатике:

- Кем работаете/кем хотелось бы работать
- Сколько лет стажа в сфере данных

Пример:

- «DA, 1 год»
- «хочу в DE, студент»

02

ПРЕДМЕТНАЯ ОБЛАСТЬ

О нас сегодня

Кто мы?

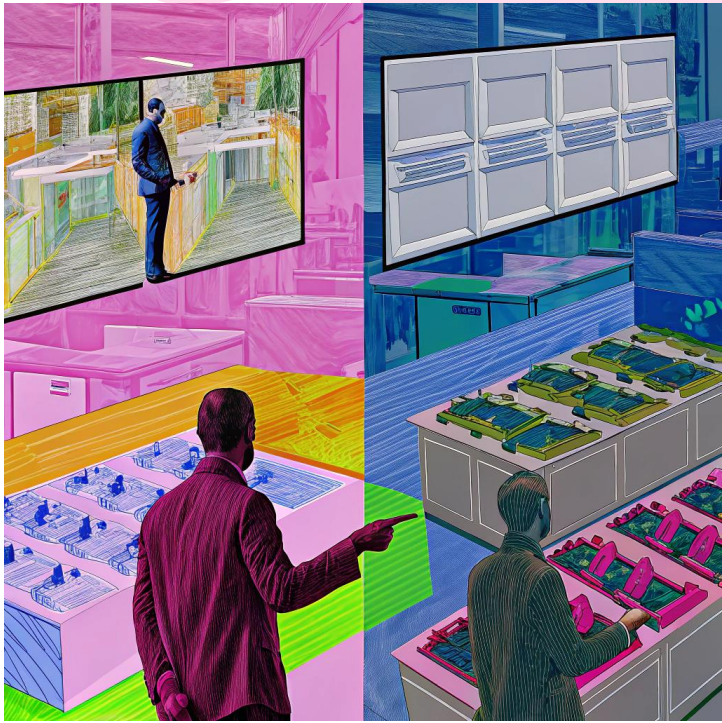
- Бизнесмены и бизнесвумены на один вечер

Что мы делаем?

- Создаем дизайн-проекты кухонек
- Делаем кухоньки на заказ

Что мы хотим?

- Хотим продавать больше кухонек
- Хотим зарабатывать денежки на дальнейшее развитие бизнеса



АНАЛИТИКА

«Найти идеи, которые позволят получить доходность выше рынка»

«Отследить, какие каналы принесли продажи и их объем»

«Как эффективнее распределить бюджет и ресурсы»

AS IS

- много разных источников данных
- данные лежат в несовместимых форматах
- большой объем данных
- требуется ручная работа

TO BE

Данные объединены, агрегированы и собраны в одном месте для оперативной аналитики, построения дашбордов и поиска инсайтов

AS IS

- много разных
  источников да
- данные лежат
  несовместимы
  форматах
- большой объе
- требуется руч

O BE

объединены,
аны и собраны
м месте для
ной аналитики,
ия дашбордов
ка инсайтов

ФЛОУ РЕАЛИЗАЦИИ ВИТРИНЫ ДАННЫХ

01

РАЗРАБОТКА МЕТОДОЛОГИИ

Дата-аналитики

02

СОСТАВЛЕНИЕ ТЗ*

Системные аналитики

03

СОЗДАНИЕ ВИТРИНЫ. ВЫВОД В ПРОД

Дата-инженеры

04

ВИЗУАЛИЗАЦИЯ. ПОИСК ИНСАЙТОВ

Дата-аналитики
Дата-сайентисты

*ТЗ – техническое задание

03

ТЕХНИЧЕСКИЕ ШТУКИ

I will be working with Big Data

You're given four excel files as a data source

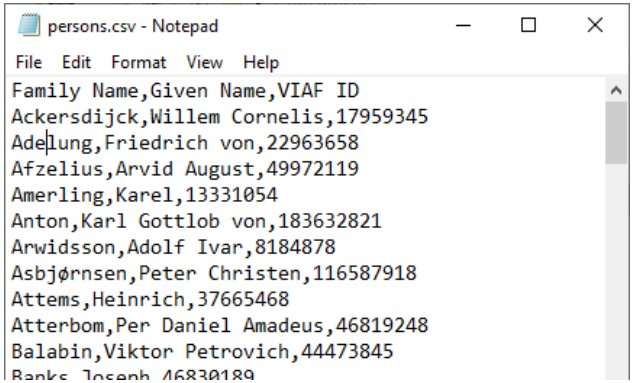
«Я буду работать с бигдатой» — И здесь тебе дают четыре эксельки в качестве источника данных — «Принимается»

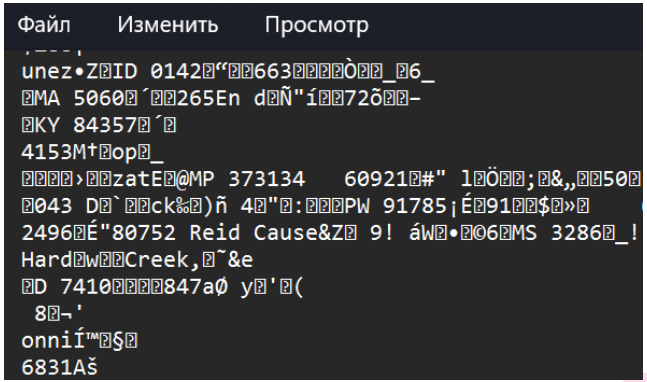
Описание исходных данных*

	Место хранения	Комментарий
visits	ClickHouse	Посещение пользователями нашего сайта
costs	Postgres	Расходы на рекламу
campaigns_dict	csv	Словарик с сопоставлением рекламных кампаний
submits	parquet	Заполненные формы на сайте для заявки/обратного звонка
deals	parquet	Заказанные дизайн-проекты

*Сгенерированные, синтетические

Форматы файлов





csv (comma-separated values)

текстовый формат с разделителями-запятыми

parquet (паркет)

бинарный колоночный формат

Что будем использовать

Postgres

транзакционная бд
как источник данных

ClickHouse

аналитическая колоночная бд
как источник данных

pgAdmin

клиент для подключения
к Postgres

DBeaver

клиент для подключения
к разным бд

Spark

движок для обработки
больших данных

Jupyter Notebook

среда интерактивной
разработки

Metabase

инструмент
для визуализации данных

Что еще нужно – драйвера

Postgres
https://jdbc.postgresql.org/download/

ClickHouse
https://mvnrepository.com/artifact/com.clickhouse/clickhouse-jdbc

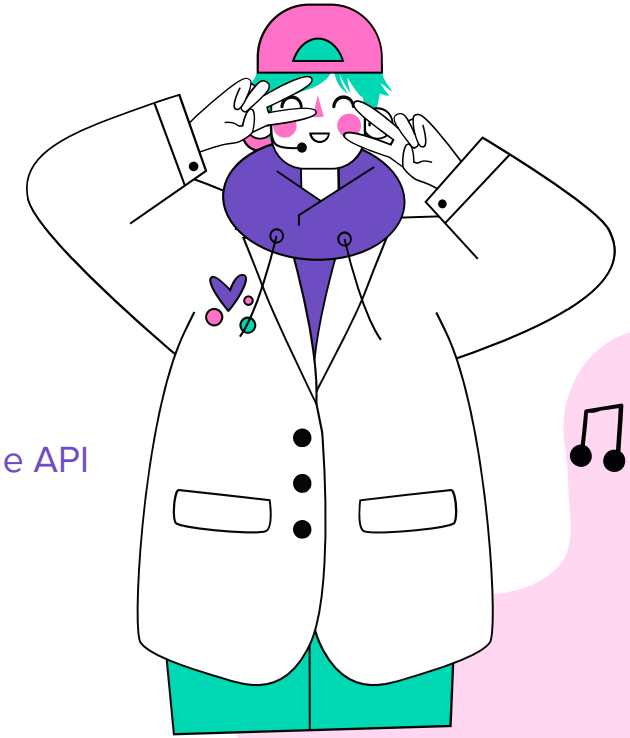
04
ПРАКТИКА

05

ИТОГИ

Чему мы научились

- понимать роль дата-инженера в потоке поставки данных
- разбираться в ТЗ и методологиях
- работать со Spark, применять функции в Spark
- читать из файлов и баз данных
- писать в файлы и базы данных
- преобразовывать диалекты SQL и pandas на Spark DataFrame API
- использовать Spark UI
- собирать витрину данных
- интерпретировать полученные результаты



06

ПОЛЕЗНОСТИ

Изученные функции в Spark

Из модуля pyspark.sql.functions:

- **col** – обращение к столбцу
- **lit** – столбец в виде конкретного значения
- **substring** – выделение подстроки
- **concat** – объединение столбцов
- **split** – разделение столбца
- **date_format** – преобразование формата даты
- **regexp_replace** – замена по регулярке
- **md5** – хеширование столбца алгоритмом md5
- **sum** – суммирование
- **countDistinct** – подсчет уникальных значений
- **when/otherwise** – case when

Изученные функции в Spark

Над датафреймом:

- **select** – выбор столбцов
- **where** – фильтр
- **groupBy** – группировка по столбцам
- **agg** – агрегация
- **join** – объединение таблиц
- **sort** – сортировка
- **withColumn** – добавление нового столбца
- **cache** – кэширование датафрейма в памяти
- **unpersist** – освобождение ресурсов
- **show** – просмотр датафрейма
- **printSchema** – структура датафрейма
- **count** – количество строк
- **distinct** – отбор уникальных значений
- **withColumnRenamed** – переименование столбца
- **drop** – удаление столбца

Изученные функции в Spark

Над столбцом:

- **alias** – псевдоним
- **cast** – преобразование к типу данных
- **isin** – нахождение значения столбца в списке значений
- **rlike** – сопоставление по регулярке
- **getItem** – получение элемента из массива
- **between** – фильтр между двумя границами

Полезные ссылки

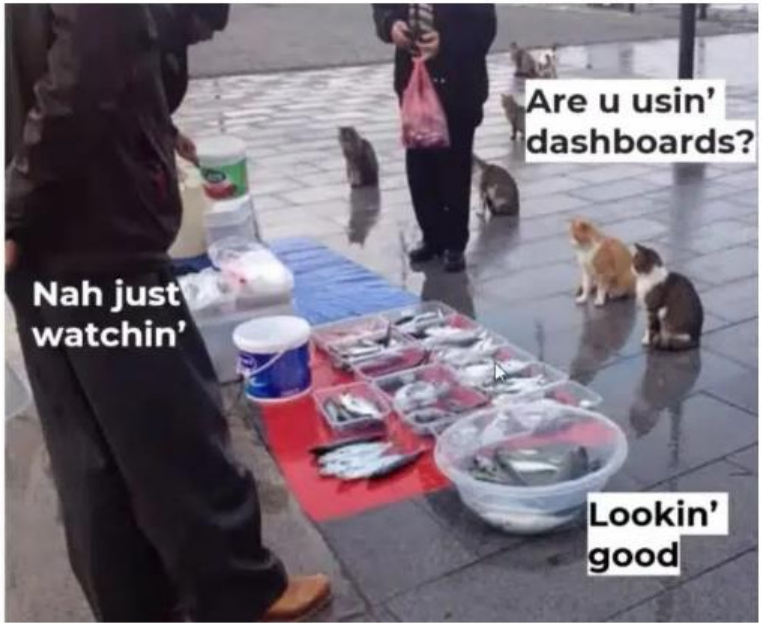
Материалы по вебинару
https://github.com/Aigul9/spark-webinar

Туториал по Spark
https://colab.research.google.com/drive/1G894WS7ltlUTusWWmsCnF_zQhQqZCDOc

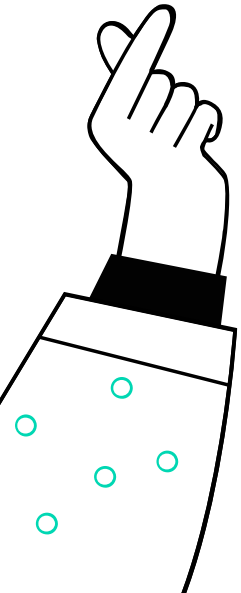
Документация по функциям в Spark
https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/index.html

Моя статья «Spark. План запросов на примерах»
https://habr.com/ru/articles/807421/

Dashboards will be useful

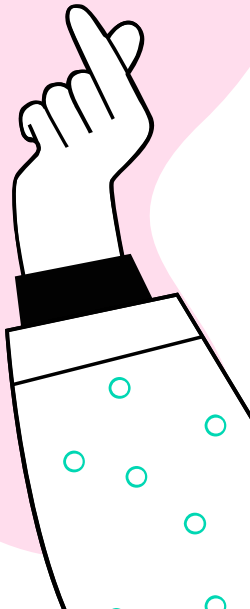


Дашборды будут полезны — «Пользуетесь дашбордами?» — «Нет, только смотрим» — «Красивое»



Собираем витрину, как в бигтехе:

от описания до результата



Aigul Sibgatullina
tg: @data_engineerette