Лекция

«Удивительный новый мир цифровых данных»

Ваш лектор:
Доцент РТУ-МИРЭА
Смирнов Михаил

E-mail: mikhaelsmirnov@gmail.com

ОПРЕДЕЛИМ ПРЕДМЕТ ДИСКУССИИ

Данные.

Форма представления информации в виде фактов.

Факты можно изучать, классифицировать, по ним можно делать выводы и получать новые знания.

Благодаря данным с вашего Apple Watch я могу узнать практически все о вашем образе жизни. И даже спрогнозировать, что вы будете завтра… или через неделю.

Базы данных.

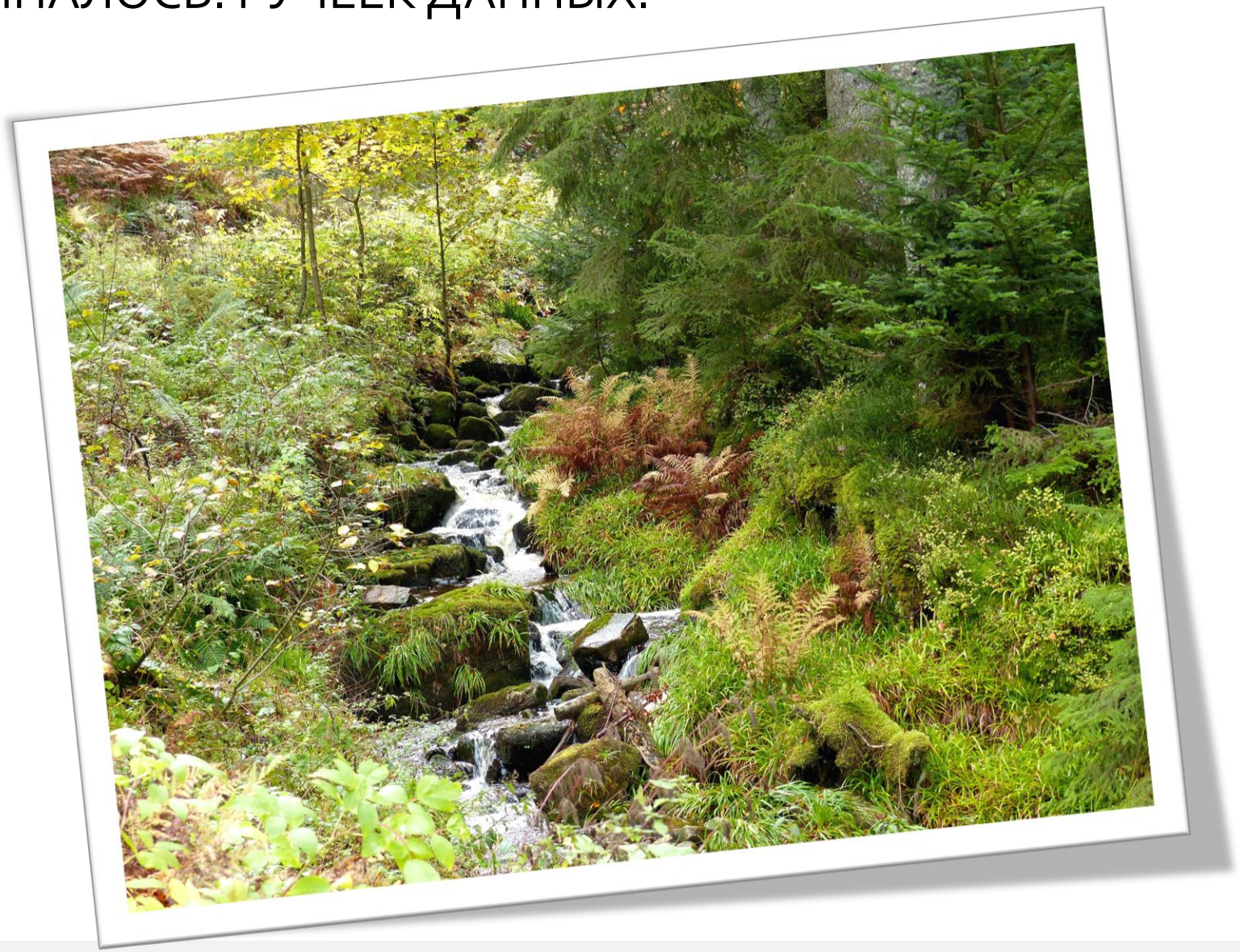
Программное обеспечение, фреймворки, технологии, главной задачей которых является обеспечение возможности безопасного хранения данных.

Их ОЧЕНЬ много.

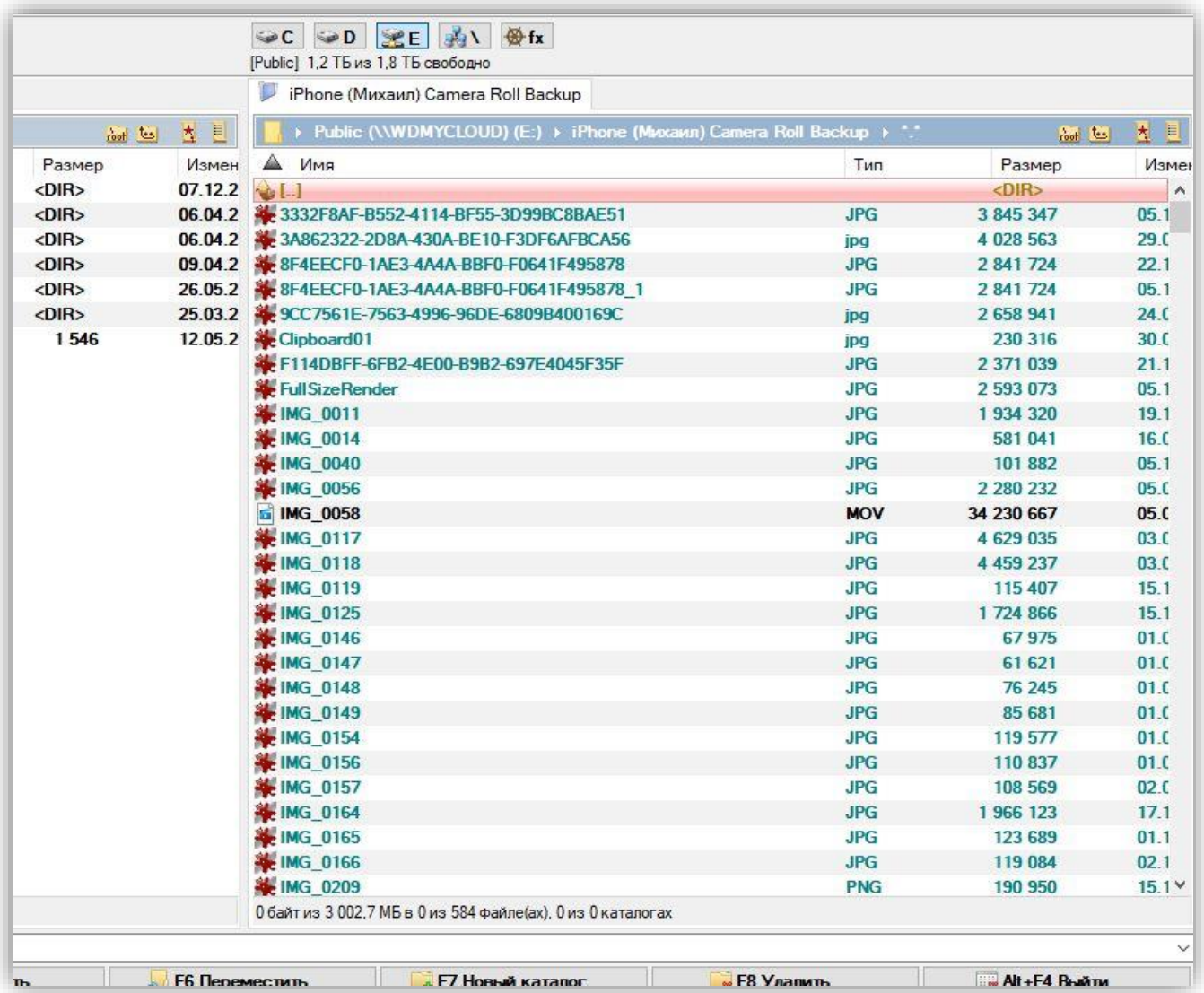
ДАННЫЕ В ЦИФРОВОМ МИРЕ

Эволюция данных и подходов к их хранению.

КАК ВСЕ НАЧИНАЛОСЬ. РУЧЕЕК ДАННЫХ.



ДАННЫЕ В ВИДЕ ФАЙЛОВ.



НУ, ЭТО ЖЕ НЕ ПРОБЛЕМА?

ДАННЫХ ВСЕ БОЛЬШЕ. РЕКА ДАННЫХ.

РЕЛЯЦИОННЫЙ ПОРЯДОК В ДАННЫХ. РЫБАЛКА ЛИНКА.

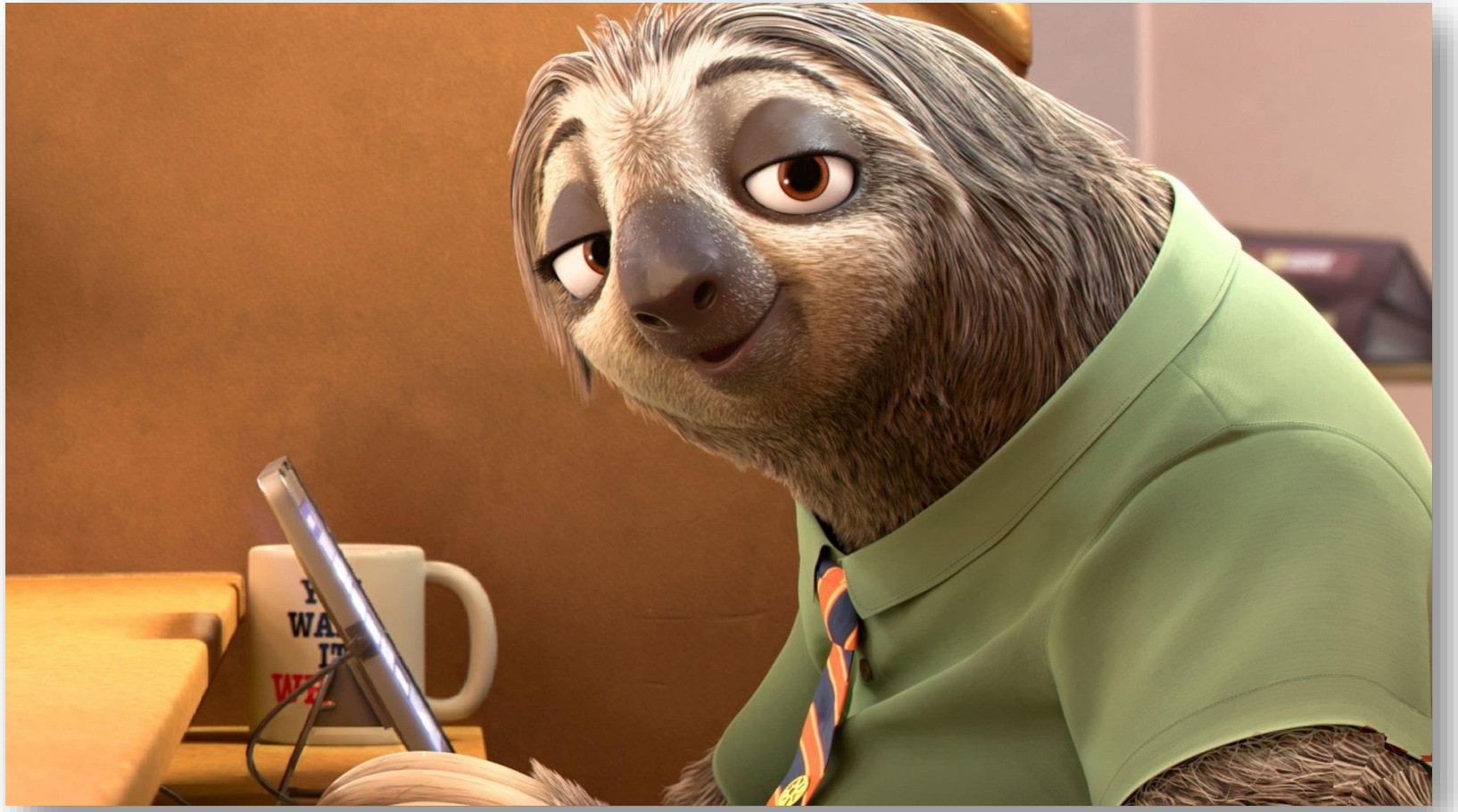
Озеро Дейа

ДАТА	КОЛ-ВО	РЫБА
17.05.2021	412	Муранья
03.07.2021	84	Сладкая рыба-пастух
03.07.2021	9	Лорд Чапу-Чапу
03.07.2021	117	Веселый карп

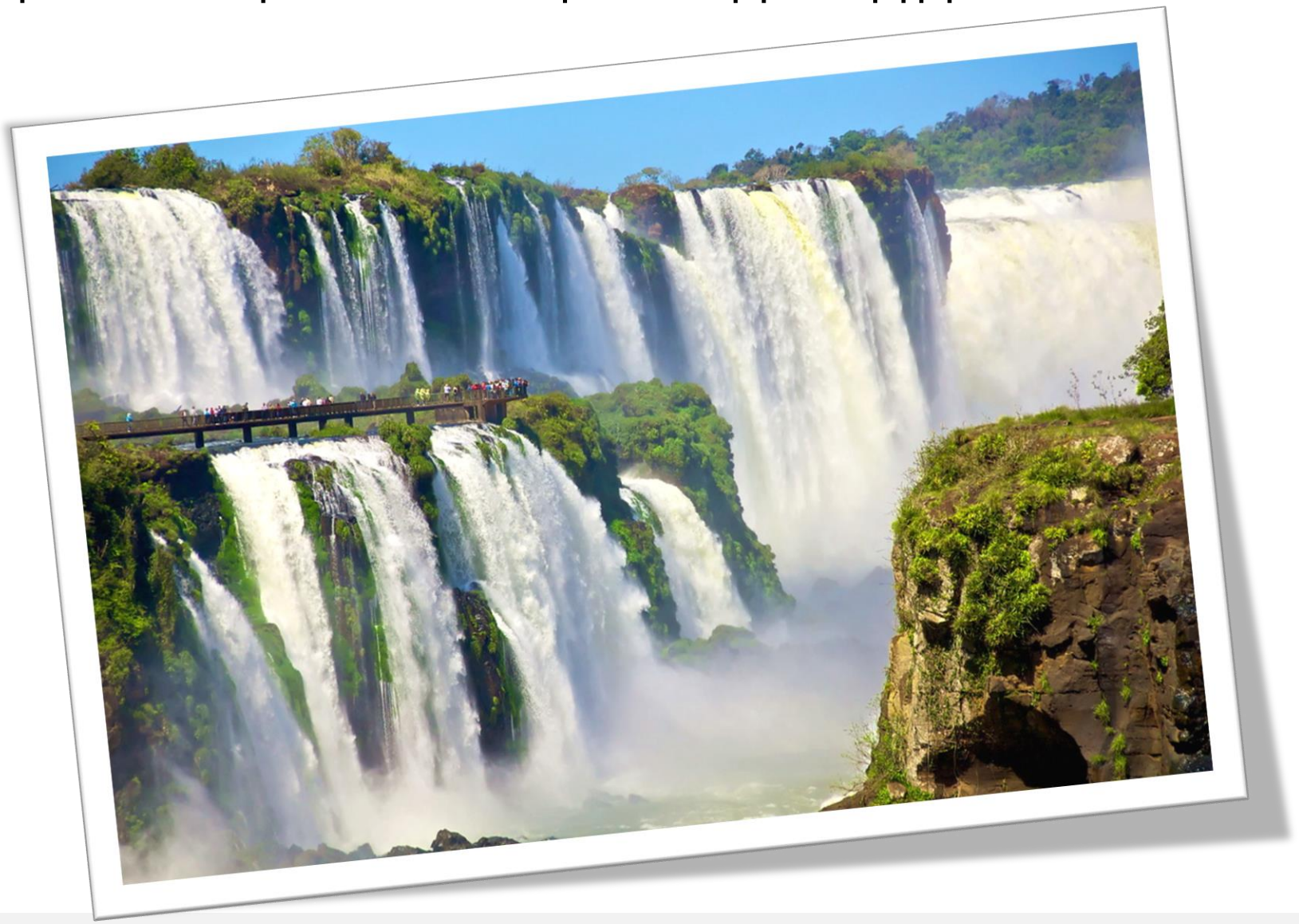
Рыба Муранья

ДАТА	КОЛ-ВО	ВОДОЕМ
17.05.2021	412	Озеро Дейа
03.07.2021	84	Озеро Джаррах
03.07.2021	9	Озеро Кора
03.07.2021	117	Озеро Нирвата

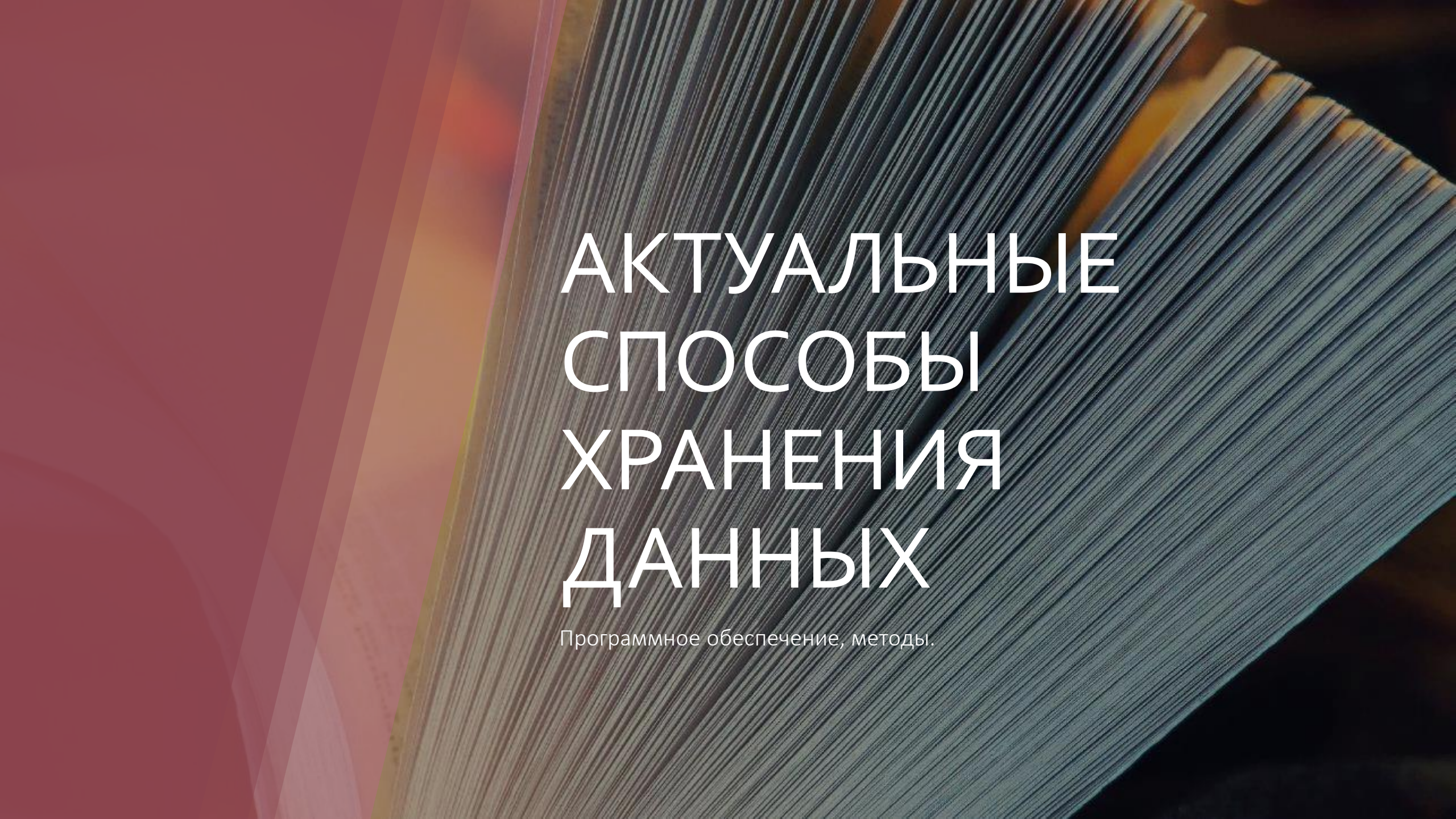
НАДЕЖНО, НО… НЕБЫСТРО.



ЭПОХА СОЦСЕТЕЙ И ЦИФРОВИЗАЦИИ. ВОДОПАД ДАННЫХ.



ТОЧКИ ПРИЛОЖЕНИЯ ДАННЫХ В VUCA-МИРЕ



АКТУАЛЬНЫЕ СПОСОБЫ ХРАНЕНИЯ ДАННЫХ

Программное обеспечение, методы.

РЕЛЯЦИОННЫЕ И ПОСТРЕЛЯЦИОННЫЕ БАЗЫ ДАННЫХ

Реляционные базы данных и хранилища данных.

Все упорядочено по строгим правилам.

Дружелюбный, похожий на естественный язык программирования (SQL).

Гарантирует правильность и целостность данных. Всегда. Не даст совершить ошибку.

Долгая подготовка, дорогие изменения.

Постреляционные базы данных и хранилища данных.

Упорядочим потом. Как-нибудь. Но это неточно.

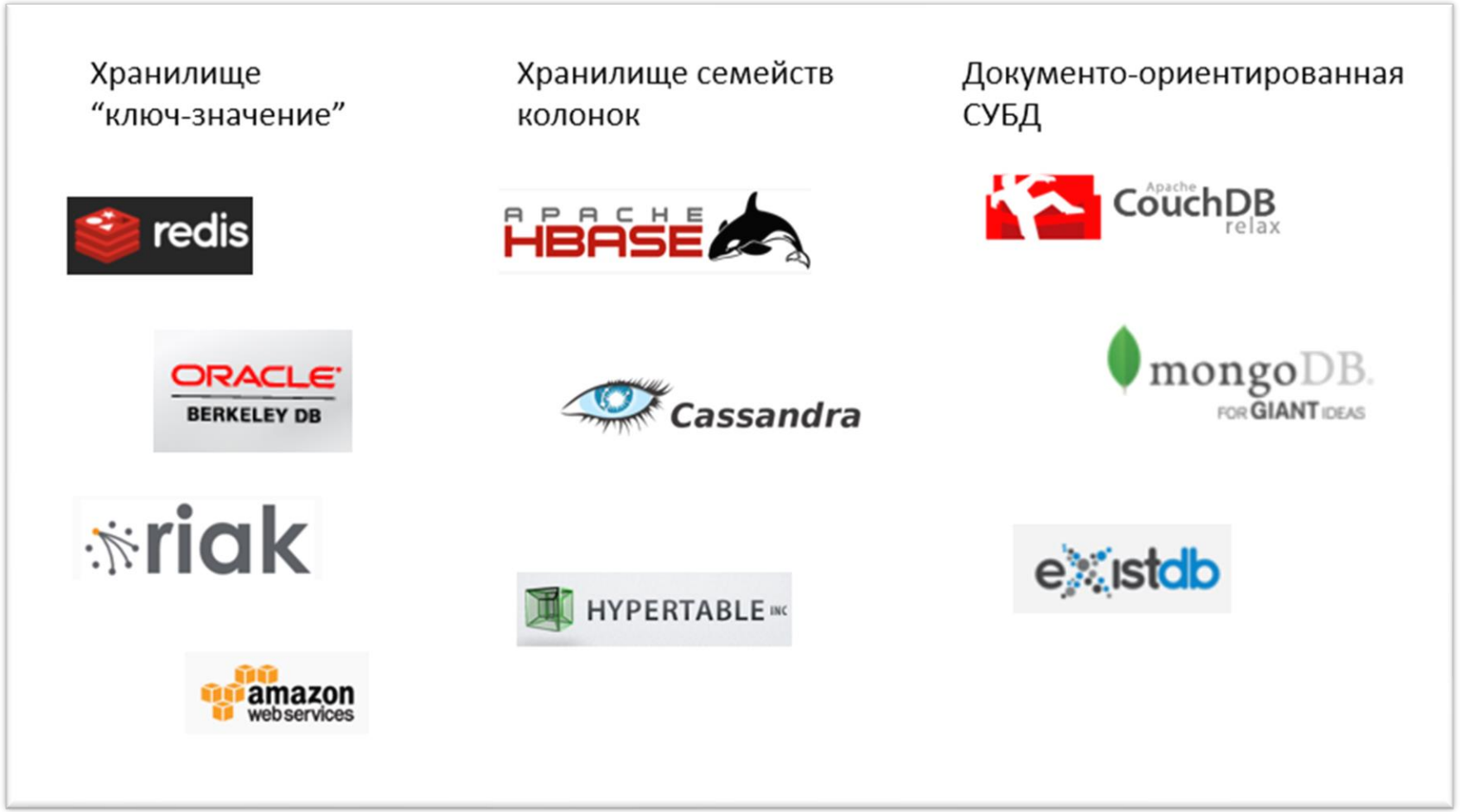
Много скобочек. Разных. Очень много! (JSON).

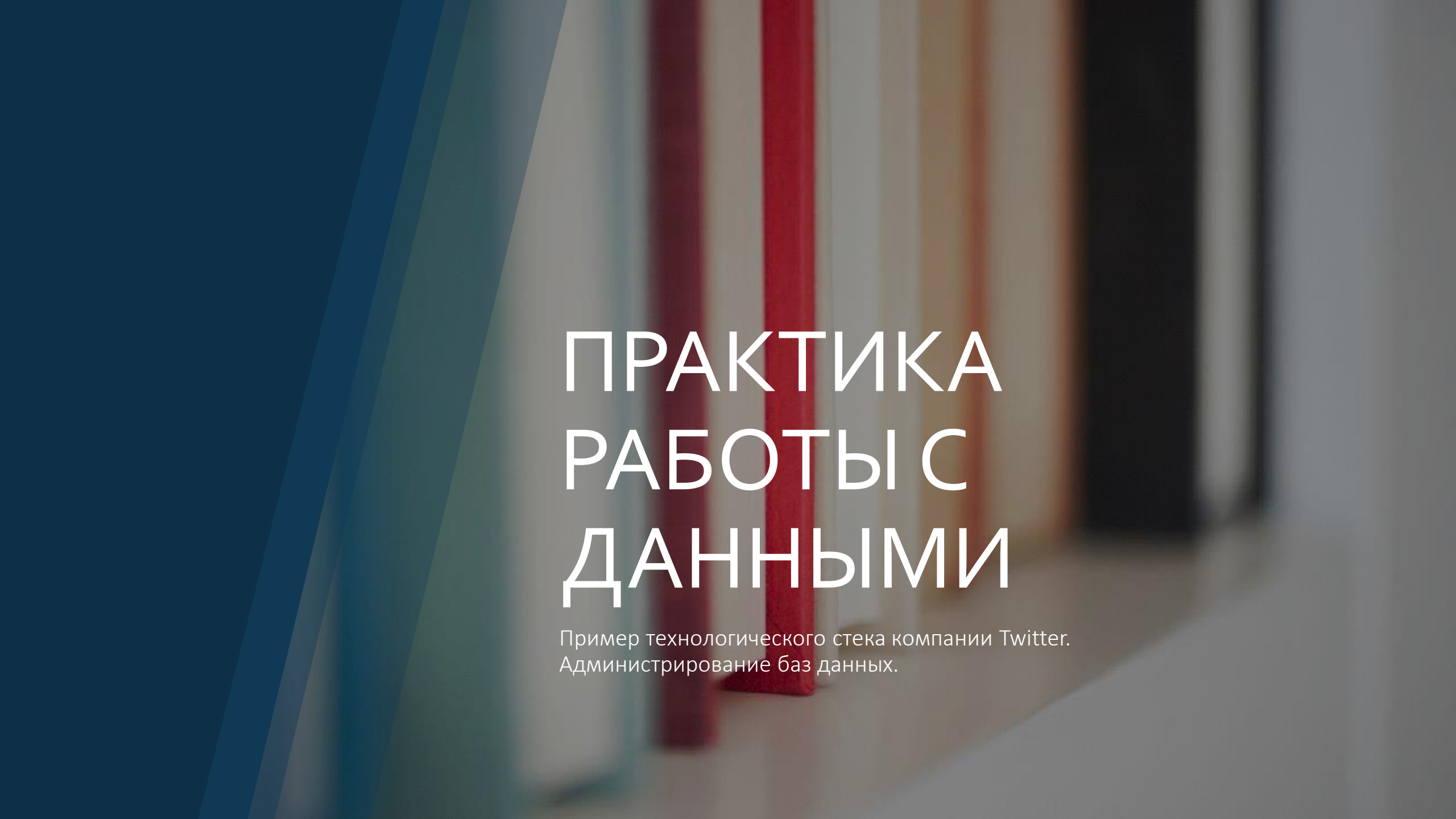
Гарантирует правильность данных. Когда-нибудь. Но это неточно.

Короткая подготовка. Моментальные изменения.

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ РЕЛЯЦИОННОЙ ПАРАДИГМЫ

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ПОСТРЕЛЯЦИОННОЙ ПАРАДИГМЫ





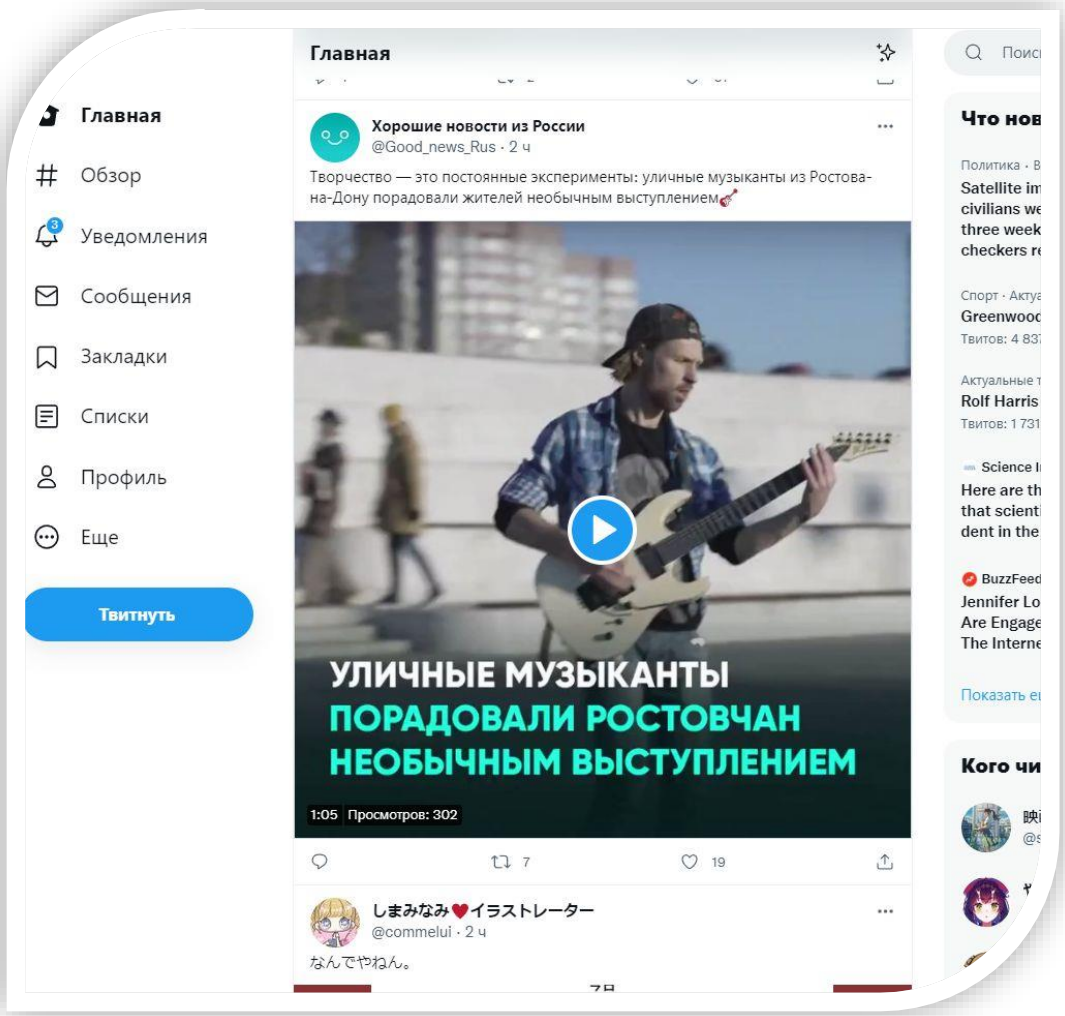
ПРАКТИКА РАБОТЫ С ДАННЫМИ

Пример технологического стека компании Twitter.
Администрирование баз данных.

ОПЫТ КОМПАНИИ TWITTER.

Десятилетний опыт построения действительно огромной и комплексной базы данных для удовлетворения нужд огромной аудитории сервиса.

ВЕБ-ПРИЛОЖЕНИЕ (СОЦИАЛЬНАЯ СЕТЬ) TWITTER

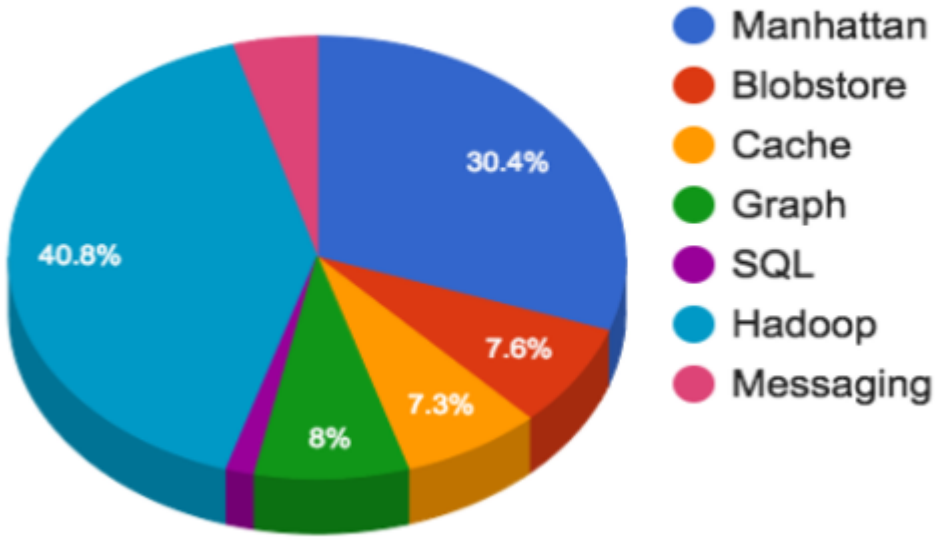


ПРАКТИЧЕСКИЕ ЗАДАЧИ, СТОЯВШИЕ ПЕРЕД СПЕЦИАЛИСТАМИ TWITTER

- За годы работы суммарный объем данных, накопленный на серверах (без мультимедиа) – около 500 петабайт (квадриллион байт, google it (c)). Где и как хранить то, что уже есть… а главное – где и как хранить то, новое, что приходит бешеными темпами?
- Обработка оперативной информации – твиты, личные сообщения, должна происходить моментально. Учитывая их количество, это десятки миллионов QPS по всему миру (в СЕКУНДУ, КАРЛ!!)
- Управление системой рекомендаций (что почитать, на кого подписаться…). Социальный граф огромного размера!
- Мультимедиа пользователей (сотни миллиардов картинок, видеороликов) – куда это девать?
- Чем кешировать метаинформацию (действия пользователей, таймлайны) для моментального доступа клиентских приложений к ней?
- Где хранить финансовую, рекламную и другую информацию строгой отчетности? Потеряется твит или личное сообщение – пользователь будет недоволен, но это не критично? Потеряется оплаченная реклама – это огромный удар по репутации компании!

https://blog.twitter.com/engineering/en_us/topics/infrastructure/2017/the-infrastructure-behind-twitter-scale

СТРУКТУРА ТЕХНОЛОГИЧЕСКОГО СТЕКА TWITTER ДЛЯ ХРАНЕНИЯ ДАННЫХ (ПОСТОЯННО ЭВОЛЮЦИОНИРУЕТ)



1. Кластеры Hadoop для вычислений и HDFS.
2. Кластеры Manhattan для всех хранилищ key-value с малой задержкой.
3. Хранилища Graph для шардированных кластеров MySQL.
4. Кластеры Blobstore для всех крупных объектов (видео, изображения, бинарные файлы...).
5. Кластеры кэширования.
6. Кластеры обмена сообщениями.
7. Реляционные хранилища (MySQL, PostgreSQL и Vertica).

https://blog.twitter.com/engineering/en_us/topics/infrastructure/2017/the-infrastructure-behind-twitter-scale

А ЧТО С ЭКСПЛУАТАЦИЕЙ ГОТОВЫХ БАЗ ДАННЫХ?

Задачи администраторов программного обеспечения баз данных.

ЗАДАЧИ АДМИНИСТРАТОРОВ БАЗ ДАННЫХ

Чтобы все работало.

Постоянная проверка данных.

Мониторинг правильности работы интерфейсов и программного обеспечения.

Автоматизация задач всем доступными способами, чтобы можно было попить кофейку и посмотреть сериальчик, пока хорошо отлаженная система работает (прохладная история про мастера PERL скриптов из комьюнити).

Чтобы все работало максимально хорошо.

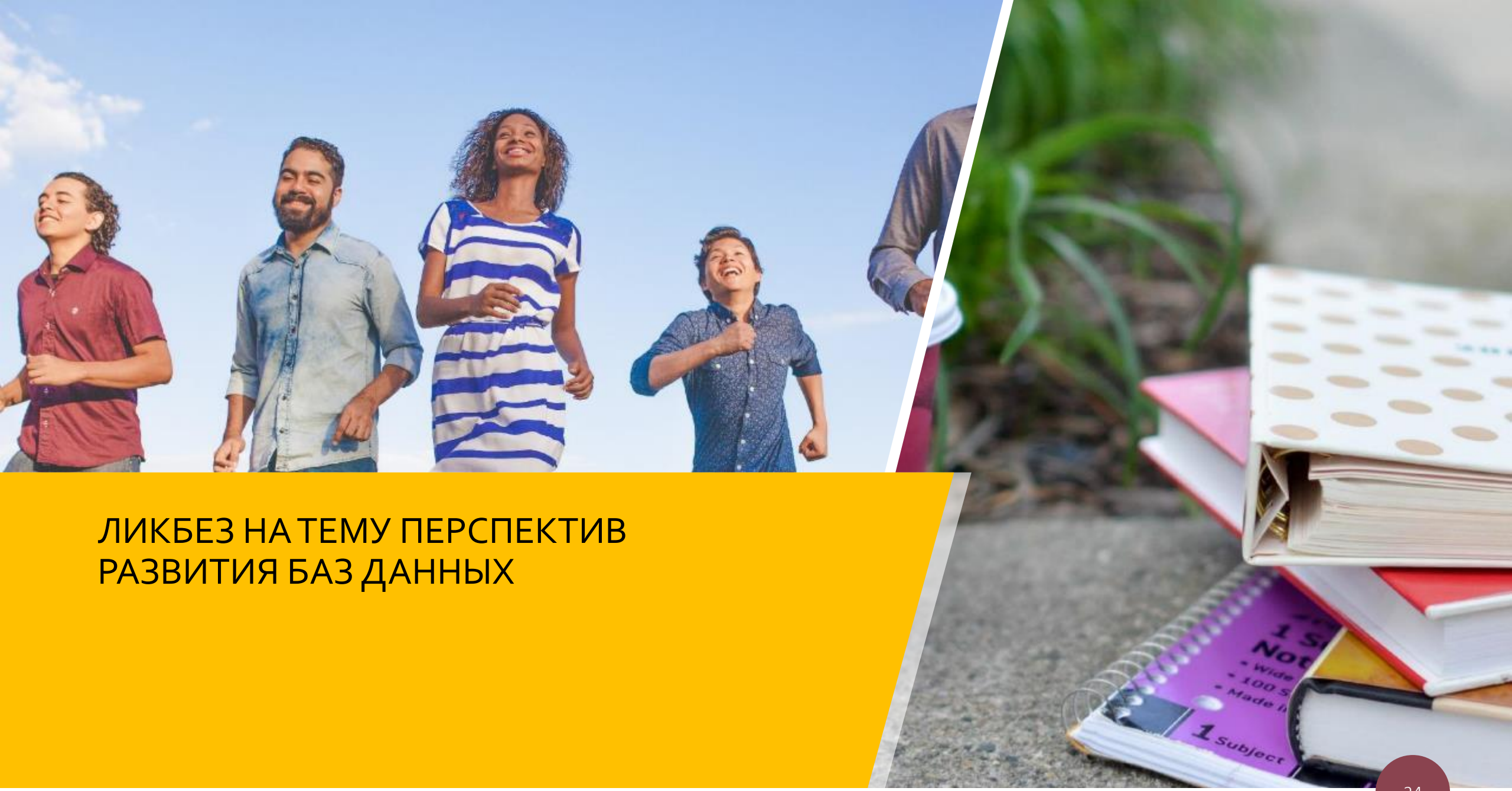
Эволюция программного обеспечения.

Эволюция программного кода.

Оптимизация работы компонентов системы хранения данных.

Чтобы никто это не сломал.

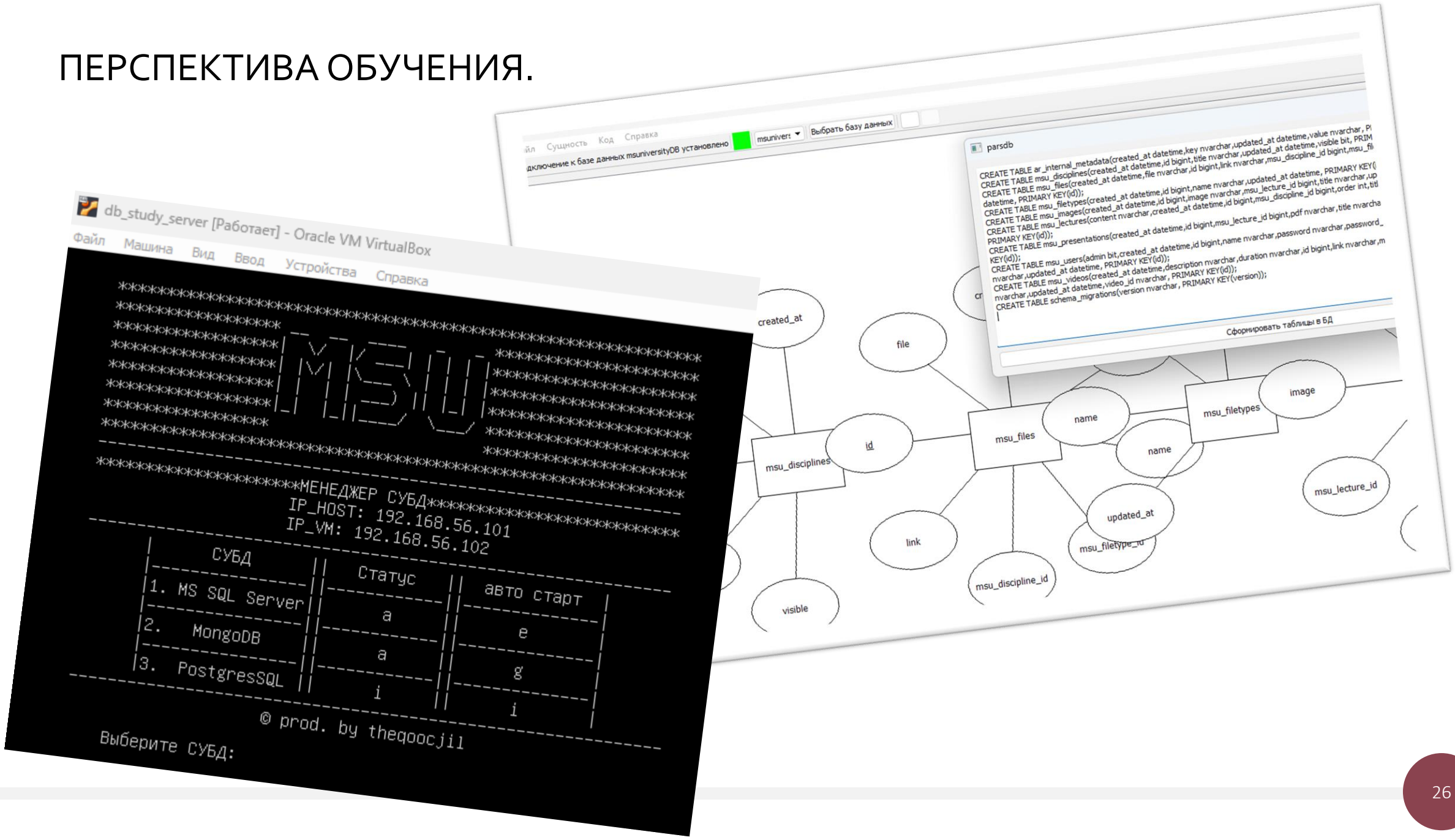
Мероприятия, связанные с постоянной защитой данных пользователей (зачастую не приводящие к положительному результату, но все-таки необходимые).



ЛИКБЕЗ НА ТЕМУ ПЕРСПЕКТИВ РАЗВИТИЯ БАЗ ДАННЫХ

ПЕРСПЕКТИВА ОБУЧЕНИЯ.

ПЕРСПЕКТИВА ОБУЧЕНИЯ.



ПЕРСПЕКТИВА КАРЬЕРЫ.

Добрый день!
Я работаю data scientist'ом в МегаФоне. В обязанности входит разработка моделей, их валидация и продуктивизация. Иногда пишу ETL. В стэк входит: PySpark, Oracle, Hive для хранения и обработки данных, Jupyter Notebook для написания пайплайнов обучения/etl, для продуктивизации - airflow, mlflow, docker
16:25

Добрый день! По поводу вашего вопроса работы с данными:
Вел документацию по проектам в Confluence, Team Foundation Server(Azure DevOps Server)
Выписывал лицензии к продукту через Microsoft SQL Server

Добрый день. Работаю в selsup.ru. Бэк почти фул на джаве, есть чуть-чуть 1С. В качестве бд преимущественно используются mysql, но есть и postgres. Liquibase в миграциях. Есть еще селфхост ydb в контейнере (прошу переписать на монгу, но пока нет), в качестве объектного хранилища minio
изменено 16:28

Добрый день! Бэкендер в Сбер, Oracle/Postgre
19:12

Добрый день, на данный момент занимаемся перепроектированием бд с переносом информации из одной базы в другую, язык PostgreSQL, субд DBeaver. Используем маппинг S2T, для построения лог модели использовали ER Assistant, MySQL Workbench и для общего доступа Drawlo. В дальнейшем будем составлять отчёты в Apache Superset.

Здравствуйте, Михаил Вячеславович. На работе саппорт ядра bss/oss системы + разработка платформы вебекс ботов. Языки Java+Python+Go, немного кубера со всеми вытекающими. База основная Oracle+SQL Developer(очень редко, как правило хватает внутренней дев странички для обращений к базе), также юзаем postgre + pgAdmin. И немного работы с s3 хранилищем(что крутится на проде не знаю, внутри используем minio), у нее ui есть из коробки. По задачам как правило починка запросов, упавших по таймауту, изучение ситуаций где какой индекс надо/не надо, чтобы потом динамически это определять + изредка дедлоки(за 2 года 1 раз встретил и то потому что на другом конце фигней занимались и сами его искусственно сделали)
17:20

Добрый день) Работа с данными - достаточно широкое понятие. Программирую, следовательно, работаю с данными 😄 Работаю в ИнтерЭВМ. Это государственная организация. Пишу и поддерживаю сервер на языке Go для базы данных на PostgreSQL и фронта на реакте. О сущности самой работы не скажу, потому что это секрет. Параллельно преподаю у нас, в мирэа, на небольшую ставку. Это ведь тоже работа с данными! Как-то так
18:55

Вечер добрый!
Занимаюсь настройкой коробочной системы налогового мониторинга. Настройка по большей части происходит через sql скрипты. Использую dbeaver для взаимодействия с бд СУБД PostgreSQL для написания скриптов vscode, git + gitlab для хранения, версионирования и ci плюс для централизованного наката/отката и контроля версий в базе используем liquibase
19:58

Тогда Сбер Решения - инженер по разработке.
Поддерживаю внутренний продукт системы бух учета, а конкретно модуль расчета зп) Обновляю старую или пишу новую логику для отчетности. Поэтому почти всегда обращаюсь к бд.
Язык - C#. Пользуемся MSQL-Server. Запросы обычно строкой в коде пишу и потом через ado.net они на сервер отправляются, в части проекта пользуюсь EFCore.
12:41

Добрый день! Не уверен, что мой стек будет полезен для вашей статистики, но все же решил написать.
Работаю разработчиком в ПАО Группа Астра, в продукте ALD Pro. Занимаюсь созданием нового функционала.
Для работы с данными использую:
* протокол LDAP,
* Postgres.
изменено 11:55

Субд - Clickhouse
Занимаюсь созданием отчетов для бизнеса. А именно - приходит запрос от заказчика, например, показать графики изменения цен на определенные товары в определенный период и с такими-то условиями. Моя задача создать витрину данных(по сути просто таблицу с нужными агрегированными данными), а затем при помощи apache superset преобразовать их в красивые графики с фильтрами и т.д. Также в apache airflow надо написать даг на питоне, который будет своевременно оьновлять данные в этом отчете
20:02

СПАСИБО

Смирнов Михаил

mikhaelsmirnov@gmail.com

http://msuniversity.ru