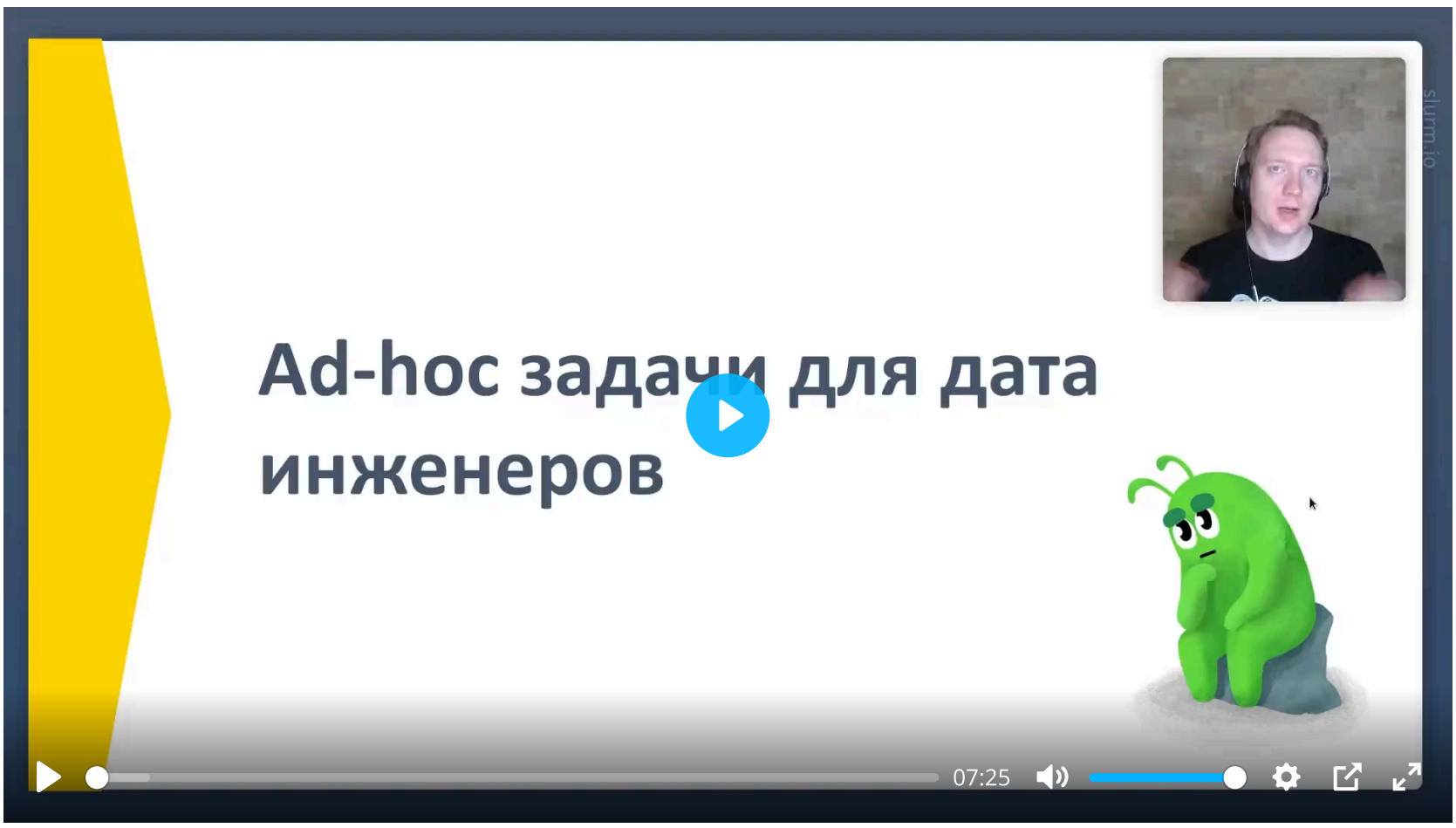

Текстовая расшифровка видео:

AD-HOC ЗАДАЧИ ДЛЯ ДАТА-ИНЖЕНЕРОВ

План:

- Данные: где и как хранятся;
- Bash;
- Вдохновение.

Данные: где и как хранятся

- В базах данных наподобие PostgreSQL;
- Чуть более специализированные форматы – HDF5;
- Различные API;
- Файлы на HDFS или локальных файлах;
- Запакованные каким-то архиватором (tag.gz, lzo, xz, zip);
- С датами или таймстемпами в названиях;
- Иногда в сложной структуре каталогов;
- В формате CSV/TSV (колонки данных) или JSON.

Bash

Как мы знаем, **Linux** – это операционная система, которая была изначально полностью серверной. Это привело к тому, что буквально все операции по работе с файлами и администрированию системы можно сделать через терминал. Так, мы получили высокопроизводительную систему и всеобъемлющий набор инструментов.

Операции:

- Параллельное копирование/перемещение файлов;
- Подсчет простых статистических метрик по сырым данным;
- Извлечение колонок из данных;
- Распаковка архивов на лету;
- Конвертация форматов.

Почти все эти операции может потребоваться в тот или иной момент сделать дата-инженеру, и почти ни для какой из них не нужно сразу бросаться писать код (на Python или другом языке), потому что часто они быстрее решаются подручными средствами. Как говорится, “лучший код - это тот, который не надо писать”.

Основные плюсы командной строки:

- Есть практически на любой *nix/BSD системе.
- Любые операции элементарно автоматизируются написанием скриптов.
- Большинство кода уже написано за нас.
- Огромное количество программ имеют CLI либо версию, работающую в окне терминала (rtorrent, midnight commander, мессенджеры, почтовые клиенты, архиваторы, браузеры).

Вдохновение

Вдохновением для данной темы послужила статья «Консольные утилиты в 235 раз быстрее Hadoop-кластера».

Ознакомиться с ней вы можете по ссылке

Также источником вдохновения послужила книга «Data Science at the Command Line» J. Janssens.

Как вам урок? 😠 😐 🔥

Далее >