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.

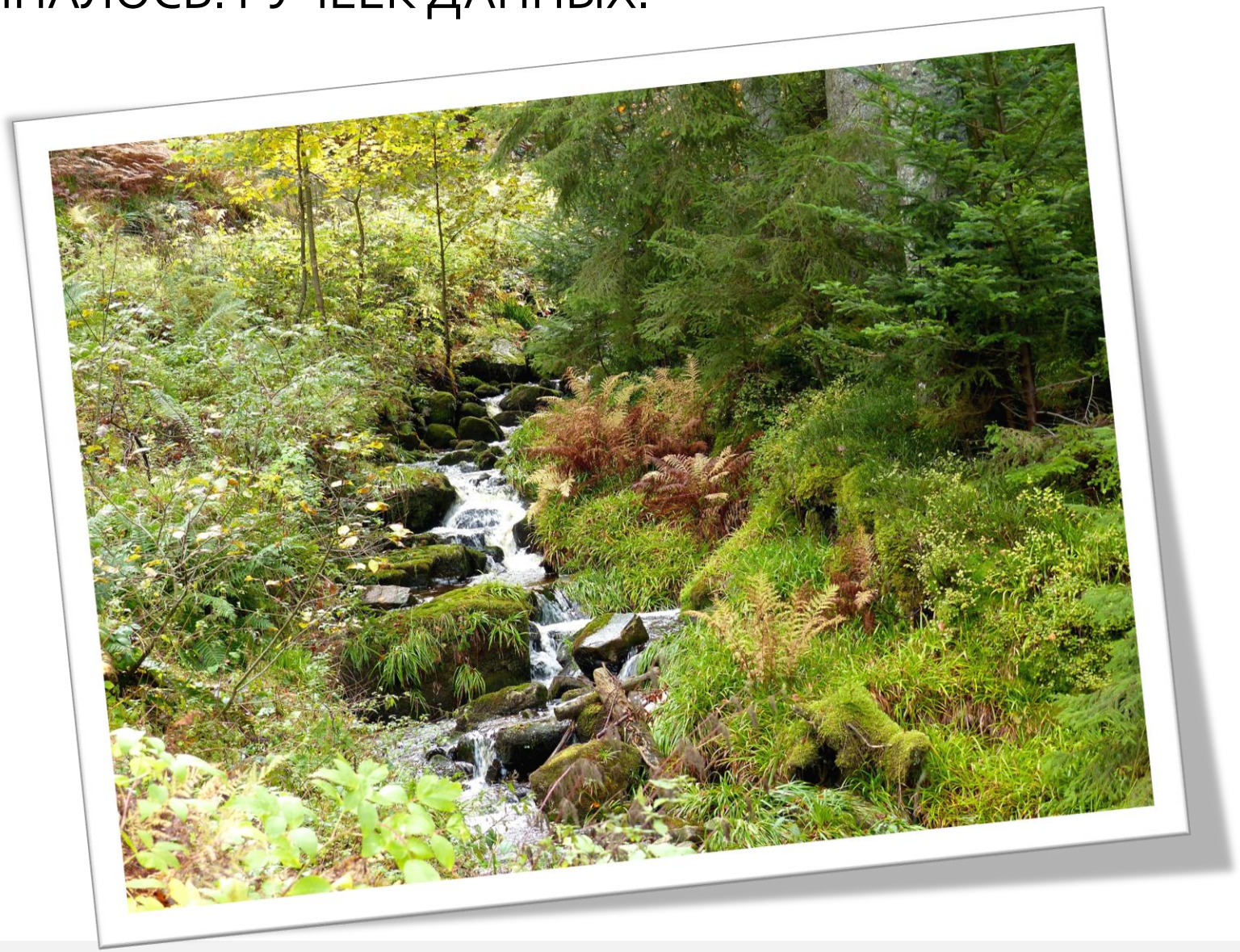
Программное обеспечение, фреймворки, технологии, главной задачей которых является обеспечение возможности безопасного хранения данных.

Их ОЧЕНЬ много.

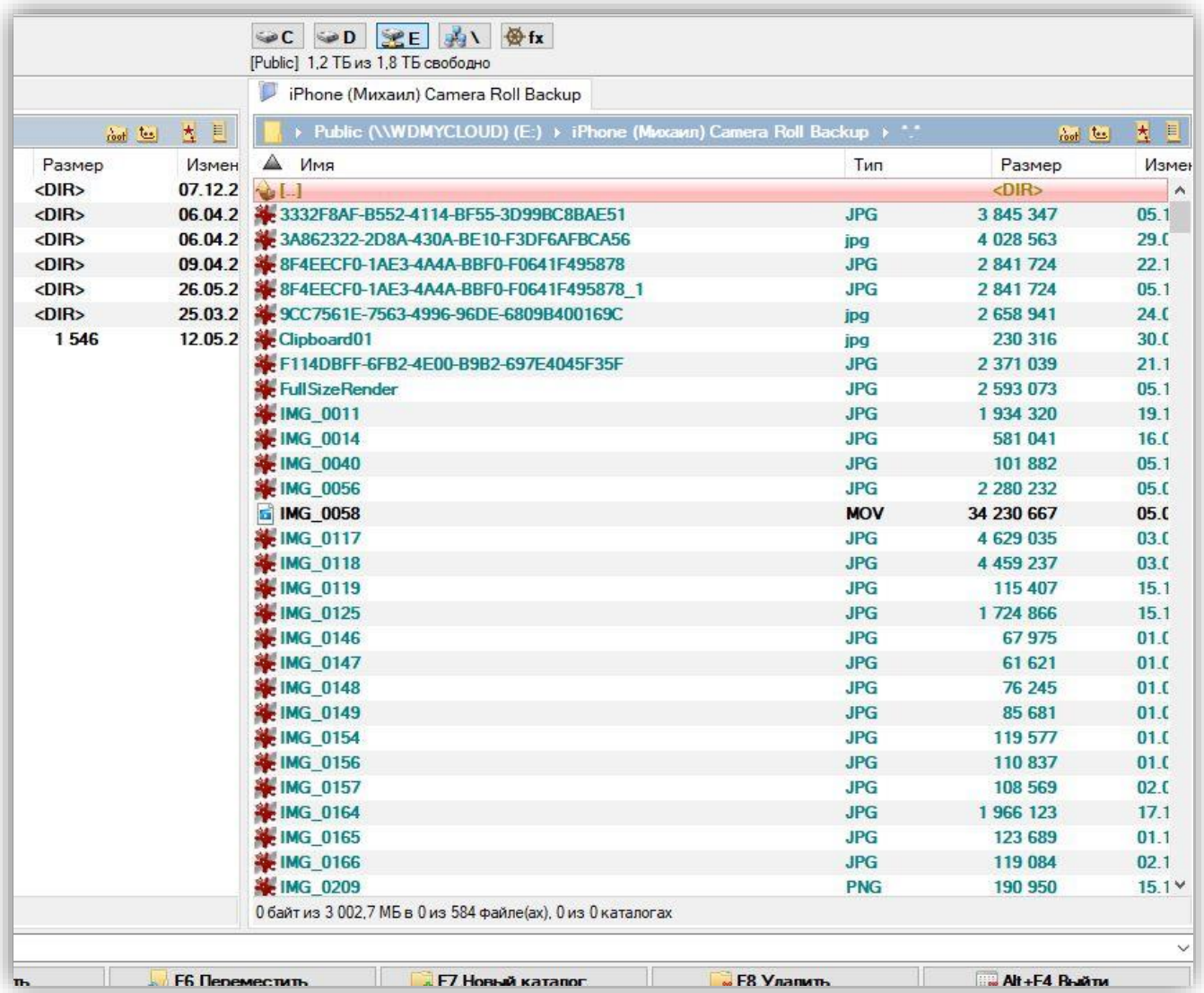
ДАННЫЕ В ЦИФРОВОМ МИРЕ

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.

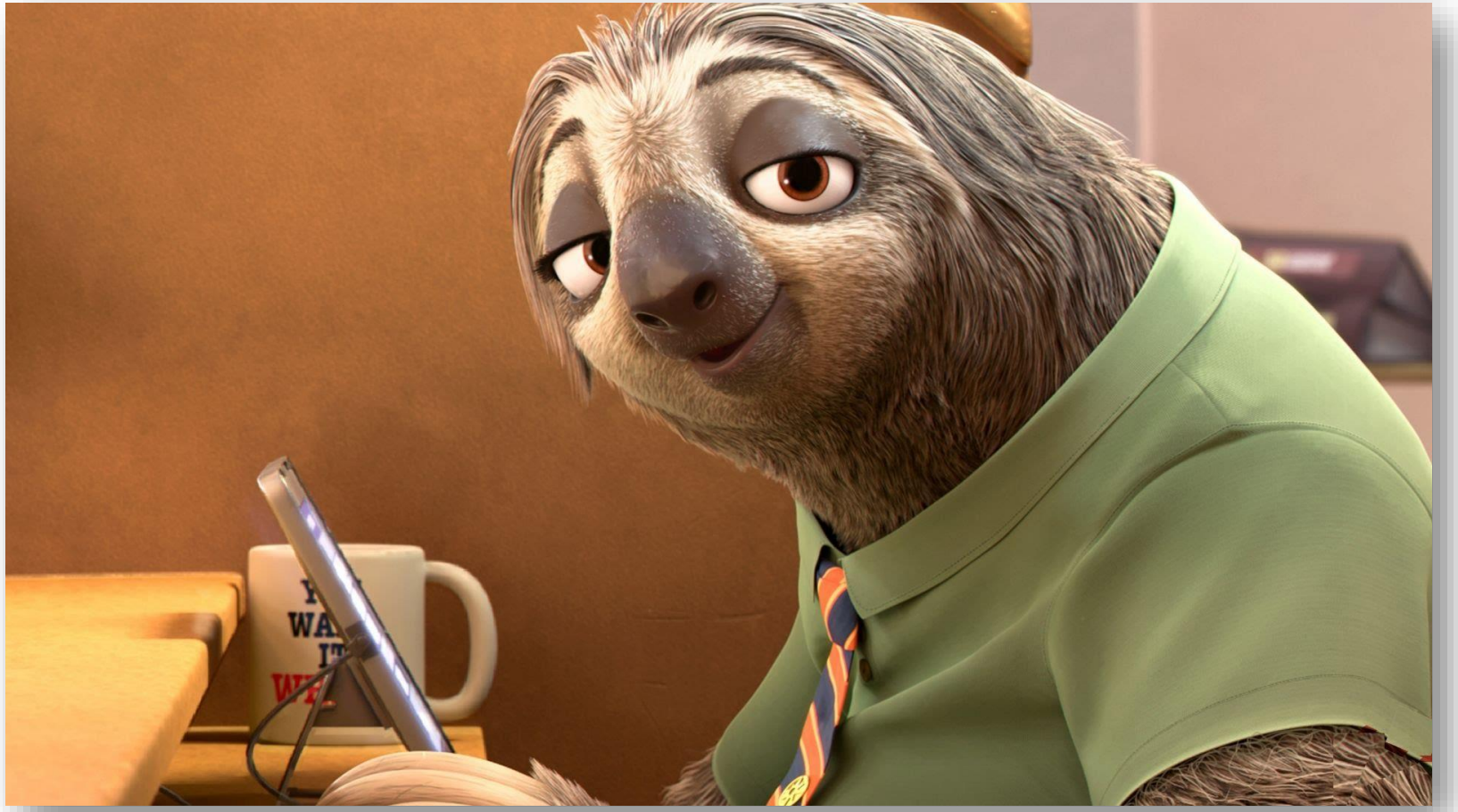
Озеро Дейа

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

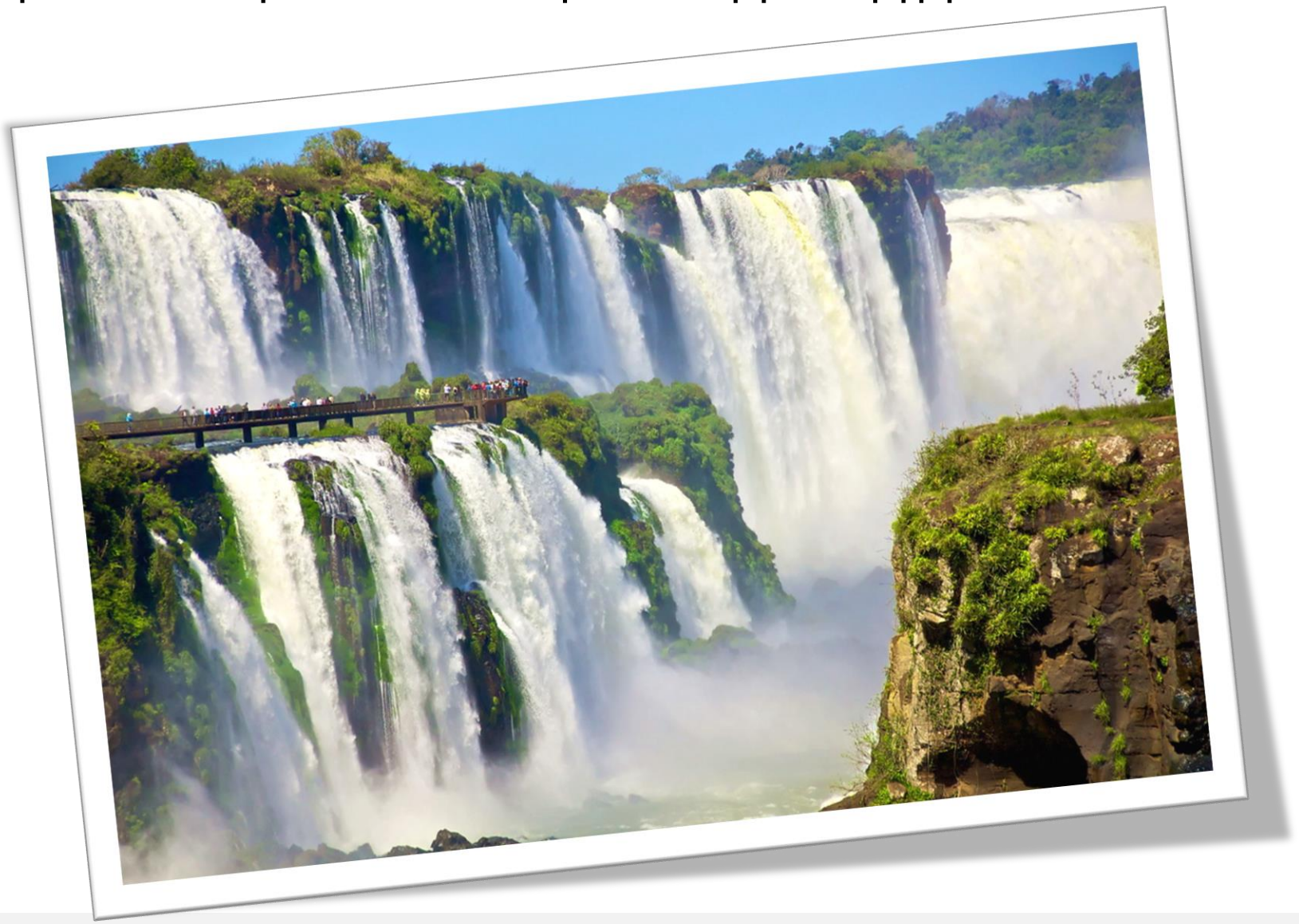
Рыба Муранья

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

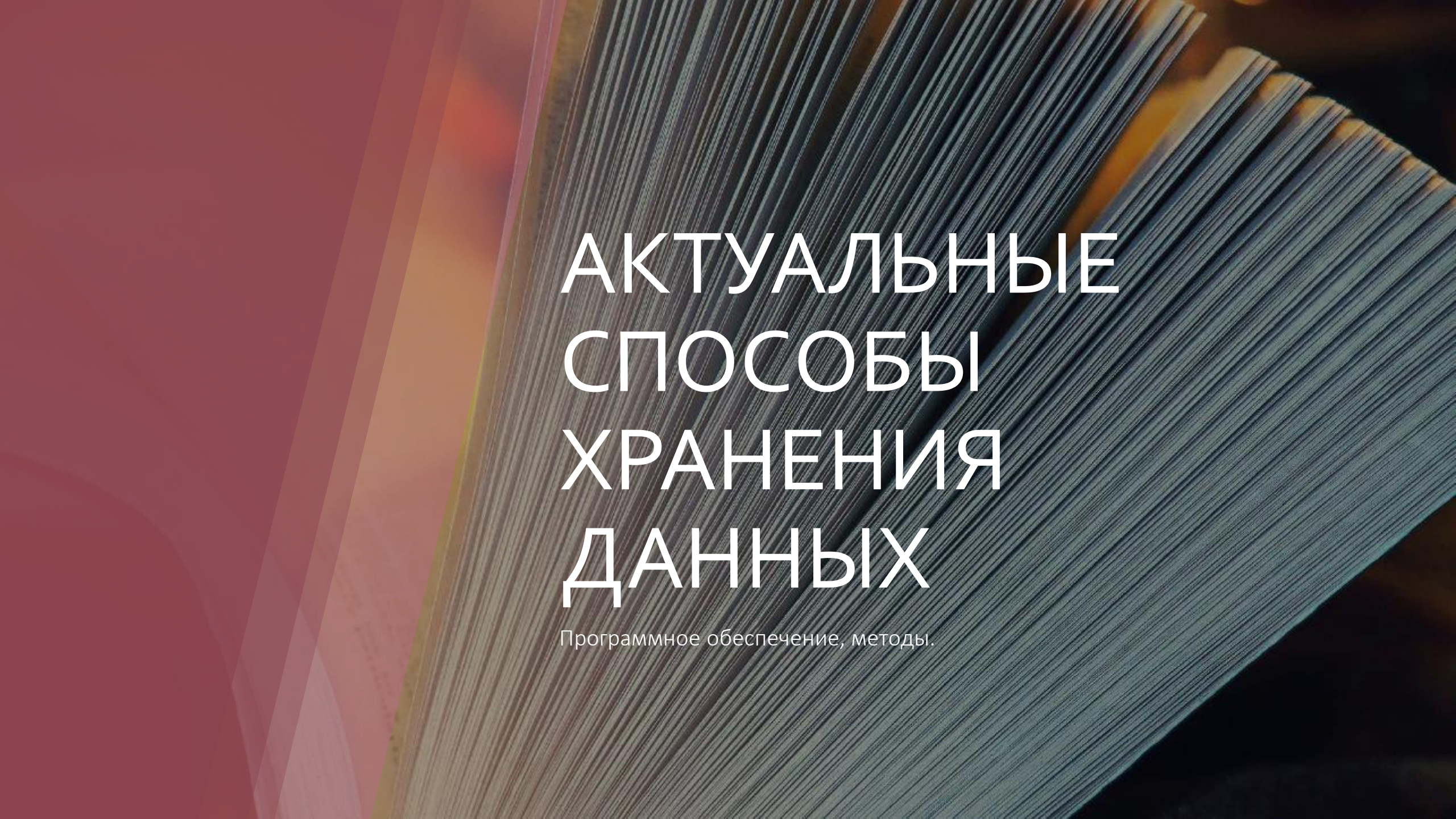
НАДЕЖНО, НО… НЕБЫСТРО.



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.

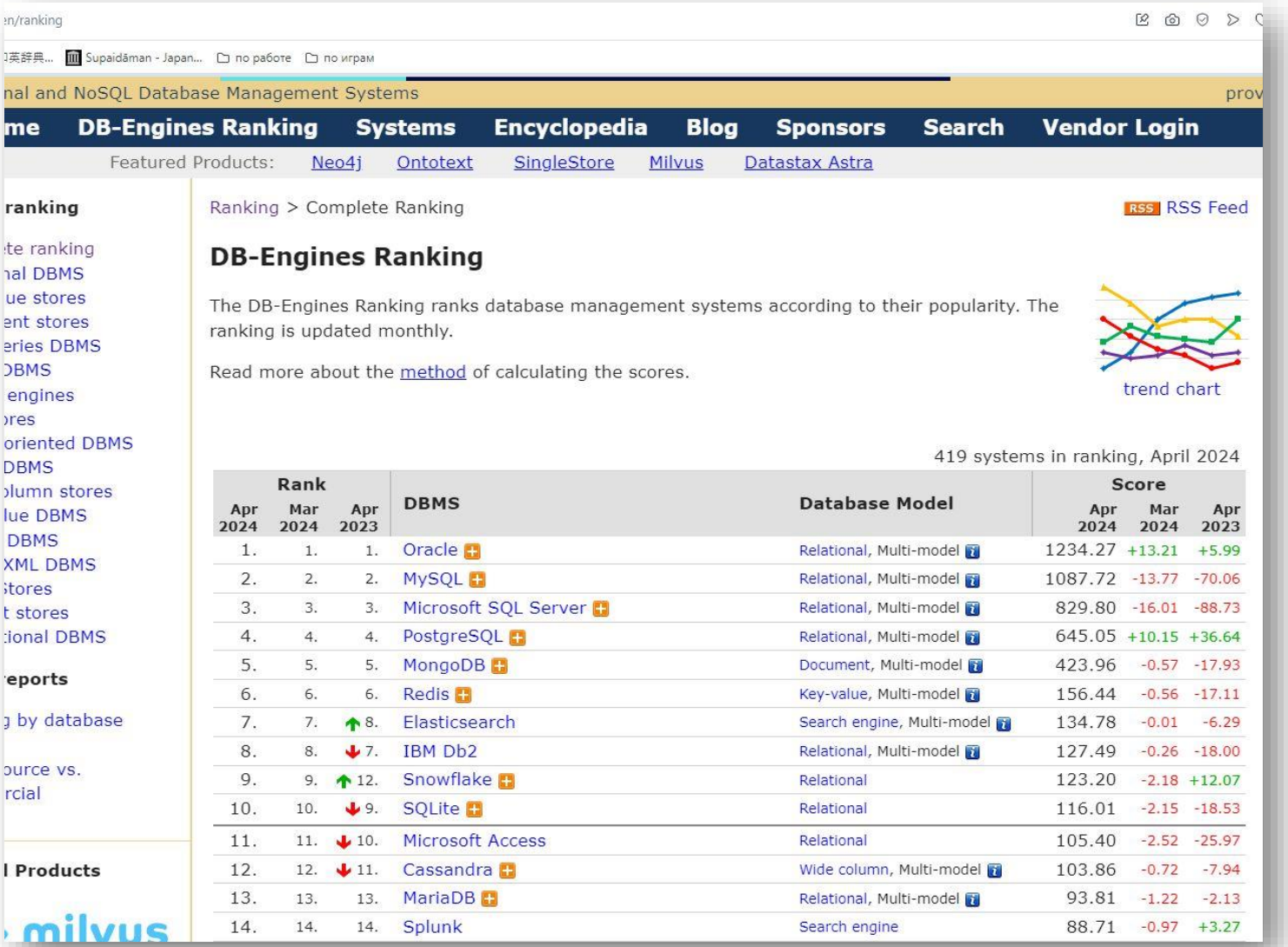
Упорядочим потом. Как-нибудь. Но это неточно.

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

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

Короткая подготовка. Моментальные изменения.

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ БАЗ ДАННЫХ



nal and NoSQL Database Management Systems

me | DB-Engines Ranking | Systems | Encyclopedia | Blog | Sponsors | Search | Vendor Login

Featured Products: Neo4j Ontotext SingleStore Milvus Datastax Astra

Ranking > Complete Ranking

RSS RSS Feed

DB-Engines Ranking

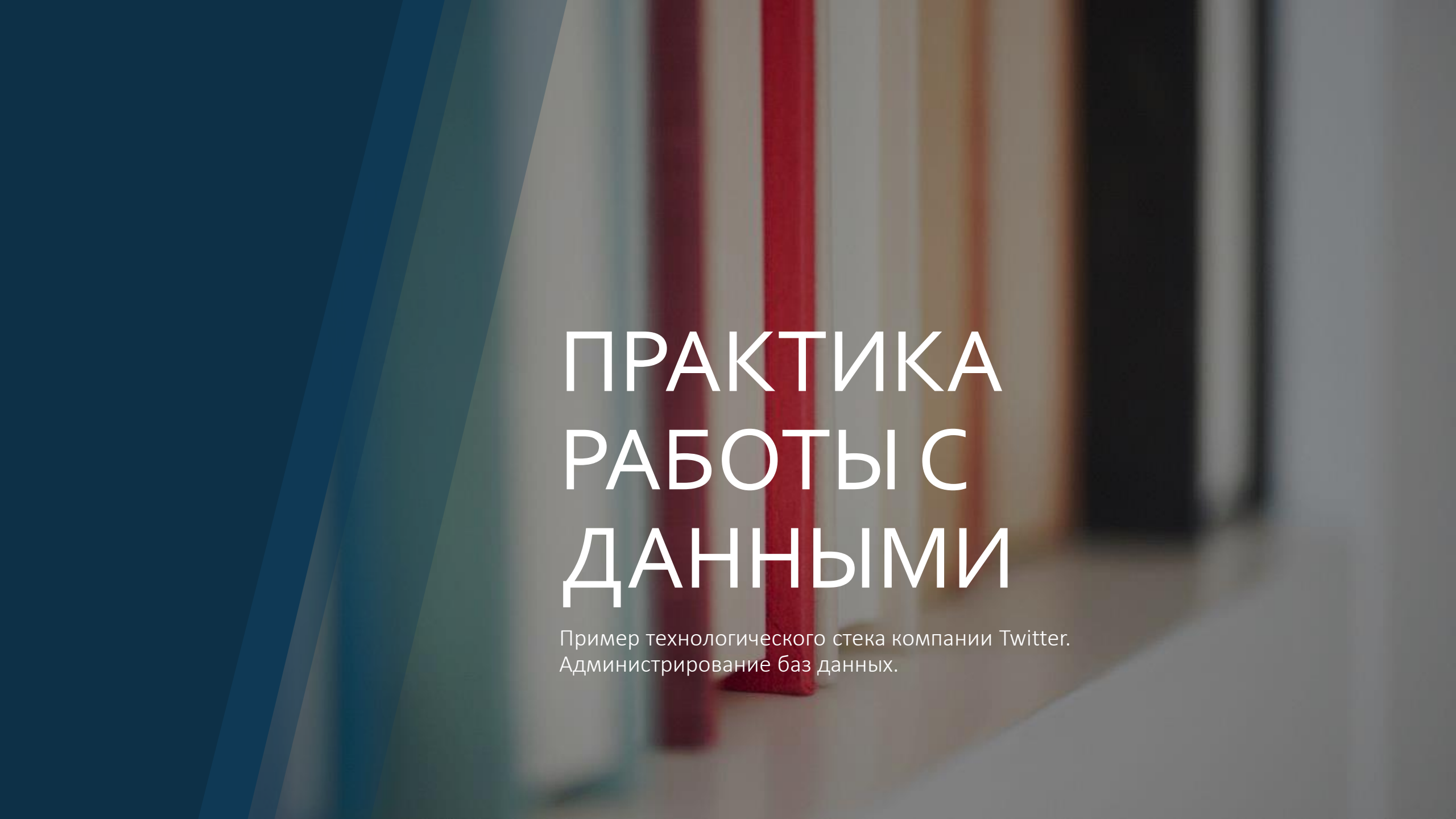
The DB-Engines Ranking ranks database management systems according to their popularity. The ranking is updated monthly.

Read more about the method of calculating the scores.

trend chart

419 systems in ranking, April 2024

Rank Apr 2024	Rank Mar 2024	Rank Apr 2023	DBMS	Database Model	Score Apr 2024	Score Mar 2024	Score Apr 2023
1.	1.	1.	Oracle	Relational, Multi-model	1234.27	+13.21	+5.99
2.	2.	2.	MySQL	Relational, Multi-model	1087.72	-13.77	-70.06
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	829.80	-16.01	-88.73
4.	4.	4.	PostgreSQL	Relational, Multi-model	645.05	+10.15	+36.64
5.	5.	5.	MongoDB	Document, Multi-model	423.96	-0.57	-17.93
6.	6.	6.	Redis	Key-value, Multi-model	156.44	-0.56	-17.11
7.	7.	↑ 8.	Elasticsearch	Search engine, Multi-model	134.78	-0.01	-6.29
8.	8.	↓ 7.	IBM Db2	Relational, Multi-model	127.49	-0.26	-18.00
9.	9.	↑ 12.	Snowflake	Relational	123.20	-2.18	+12.07
10.	10.	↓ 9.	SQLite	Relational	116.01	-2.15	-18.53
11.	11.	↓ 10.	Microsoft Access	Relational	105.40	-2.52	-25.97
12.	12.	↓ 11.	Cassandra	Wide column, Multi-model	103.86	-0.72	-7.94
13.	13.	13.	MariaDB	Relational, Multi-model	93.81	-1.22	-2.13
14.	14.	14.	Splunk	Search engine	88.71	-0.97	+3.27



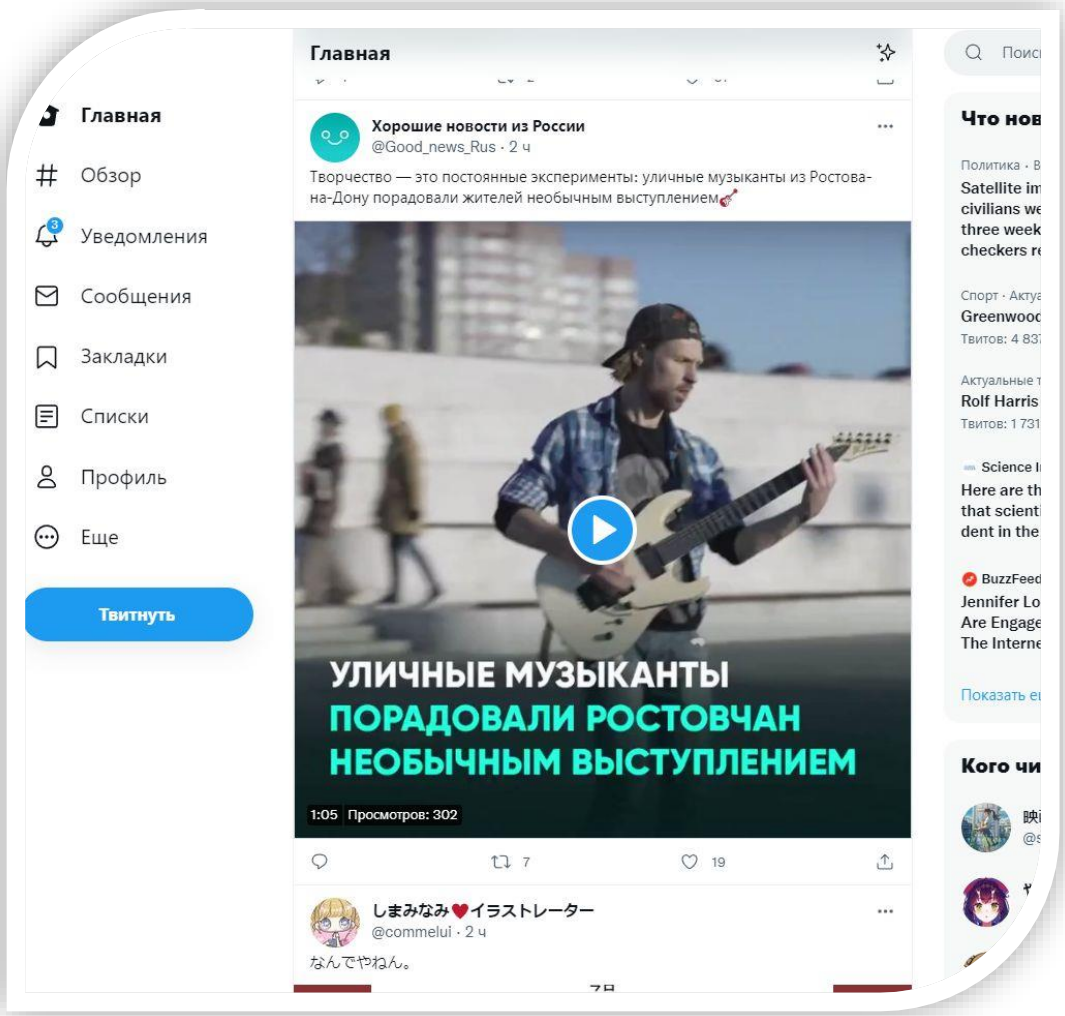
ПРАКТИКА РАБОТЫ С ДАННЫМИ

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

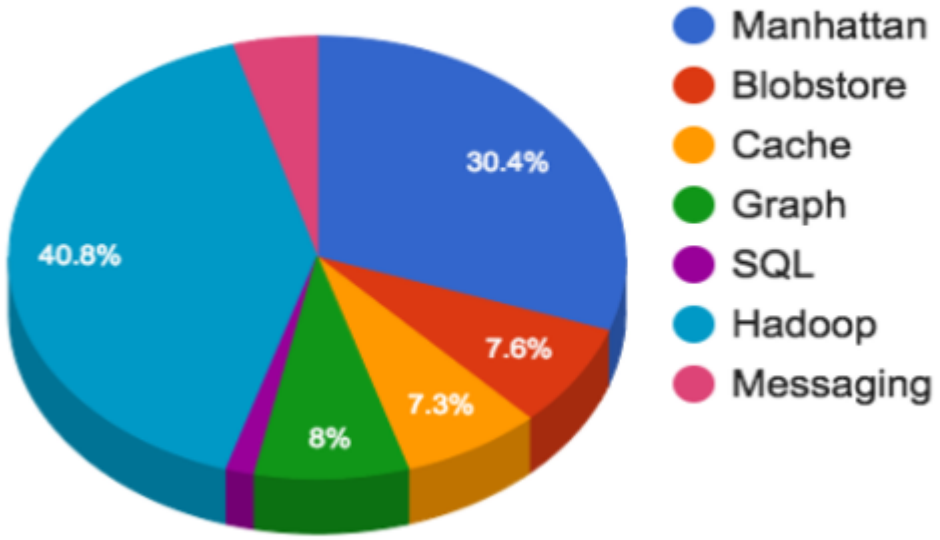


ПРАКТИЧЕСКИЕ ЗАДАЧИ, СТОЯВШИЕ ПЕРЕД СПЕЦИАЛИСТАМИ TWITTER

- 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.

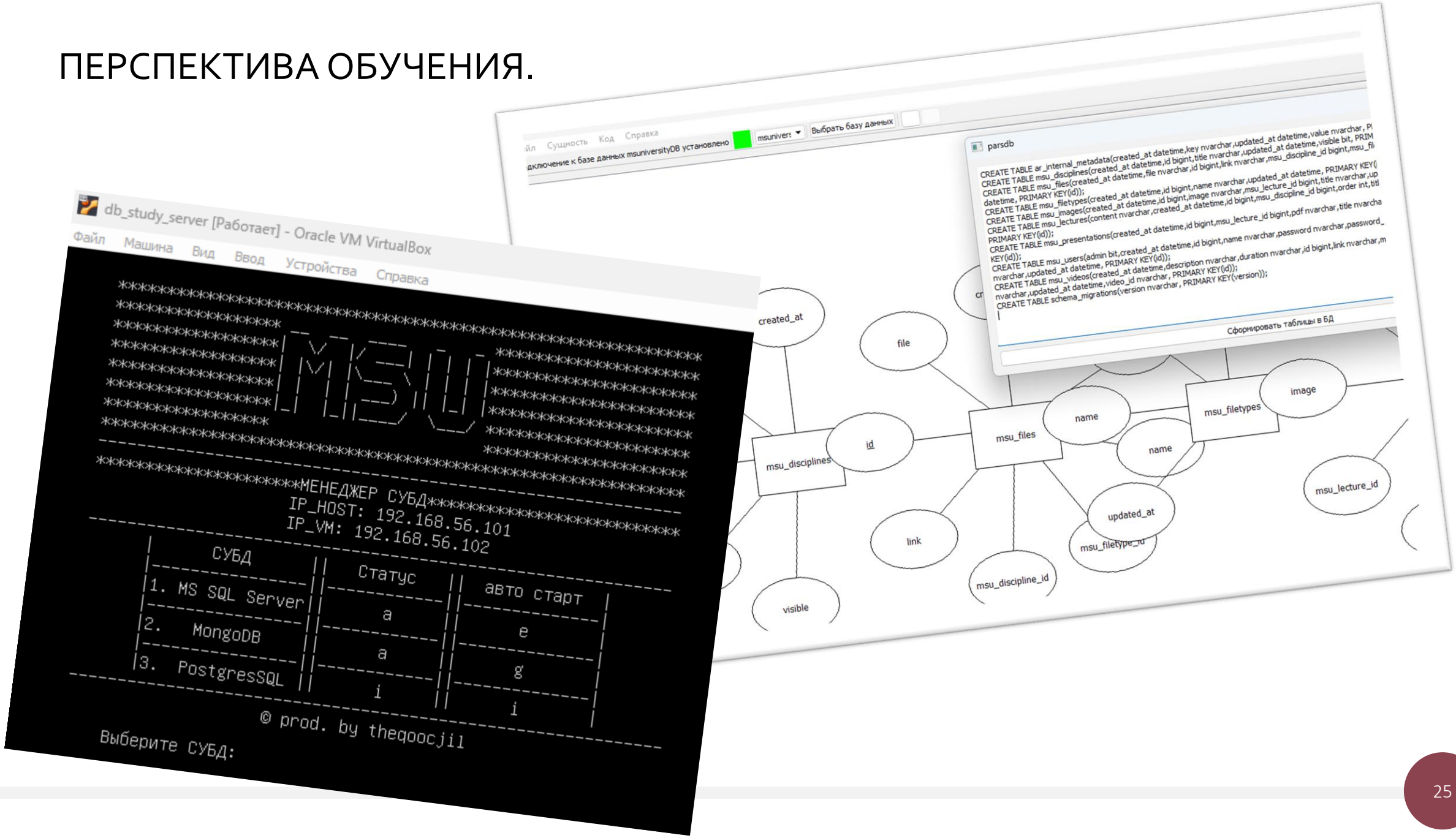
Чтобы никто это не сломал.

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

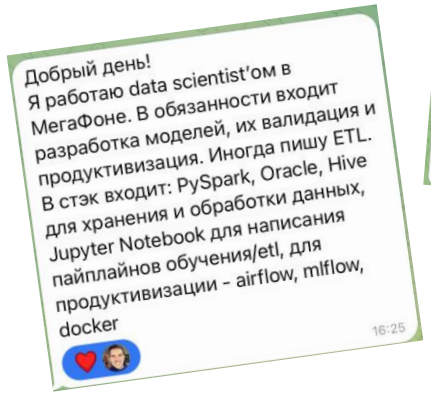
ЛИКБЕЗ НА ТЕМУ ПЕРСПЕКТИВ РАЗВИТИЯ БАЗ ДАННЫХ

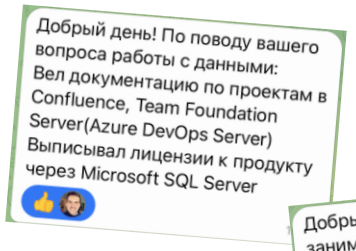
ПЕРСПЕКТИВА ОБУЧЕНИЯ.

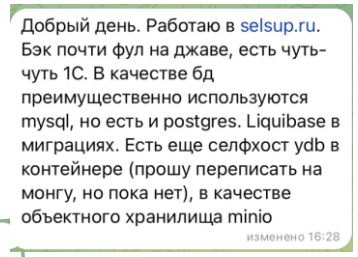
ПЕРСПЕКТИВА ОБУЧЕНИЯ.

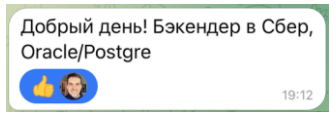


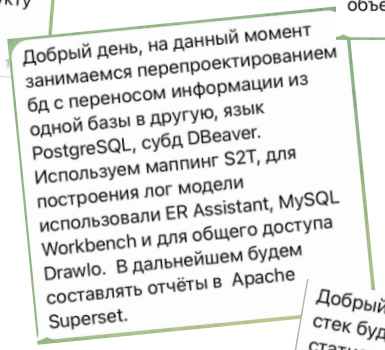
ПЕРСПЕКТИВА КАРЬЕРЫ.

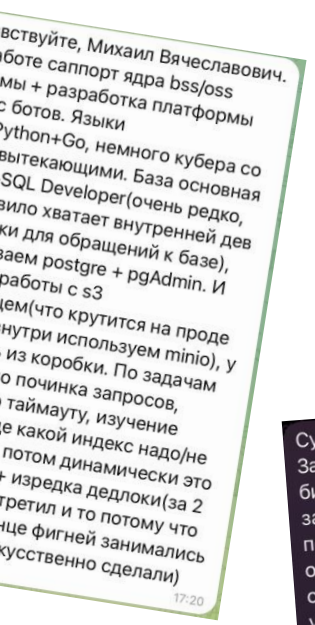


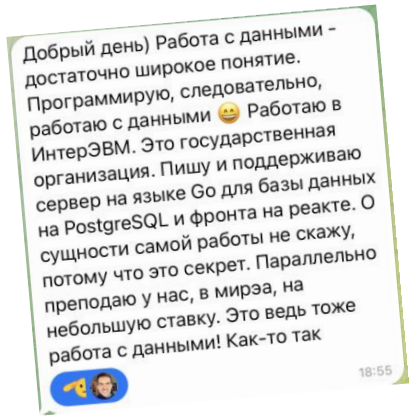


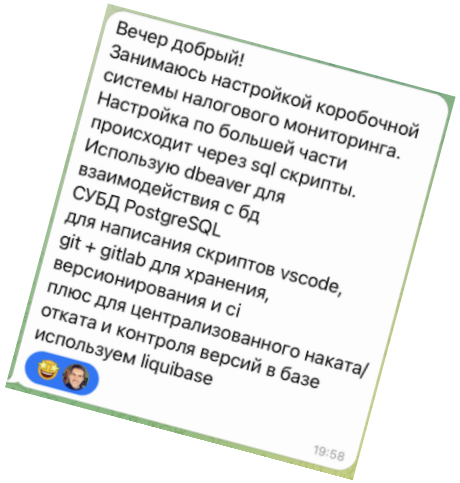


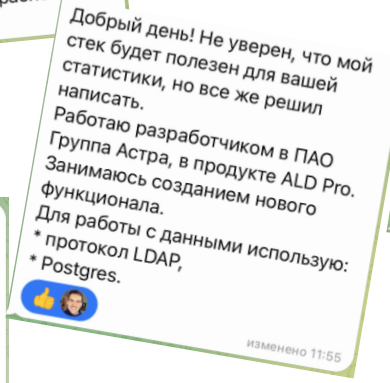


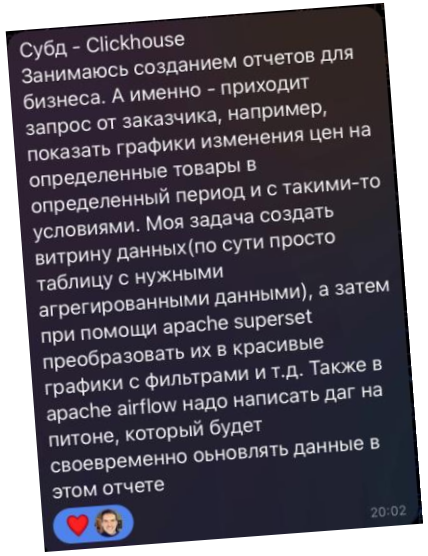












СПАСИБО

Смирнов Михаил

mikhaelsmirnov@gmail.com

http://msuniversity.ru