The background features a large, soft pink cloud-like shape. Scattered around the text are several decorative elements: a single musical note in the top left, a double musical note in the top right, a single heart in the top left, a double heart in the top right, a single musical note in the bottom left, and a double musical note in the bottom right. On the far left and right edges, there are illustrations of hands in white sleeves with black cuffs, making a 'finger heart' gesture. The sleeves are decorated with small teal circles.

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 в сфере данных
- 3 года в дата-аналитике
- 2 года в дата-инженерии
- аудит, антифрод, маркетинг
- веду тг-канал «дата инженеретта»
- 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»

The background is a solid light pink color. It features several white, organic, cloud-like shapes of various sizes. Scattered across the pink background are three black musical notes: a single eighth note in the upper right, and two pairs of beamed eighth notes, one in the lower left and one in the lower center.

02

ПРЕДМЕТНАЯ ОБЛАСТЬ



О нас сегодня

Кто мы?

- Бизнесмены и бизнесвумены на один вечер
- Есть свой сайт <https://our-cool-website.com>

Что мы делаем?

- Создаем дизайн-проекты кухонек
- Делаем кухоньки на заказ

Что мы хотим?

- Хотим продавать больше кухонек
- Хотим зарабатывать денежки на 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 данных
- данные лежат в несовместимых форматах
- большой объем данных
- требуется ручная обработка



ты спишь?

вставай, мы
все витрины
сломали

вообще все витрины
сломали

Не знаю когда починим, но
данные ты не получишь



TO BE

объединены,
данные сгруппированы и собраны
в одном месте для
простой аналитики,
система дашбордов
для визуализации инсайтов



ФЛОУ РЕАЛИЗАЦИИ ВИТРИНЫ ДАННЫХ



*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




```
persons.csv - Notepad
File Edit Format View Help
Family Name,Given Name,VIAF ID
Ackersdijck,Willem Cornelis,17959345
Ade lung,Friedrich von,22963658
Afzelius,Arvid August,49972119
Amerling,Karel,13331054
Anton,Karl Gottlob von,183632821
Arwidsson,Adolf Ivar,8184878
Asbjørnsen,Peter Christen,116587918
Attems,Heinrich,37665468
Atterbom,Per Daniel Amadeus,46819248
Balabin,Viktor Petrovich,44473845
Banks,Joseph,16830189
```

csv (comma-separated values)

текстовый формат с разделителями-запятыми

```
Файл  Изменить  Просмотр
-----
unez•ZID 0142"“6630000_06_
MA 5060'2265En dÑ"í720-
KY 84357'
4153M+op_
zatzatE@MP 373134 60921# 1Ö;8,,50
043 D`ckk)ñ 4":PW 91785;É91$»
2496É"80752 Reid Cause&Z 9! áw•06MS 3286_!
HardwCreek,~&e
D 7410000847aø y' (
8- '
onnií™$
6831Aš
```



parquet (паркет)

бинарный колоночный формат



Что будем использовать



Postgres

транзакционная бд
как источник данных

ClickHouse

аналитическая колоночная бд
как источник данных

Docker

платформа
контейнеризации

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

<https://github.com/Aigul9/spark-webinar/tree/main>




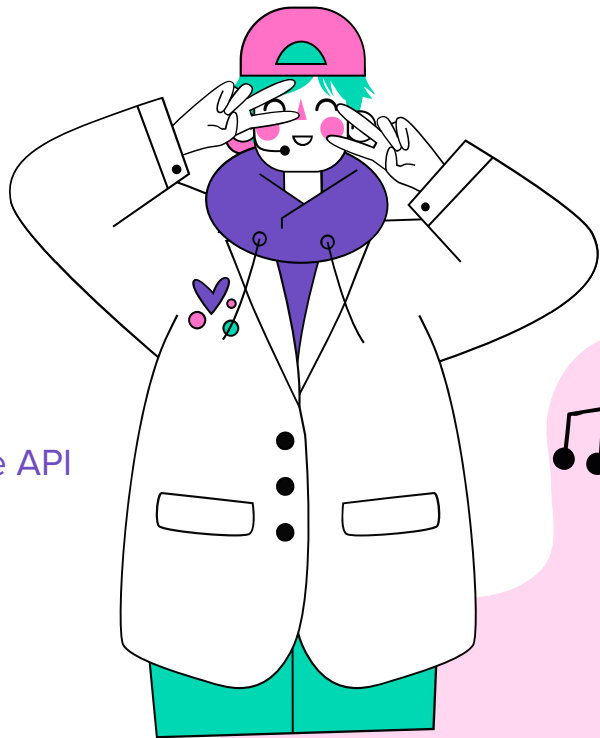
05

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>

Тutorial 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»



The background features a large, soft pink cloud-like shape. Scattered around the text are several decorative icons: a single musical note in the top left, a double musical note in the top right, a single musical note in the bottom left, and a double musical note in the bottom right. There are also two pink hearts, one on the left and one on the right. On the far left and right edges, there are illustrations of hands in white sleeves with black cuffs and teal polka dots, making a 'finger heart' gesture. The main text is centered and reads:

Собираем витрину, как в бигтехе: от описания до результата

Aigul Sibgatullina
tg: @data_engineerette