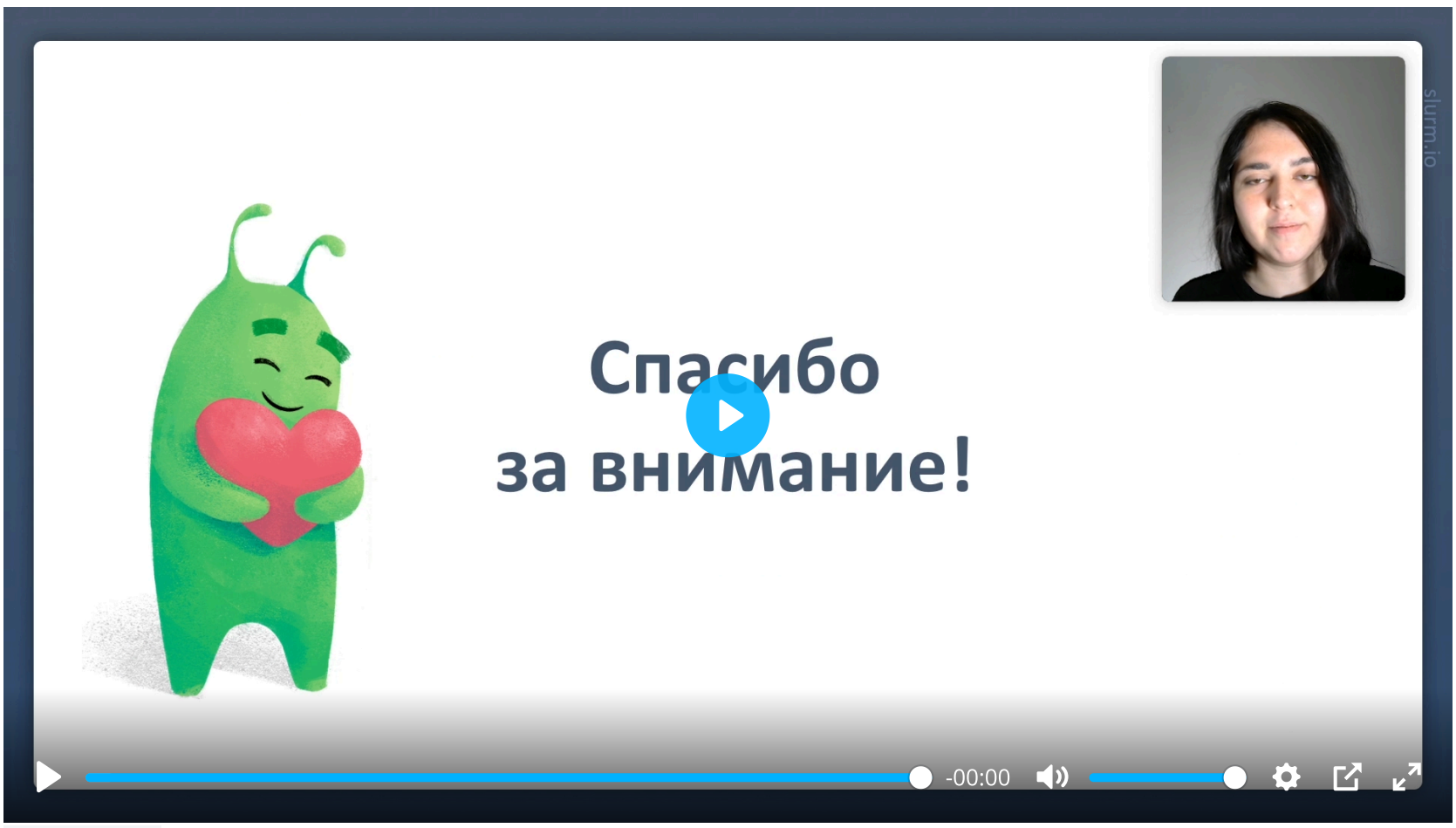
1.5 Инженерия данных. Профессия инженера данных



Презентация

Текстовая расшифровка видео:

ИНЖЕНЕРИЯ ДАННЫХ. ПРОФЕССИЯ ИНЖЕНЕРА ДАННЫХ

Добрый день, сегодня мы поговорим об инженерии и о профессии инженера данных.

План:

1. Проблематика Big Data-проекта;

2. Задачи инженерии данных;

3. Типичная команда Big Data-проекта;

4. Обязанности инженера данных, его умения и навыки.

Проблематика Big Data

Big Data-проекта отличается от проектов, в которых нет больших данных, что подразумевает под собой некоторые принципы, которым необходимо следовать:

1. Volume;
2. Velocity;
3. Variety;
4. Veracity;
5. Value.

Иногда к ним добавляются еще два принципа:

1. Variability;
2. Visualization.

Volume

Volume – объем данных:

- Растет по экспоненте;
- Постоянно обновляются, генерируются новые данные.

Velocity

Velocity – скорость прироста данных:

- Необходимо справляться с огромной скоростью создания данных;
- Необходимо анализировать данные в режиме реального времени.

Variety

Variety – разнообразие данных:

- Проекты Big Data включают данные в самых разных форматах;
- Каждый из типов данных требует различные типы анализа и подходящие инструменты.

Veracity

Veracity – достоверность данных:

- Любой анализ данных бесполезен, если данные недостоверны;
- Неточность данных может привести к неправильным решениям и потерям для бизнеса.

Value

Value – ценность данных:

- Необходимо извлечь максимум пользы из результатов анализа больших данных;
- Важно использовать для принятия решений идеи, полученные из анализа данных.

Variability

Variability – изменчивость данных:

- Значение одних и тех же данных может различаться в зависимости от контекста;
- Алгоритмы должны быть в состоянии понять контекст и расшифровать точное значение слова в этом контексте.

Visualization

Visualization – визуализация данных:

- Визуализация делает большие данные доступными для человеческого восприятия;
- Визуализация больших объемов сложных данных понятнее для человека, чем электронные таблицы и отчеты.

Выводы:

Data Engenieer должен учитывать в своей работе следующее:

- Данные объемные (объем);
- Данные постоянно прирастают (скорость прироста данных);
- Данные могут быть разнообразными (разнообразие данных);

- Данные могут иметь разные значения для разных задач (изменчивость данных);
- Данные должны быть достоверными.

Задачи инженерии данных:

Задачи инженерии данных следующие:

- Получение и интеграция данных из информационных систем;
- Извлечение данных из устройств;
- Организация удобного хранения данных;
- Проверка полноты и корректности данных;
- Оценка качества полученных данных;
- Создание инфраструктуры для потоковой обработки машинно-генерируемых и человеко-генерируемых данных;
- Объединение в представления для конечного пользователя;
- Обеспечение последующего доступа к данным.

Типичная команда Big Data-проекта

Вначале приходят сырые необработанные данные; **Data Engenieer** (инженер данных) их преобразовывает в пригодный для использования вид и хранит в storage, куда имеют доступ **Data Scientist** (специалист по данным) и **Data Analyst** (дата аналитик); Data Scientist (специалист по данным) прогнозирует будущее по перспективе, а Data Analyst (дата аналитик) генерирует новые идеи из этих данных; **BI Analyst** (BI аналитик) предоставляет эти данные бизнесу и убеждает в принятии решения, которое позволит бизнесу развиваться.

Обязанности Data Engenieer

Обязанности Data Engenieer следующие:

- Отвечать за извлечение, интеграцию и объединение данных из разрозненных источников;
- Производить очистку, преобразование и подготовку данных для дальнейшего использования;
- Проектировать хранилища данных, организовывать хранение и управление данными;
- Обеспечивать доступ к данным в форматах, пригодных для бизнес-приложений и потребителей данных.

Необходимые знания и навыки:

- Хорошие знания в области программирования;
- Глубокие знания систем и технологических архитектур;
- Глубокое понимание реляционных баз данных и нереляционных хранилищ данных.

Обязанности Data Analyst

Обязанности Data Analyst следующие:

- Переводить данные и цифры на язык бизнеса;
- Находить закономерности и применять статистические методы для анализа данных;
- Валидировать и очищать данные;

- Визуализировать данные для представления результатов анализа.

Необходимые знания и навыки:

- Хорошее знание электронных таблиц;
- Навыки составления запросов;
- Умение работать со статистическими инструментами для создания дашбордов;
- Некоторые навыки программирования;
- Сильные аналитические и коммуникативные навыки.

Обязанности Data Scientist

Обязанности Data Scientist следующие:

- Анализировать данные для выявления сложных связей и корреляций;
- Создавать модели машинного или глубокого обучения;
- Оценивать эффективность полученных моделей;
- Восстанавливать ход процессов по цифровым следам.

Необходимые знания и навыки:

- Знания математики и статистики;
- Хорошее понимание языков программирования и баз данных;
- Навыки построения моделей данных;
- Знания в предметной области (domain knowledge).

Обязанности BI Analyst

Обязанности BI Analyst следующие:

- Изучать возможные последствия для бизнеса;
- Анализировать риски;
- Анализировать внутренние процессы;
- Предлагать новые направления для развития бизнеса клиента.

Необходимые знания и навыки:

- Знания и навыки работы в BI-инструментах;
- Навыки составления запросов, некоторые знания языков программирования;
- Знания в предметной области (domain knowledge), знания в бизнесе (business knowledge);
- Сильные коммуникативные навыки;
- Навыки представления информации.

Обязанности дата-инженера

Обязанности дата-инженера следующие:

1. Проектировать и внедрять хранилища данных;
2. Проектировать пайплайны для извлечения, преобразования и загрузки данных;
3. Внедрять и поддерживать распределенные системы для крупномасштабной отработки данных;
4. Внедрять решения для оценки качества, защиты конфиденциальности и безопасности данных;

5. Обеспечивать надежность и масштабируемость систем;
6. Обеспечивать соблюдение нормативных требований, мониторинг, резервное копирование и восстановление данных;
7. Разрабатывать интерфейсы и дашборды для предоставления данных бизнес-приложениям и пользователям.

Hard Skills инженера данных

Инфраструктура:

- Работа с операционными системами (UNIX/Linux и Windows);
- Работа с виртуальными машинами, сетями и службами приложений, таких как балансировка нагрузки и мониторинг производительности приложений;
- Работа с облачными сервисами (Amazon, Google, IBM, Microsoft).

Базы и хранилища данных:

- Базы и хранилища данных (RDBMS, IBM, MSSQL, MySQL, Oracle Database/ PostgreSQL);
- Базы данных NOSQL (Redis, MongoDB, Cassandra, Neo4J);
- Хранилища данных (Oracle Exadata, IBM Db2 Warehouse on Cloud, IBM Netezza Performance Server, Amazon RedShift).

ETL-инструменты и инструменты обработки больших данных:

- Решения для создания пайплайнов данных (Apache Beam, AirFlow, DataFlow);
- Инструменты ETl (Apache Nifi, IBM, Infosphere Information Server, AWS Glue, Improvado);
- Инструменты для обработки больших данных (Hadoop, Hive, Spark).

Языки запросов и программирования:

- Языки запросов, манипулирования и обработки данных (SQL и SQL-подробные языки запросов);
- Языки программирования (Python, R, Java);
- Языки оболочки и сценариев (Unix/Linux Shell и PowerShell)

Soft Skills инженера данных

Soft Skills инженера данных следующие:

- Способность преобразовывать бизнес-требования в технические;
- Способность работать с соблюдением полного жизненного цикла разработки программного обеспечения;
- Понимание потенциального применения данных в бизнесе;
- Понимание рисков плохого управления данными;
- Навыки межличностного общения, командной работы и сотрудничества.

Итоги

На данном уроке мы подняли следующие темы:

- Формула 5 «V» Big Data-проекта;
- Задачи инженерии данных;

- Типичные роли в BigData команде и их взаимодействие;
- Обязанности инженера данных;
- Hard Skills и Soft Skills инженера данных.

Как вам урок?

Далее >