Технологии работы с большими данными

Лекция 1. Традиционные аналитические подходы

нетология

Алексей Кузьмин

Директор разработки; ДомКлик.ру

О спикере:

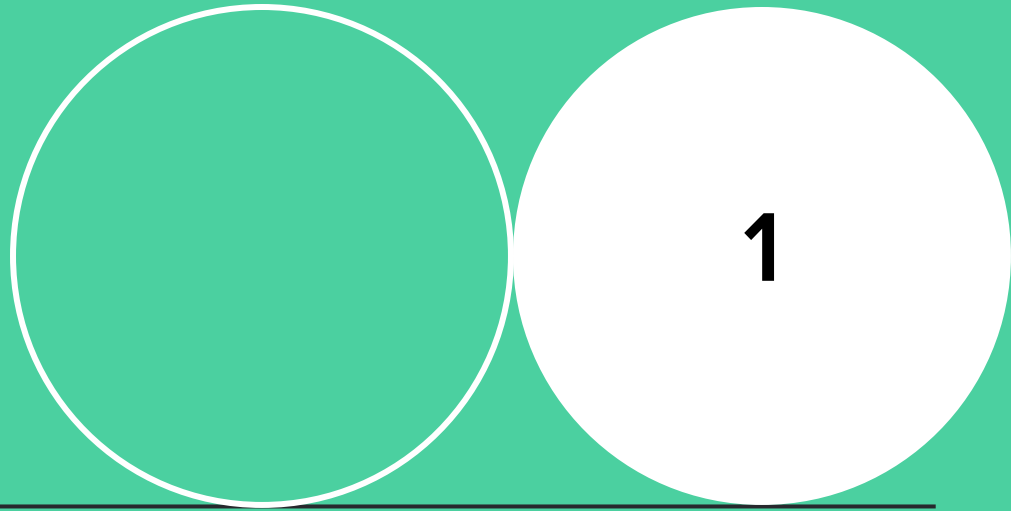
- Руковожу направлением работы с данными и Data Science
- Преподаю в Нетологии
- Работаю в IT с 2010 года (ABBYY, ДомКлик)
- Окончил МехМат МГУ в 2012 году

Я в Слаке:

@Alexey Kuzmin

О Курсе

Что будем изучать?



Алексей Кузьмин

Технологии работы с большими данными

нетология

Зачем мы тут?

1. Узнать, как развивались методы работы с данными
2. Понять, почему технологии обработки больших данных появились тогда, когда появились и какой value они могут дать на проекте
3. Своими руками попробовать все этапы и уровни анализа данных
4. Узнать, когда и как стоит внедрять анализ больших данных у себя в компании

Структура курса

1. Традиционные аналитические подходы
2. DataScience
3. Мотивация и инструменты больших данных
4. NoSQL-подход
5. MapReduce - теория
6. PySpark - практика 1
7. PySpark - практика 2
8. Культура сбора данных
9. Организация команды для работы с данными
10. Примеры и разбор кейсов анализа данных
11. Лабораторная работа и итоговая работа

Работа с данными

Источники данных

Сбор данных
Лекция 8

SQL-БД
Лекция 1

NoSQL
Лекция 4

MapReduce
Лекция 5-7

Мотивация Big Data
Лекция 3

Data Science
Лекция 2

Отчеты
Лекция 1

Модели
Лекция 2

Люди и процессы
Лекция 9

Примеры кейсов
Лекция 10

Как будет проходить обучение?

1. Лекции + практическая часть на каждой лекции
2. Домашнее задание

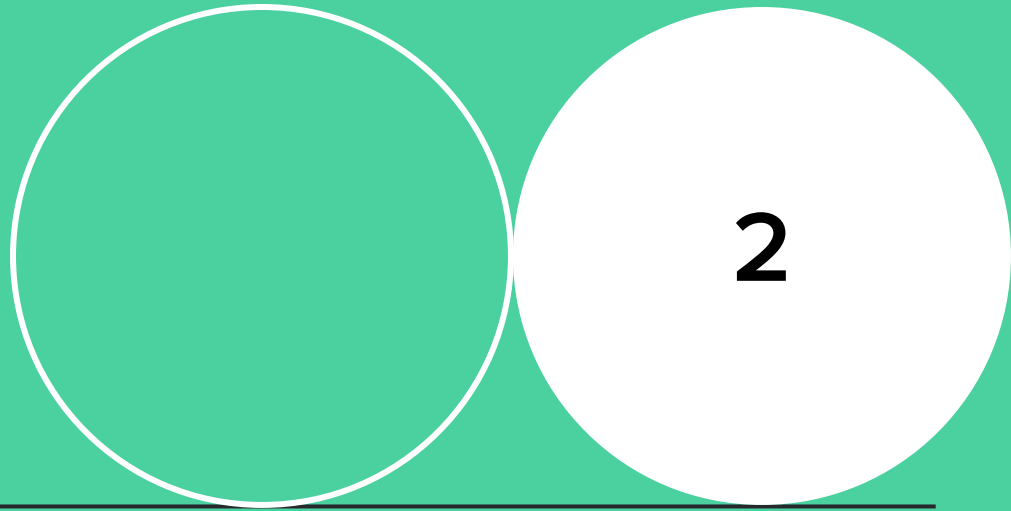
Каждое следующая лекция является расширением предыдущей, поэтому лекции лучше смотреть по порядку

Как преуспеть в практике

1. “Мне должны все рассказать” - проигрышная стратегия
    a. Мир постоянно развивается
    b. Постоянная актуализация своих знаний - неотъемлемый атрибут любого успешного человека
2. Лекции + практические задания дают основную канву и основные принципы
3. Но для успешного выполнения домашнего задания иногда придется расширять эти знания и искать ответы в интернете
4. Выигрышная стратегия
    a. Поискать самостоятельно
    b. Не получилось - спросить в группе
    c. Опять не получилось - спросить у преподавателя

Традиционная аналитика

Любое движение имеет свое начало



Алексей Кузьмин

Технологии работы с большими данными

нетология

Бизнес-анализ

Бизнес-анализ– это ключевой инструмент управления

Позволяет получить достоверную картину текущего состояния дел в компании.

От качества и эффективности процессов анализа зависит достоверность и качество принимаемых управленческих решений

Задачи бизнес анализа

- обеспечение достоверной информации в нужном разрезе для принятия управленческих решений;
- определение уровня текущей эффективности бизнес-процессов;
- ...

Предоставление отчетов на основе данных о состоянии дел в компании

Инструменты бизнес-аналитика

Даже в 2019 году основным инструментом бизнес-аналитика остается Excel

Преимущества:

- Низкий порог вхождения
- Возможность работать “мышкой”
- Удобство доставки и демонстрации заказчику
- Относительно богатый инструментарий

Практика 1

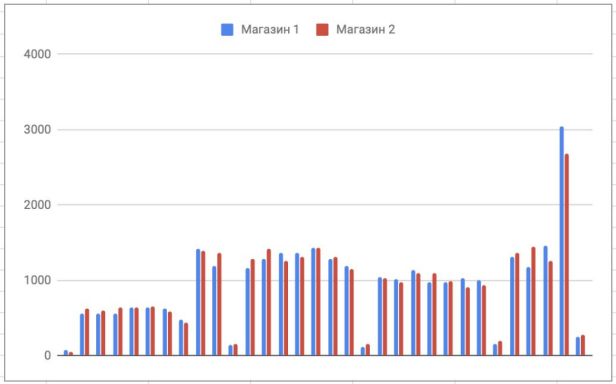
Возьмем файл со статистикой продаж - https://docs.google.com/spreadsheets/d/12f9afcvue9X3y6VxS_Dd0NU4XLo6LaJn7D4hw1uNDjw/edit?usp=sharing

Попробуем ответить на вопросы:

- Какие доходы нашего пункта проката по дням
- Построить график доходов каждого магазина по дням

Решение практики

SUM of Сумма	Магазин			
Дата - Day-Month		Магазин 1	Магазин 2	Grand Total
14-Feb		73.84	42.89	116.73
15-Feb		563.51	625.41	1188.92
16-Feb		558.65	595.53	1154.18
17-Feb		555.72	632.45	1188.17
18-Feb		639.51	636.47	1275.98
19-Feb		634.48	656.42	1290.9
20-Feb		628.57	590.52	1219.09
21-Feb		481.91	435.96	917.87
1-Mar		1423.68	1384.56	2808.24
2-Mar		1186.21	1363.84	2550.05
16-Mar		139.64	159.64	299.28
17-Mar		1158.24	1283.92	2442.16
18-Mar		1288.01	1413.75	2701.76
19-Mar		1363.72	1253.97	2617.69
20-Mar		1362.97	1306.92	2669.89
21-Mar		1433.65	1434.62	2868.27
22-Mar		1276.91	1309.88	2586.79
23-Mar		1193.2	1149.23	2342.43
5-Apr		118.74	154.62	273.36
6-Apr		1043.57	1033.57	2077.14
7-Apr		1010.64	973.64	1984.28
8-Apr		1135.47	1092.37	2227.84
9-Apr		970.61	1097.25	2067.86
10-Apr		980.59	992.59	1973.18
11-Apr		1028.59	911.73	1940.32
12-Apr		994.72	935.76	1930.48
26-Apr		154.64	192.57	347.21
27-Apr		1308.89	1364.68	2673.57
28-Apr		1175.05	1447.68	2622.73
29-Apr		1457.64	1259.96	2717.6
30-Apr		3044.24	2679.65	5723.89
14-May		243.1	271.08	514.18
	0			0
Grand Total	**0**	**30628.91**	**30683.13**	**61312.04**



Самостоятельная практика

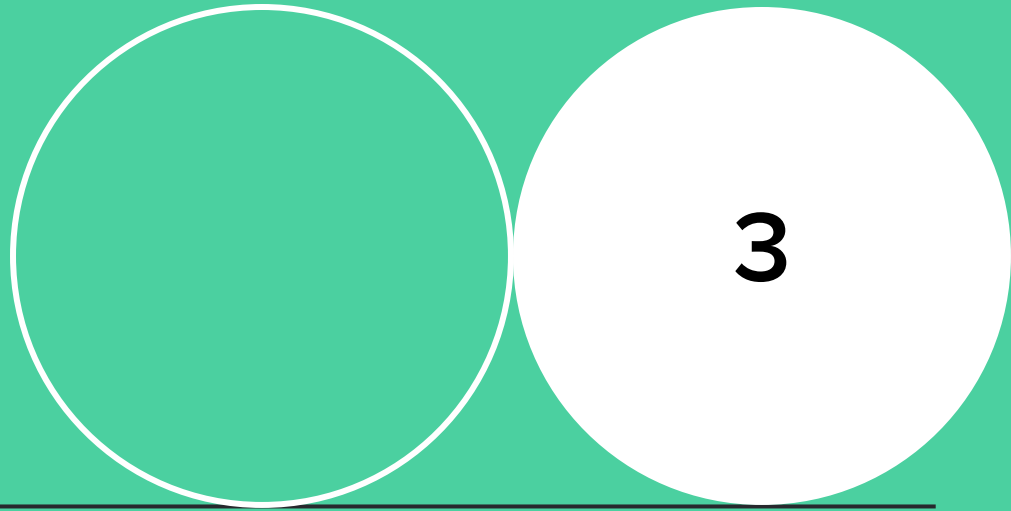
- Вывести top самых прибыльных клиентов

Дополнительные возможности Excel

- Связывание данные в нескольких таблицах
- Продвинутые методы анализа данных
- Работа с текстом
- Защиты и контроль качества данных

Откуда берутся данные?

Не Excel’ем единым



Алексей Кузьмин

Файлы

- Исторически данные брались из разных файлов
- Основные форматы:
  - Excel -> бинарный файл с данными excel
  - CSV -> табличный файл, содержащий строки, разделенные запятой

Пример CSV

```
"Покупатель","Сумма","Название фильма","Дата платежа","Магазин"
Peter Menard,7.99,Rules Human,2007-02-15 22:25:46,Магазин 1
Peter Menard,1.99,Majestic Floats,2007-02-16 17:23:14,Магазин 2
Peter Menard,7.99,Maiden Home,2007-02-16 22:41:45,Магазин 1
Peter Menard,2.99,Hyde Doctor,2007-02-19 19:39:56,Магазин 1
Peter Menard,7.99,Massacre Usual,2007-02-20 17:31:48,Магазин 1
Peter Menard,5.99,Annie Identity,2007-02-21 12:33:49,Магазин 1
Harold Martino,5.99,Wash Heavenly,2007-02-17 23:58:17,Магазин 2
Harold Martino,5.99,Lola Agent,2007-02-20 02:11:44,Магазин 2
Harold Martino,2.99,Identity Lover,2007-02-20 13:57:39,Магазин 2
```

Файлы - не удобны

- Данные актуальны только на момент создания файла
- Данные из разных источников плохо бьются между собой
- ПО не хранит данные в файлах!
- Почему бы не брать данные там, где их хранит ПО?

На выручку приходят Базы Данных...

Зачем нужны Базы Данных?

Основная функция базы данных – предоставление единого хранилища для всей информации, относящейся к определенной теме.

- Вместо того чтобы выискивать нужные сведения в документах Word, таблицах Excel, текстовых файлах, сообщениях электронной почты и самоклеющихся заметках, их можно взять из единой базы.

- База данных может содержать все что угодно, будь-то список приглашенных на свадьбу гостей или информация о каждом клиенте, посетившем Web-сайт электронного магазина и разместившего там свои заказы.

Реляционные БД

- набор таблиц, связанных или не связанных общими ключами
- значения хранятся в строках и столбцах

Номер заказа	Код услуги	Номер телефона	Дата разговора	Код города	Продолжи-тельность	Стоимость
1	4	543-67-12	3.10.02	523	5	10,12
2	5	234-56-18	3.10.02	736	34	45,50
3	7	874-34-54	3.10.02	945	7	4,10
4	10	112-58-12	3.10.02	153	30	120,80
5	12	453-22-54	4.10.02	023	2	2
6	9	638-71-61	4.10.02	152	9	6
7	12	442-68-32	4.10.02	042	3	1
8	10	618-31-15	4.10.02	005	14	56
9	11	736-84-53	4.10.02	513	20	123
10	4	231-65-34	4.10.02	041	12	45

Реляционные БД

В таблицах хранится информация об объектах, представленных в базе данных.

К этим данным можно получить доступ многими способами, и при этом реорганизовывать таблицы БД не требуется.

SQL

Специальный язык для работы с данными в реляционной бд

- добавление строк данных
- обновление строк данных
- удаление строк данных
- извлечение наборов данных
- управление всеми аспектами работы базы данных

Основа всего - оператор SELECT

Для извлечения данных из БД применяется команда SELECT. В упрощенном виде она имеет следующий синтаксис:

```
SELECT список_столбцов FROM имя_таблицы;
```

Пример - таблица продуктов (products)

id	name	manufacturer	count	price
1	iPhone 11 Pro	Apple	3	89000
2	iPhone 11	Apple	4	59000
3	Galaxy A71	Samsung	5	29000
4	Galaxy Note 10	Samsung	3	71000

Основа всего - оператор SELECT

получить название товара и цену из этой таблицы можно командой:

```
SELECT name, price FROM products;
```

name	price
iPhone 11 Pro	89000
iPhone 11	59000
Galaxy A71	29000
Galaxy Note 10	71000

Основа всего - оператор SELECT

получить все объекты из этой таблицы можно командой:

```
SELECT * FROM products;
```

- * - специальный символ, который в операторе SELECT говорит “выбрать все колонки”

Основа всего - оператор SELECT

Спецификация столбца не обязательно должна представлять его название. Это может быть любое выражение, например, результат арифметической операции. Рассмотрим следующий запрос:

```
SELECT name, price * count FROM products;
```

name	Price * count
iPhone 11 Pro	267000
iPhone 11	236000
Galaxy A71	145000
Galaxy Note 10	213000

Для сортировки - добавим ORDER BY

Оператор ORDER BY позволяет отсортировать значения по определенному столбцу. Например, упорядочим выборку из таблицы **products** по столбцу **count**:

```
SELECT * FROM products ORDER BY count;
```

id	name	manufacturer	count	price
4	Galaxy Note 10	Samsung	3	71000
1	iPhone 11 Pro	Apple	3	89000
2	iPhone 11	Apple	4	59000

Для сортировки - добавим ORDER BY

По умолчанию данные сортируются по возрастанию, однако с помощью оператора **DESC** можно задать сортировку по убыванию. Явно задать сортировку по возрастанию можно указав оператор **ASC**.

```
SELECT * FROM products ORDER BY count DESC;
```

id	name	manufacturer	count	price
3	Galaxy A71	Samsung	5	29000
2	iPhone 11	Apple	4	59000
4	Galaxy Note 10	Samsung	3	71000

Для фильтрации - WHERE

Для фильтрации данных применяется оператор **WHERE**, после которого указывается условие, на основании которого производится фильтрация.

- = сравнение на равенство
- <> сравнение на неравенство
- < меньше чем, > больше чем
- <= меньше чем или равно, >= больше чем или равно

Для фильтрации - WHERE

Например, найдем всех товары, производителем которых является компания Apple:

```
SELECT * FROM products WHERE manufacturer = 'Apple';
```

id	name	manufacturer	count	price
1	iPhone 11 Pro	Apple	3	89000
2	iPhone 11	Apple	4	59000

Практика

Выберем все данные из таблицы

- https://pgexercises.com/questions/basic/selectall.html

Самостоятельная практика

Выполните еще одно упражнение:

- https://pgexercises.com/questions/basic/selectspecific.html

Что еще есть в SQL?

Запросы:

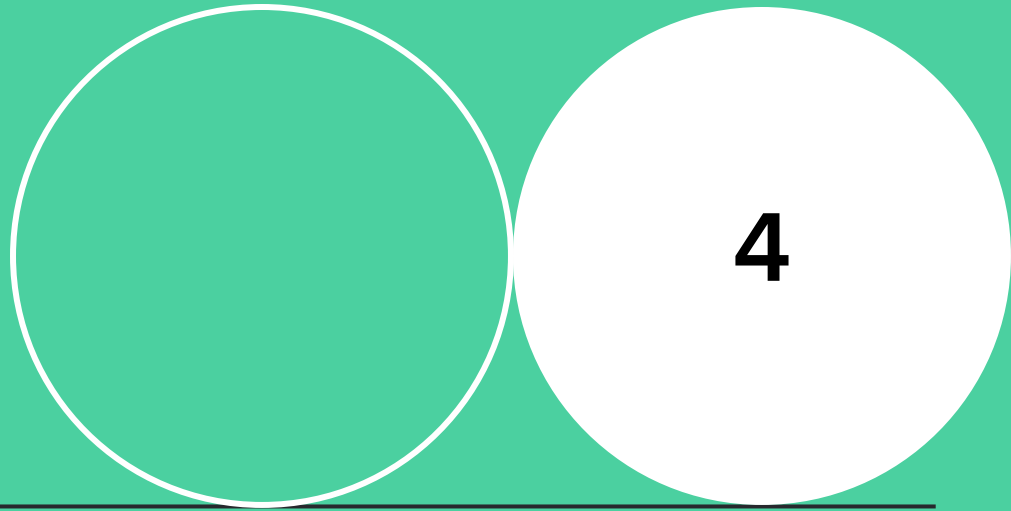
- Соединение данных из разных таблиц
- Группировка и вычисление агрегатов
- ...

Работа с данными:

- Создание и работа с таблицами и базами
- Вставка, обновление и удаление данных
- ...

BI

Наведем марафет



Алексей Кузьмин

Технологии работы с большими данными

нетология

Excel и SQL

Excel (в отличии от Spreadsheets) умеет забирать данные из БД

Это позволяет аналитикам работать напрямую с данными в базе

- Повышается актуальность
- Повышается надежность данных

К сожалению, excel делает это только в момент обновления - достаточно долгой и не всегда доступной операции

Кроме того, возможности по визуализации и анализу данных ограничены

BI

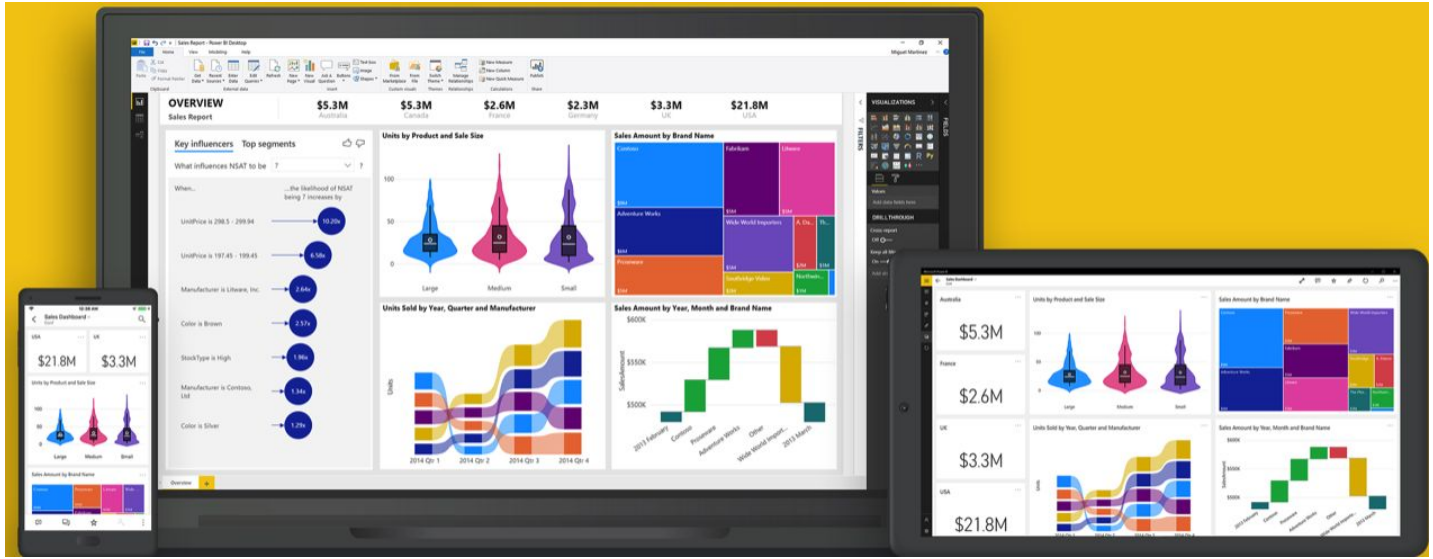
Excel файлы с данными из базы начинают занимать много места на диске, а их потребность в постоянной актуализации начинает вызывать раздражение у бизнеса.

Возможности визуализации перестают соответствовать современным потребностям (например, очень сложно нарисовать карту с самыми продающими магазинами)

Это приводит к появлению нового класса аналитических инструментов - BI-дашбордов.

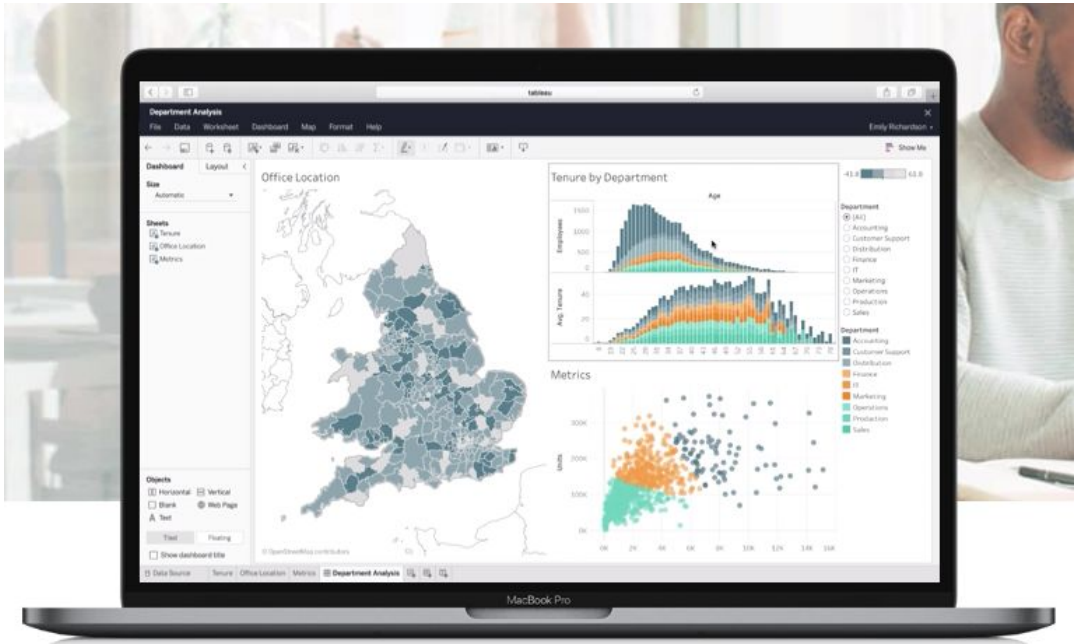
Power BI

- https://powerbi.microsoft.com/ru-ru/
- https://netology.ru/programs/power-bi#/presentation



Tableau

- https://www.tableau.com
- https://netology.ru/programs/tableau



Еще инструменты

- Google Data Studio - https://datastudio.google.com/
- Pentaho Reporting - https://www.hitachivantara.com
- Qlik Sense - https://www.qlik.com/us/products/qlik-sense
- Redash - https://redash.io
- Grafana - https://grafana.com
- ...

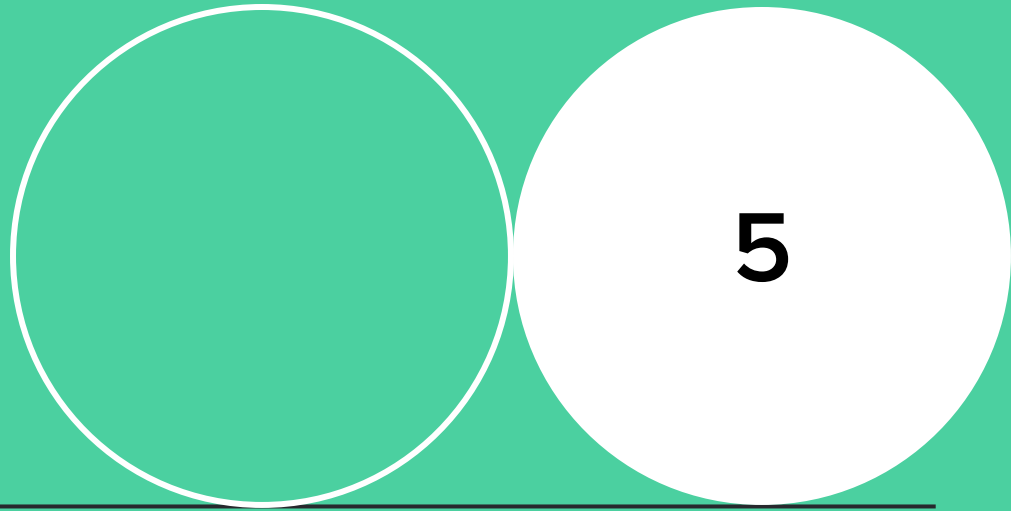
BI

Они сразу поддерживают множество средств визуализации

Более заточены именно на презентацию анализа, а не на его проведение

В платных версиях умеют актуализировать себя по расписанию или по запросу

Что дальше?



Алексей Кузьмин

Переходим от аналитики к анализу

- Хочется понимать не только, как идут дела в компании, но и почему они так идут
- Как будут идти дела завтра
- Что нужно сделать, чтобы дела шли лучше

Excel’я катастрофически начинает не хватать

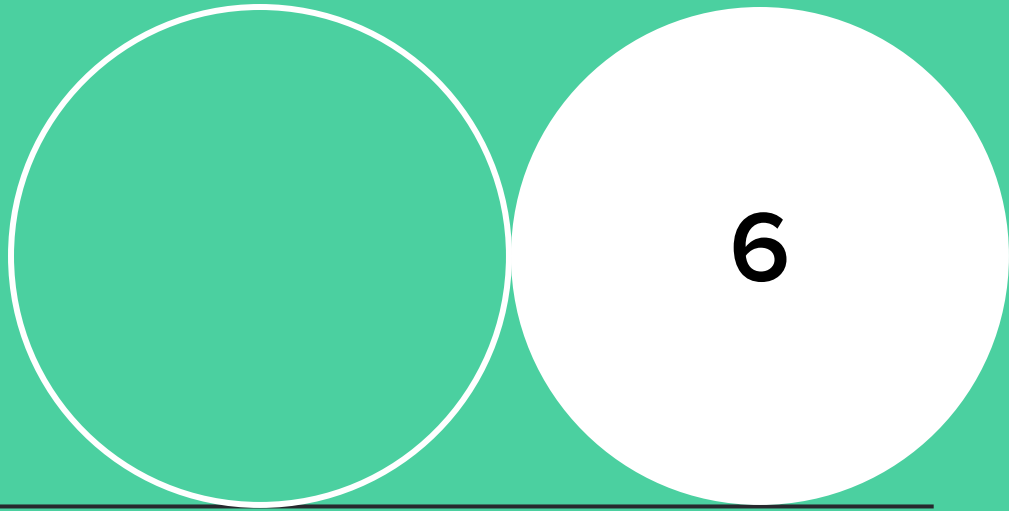
“Почему сейчас делают self-driving автомобили, а мы до сих пор не можем сказать, сколько товаров продадим завтра?”

Типичный руководитель

DS

Тут на выручку приходит DataScience со своими возможностями, но это уже тема следующей лекции...

Итоги



Алексей Кузьмин

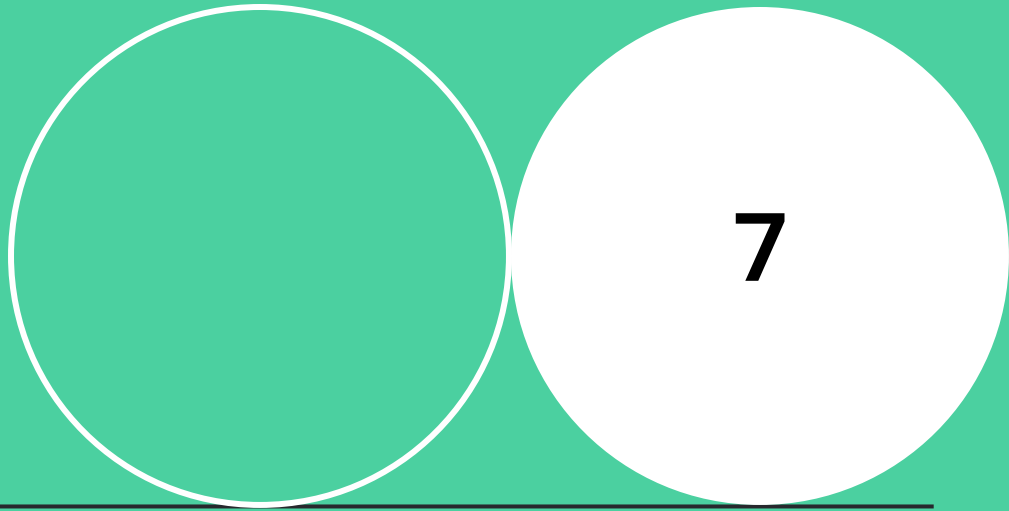
Технологии работы с большими данными

нетология

Что мы узнали сегодня

- Поговорили про классическую бизнес-аналитику
- Посмотрели пару примеров и инструментов
- Узнали, с чего вообще начинается работа с данными в любой компании

Домашнее задание



Алексей Кузьмин

Технологии работы с большими данными

нетология

Домашнее задание

- На основе данных по фильмам
  - Нарисовать круговую диаграмму (pie-chart) с доходами от 5-ти самых прибыльных фильмов за все время
  - Получить top 5 самых продаваемых фильмов в первом магазине
- Выполнить еще 2 задания на SQL
  - https://pgexercises.com/questions/basic/where.html
  - https://pgexercises.com/questions/basic/where2.html

В качестве решения в ЛК:

- Ссылку на расшаренный google spreadsheet
- Скриншоты выполненных pgexercises

Спасибо за внимание

Алексей
Кузьмин

нетология