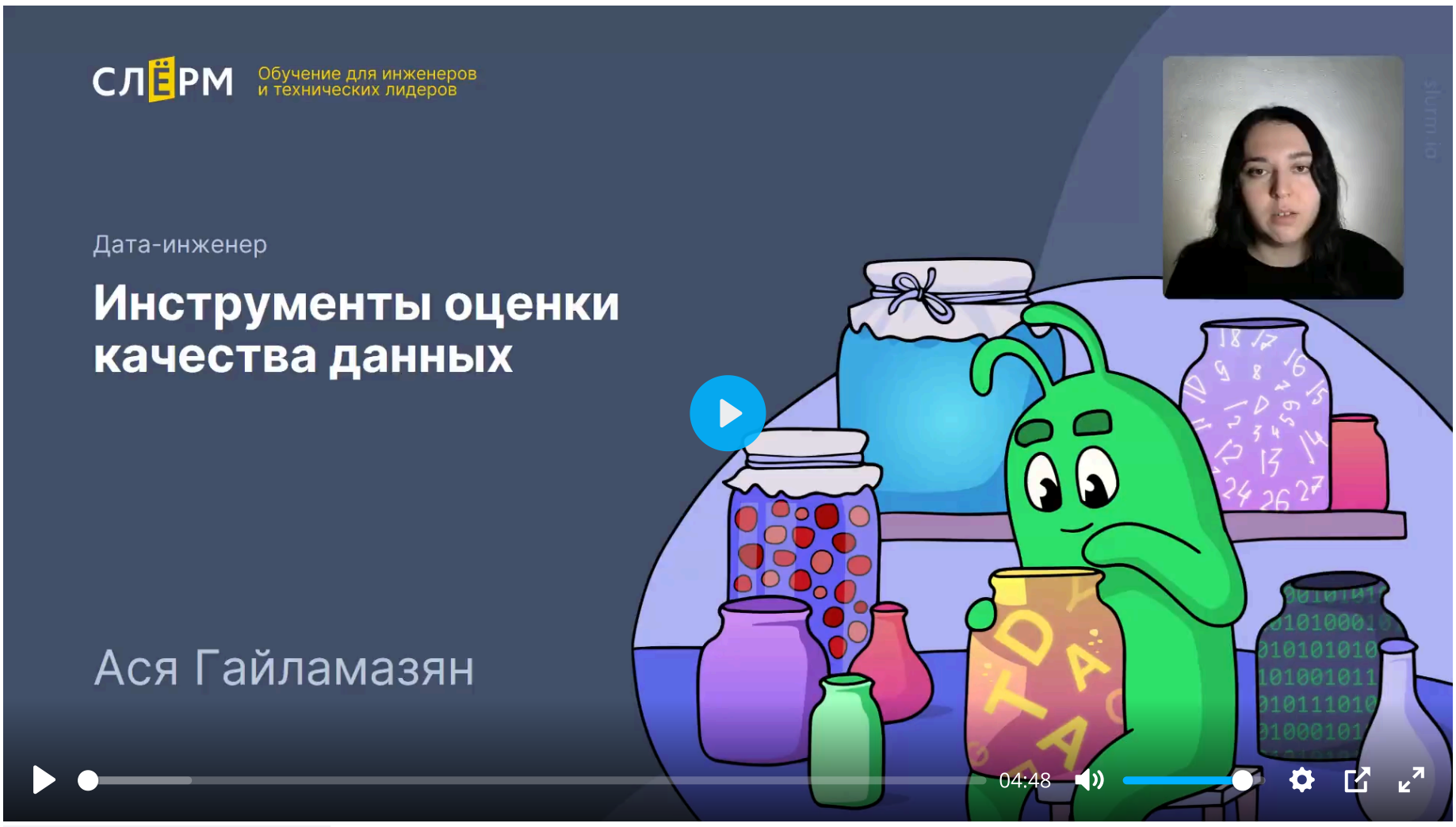


Презентация к уроку 12.2.5

Тексовая расшифровка видео:

ИНСТРУМЕНТЫ ОЦЕНКИ КАЧЕСТВА ДАННЫХ

План:

- Практический кейс очистки данных;
- Amazon Deequ;
- Talend;
- OpenRefine;
- Cucumber;
- Soda;
- Great Expectations;
- Выбор инструмента.

Практический кейс очистки данных

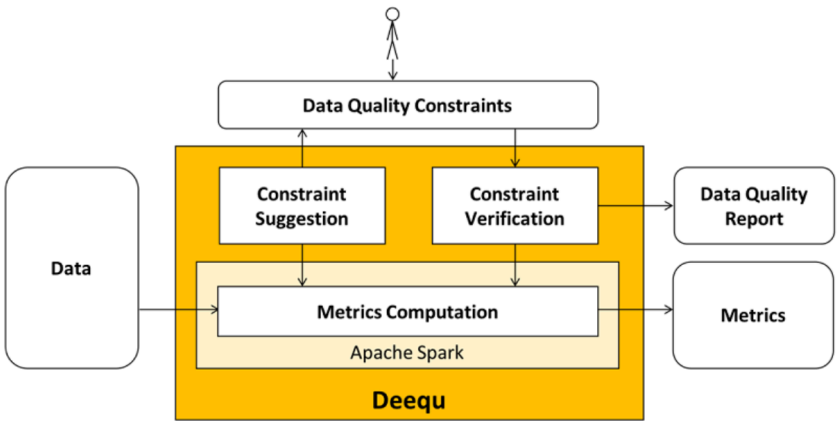
Очистка данных – это процесс, в котором выявляются и исправляются ошибки в данных, неточности, несоответствия в датасете.

Очистку данных можно делать с помощью скриптов препроцессинга (например, на Python): очищать от спама, мёржить их с помощью Pandas.

Существуют специализированные инструменты, например, ранее изученный NiFi. Он может использоваться для препроцессинга, однако имеет ограничения, поскольку лишен обширных функций оценки качества данных, как у специальных инструментов Data Quality.

Amazon Deequ

Amazon Deequ – инструмент, созданный Amazon и являющийся надстройкой над Spark. Этот инструмент создан как Open Source и представляет разные возможности для модульных тестов над данными.



Плюсы:

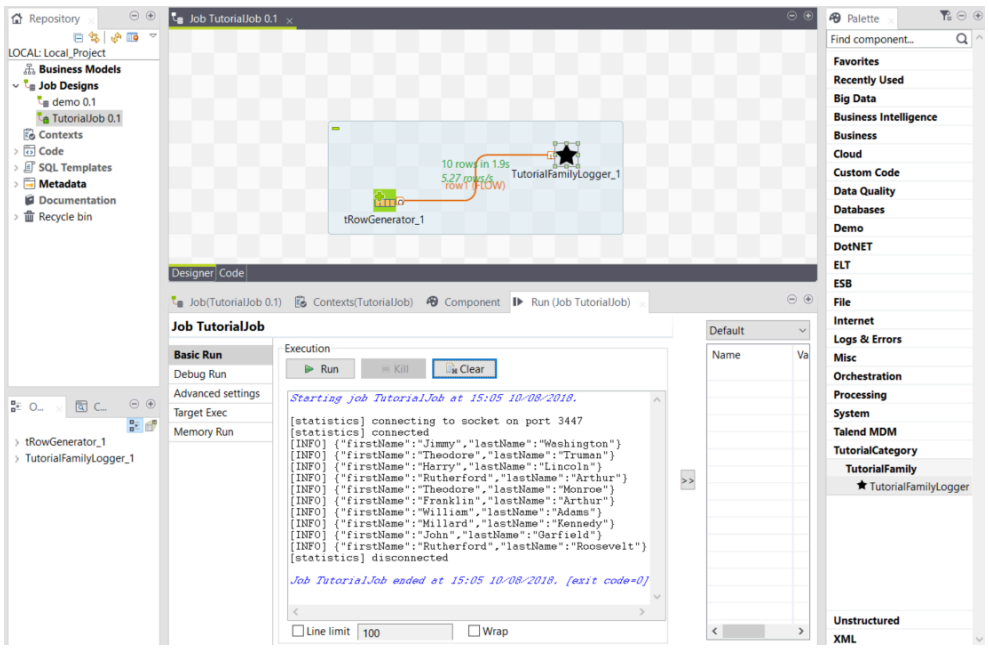
- Deequ специально разработан для оценки и проверки качества данных.
- Хорошо интегрируется с Apache Spark для масштабируемой обработки данных.

Минус:

- Недостаток функций подготовки данных.

Talend

Talend представляет большие возможности для работы с данными. Это тоже инструмент с открытым исходным кодом, а также полноценный ETL-инструмент наряду со Spark.



Плюс:

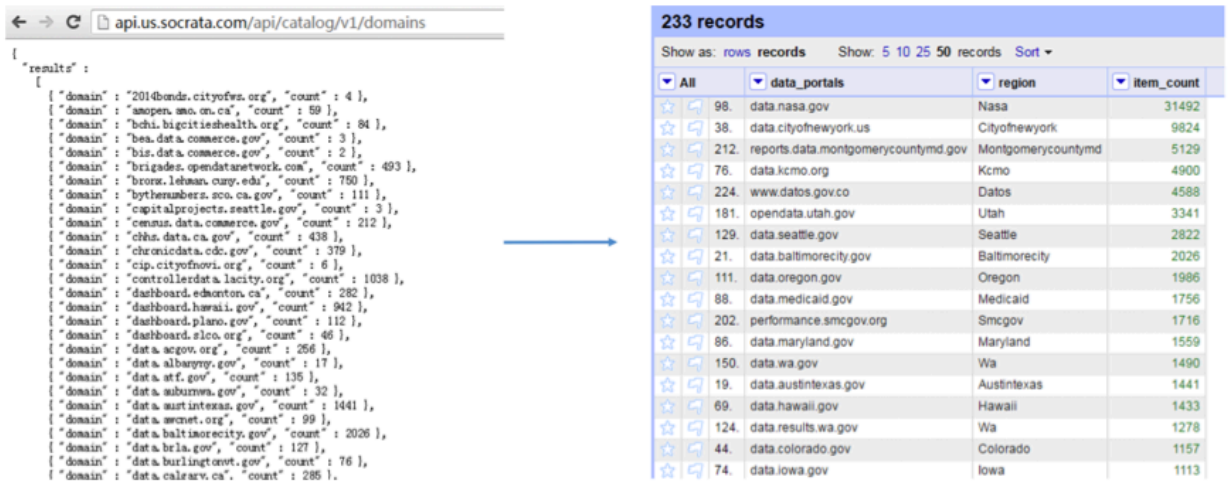
- Возможности профилирования, очистки, стандартизации и проверки данных.

Минус:

- Для использования полного набора функций может потребоваться приобретение платной корпоративной версии.

OpenRefine

OpenRefine подходит для преобразования неупорядоченных данных из одного формата в другой.



Плюсы:

- OpenRefine подходит для задач очистки и преобразования данных.
- Возможности манипулирования над текстом и строками.

Минус:

- Нет расширенных возможностей проверки данных и обнаружения аномалий.

Cucumber

Cucumber не является инструментом для произведения оценки качества данных, но он позволяет создавать тесты над данными. Благодаря BDD (Behavior driven development) можно производить тест над данными и по поведению оценивать их качество.

Плюс:

- Подходит для проведения тестов ПО.

Минус:

- Может не обеспечивать всю глубину оценки качества данных.

Soda

Soda – это инструмент с открытым кодом, который представляет множество способов оценки качества и проверки данных. Особенно хорошо Soda работает с открытыми API.

Плюс:

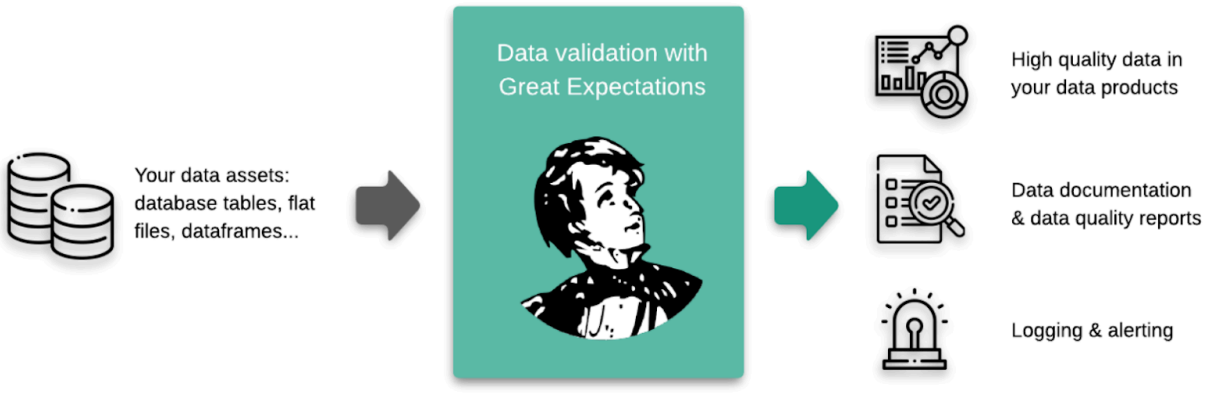
- Soda для оценки и проверки качества данных особенно хорошо работает с открытыми API.

Минус:

- Он может быть более специализированным для конкретных случаев.

Great Expectations

Great Expectations позволяет пользователям задавать ожидаемые значения для данных. Возможности варьируются от того, чтобы в одном столбце данные были не нулевые, до того, чтобы проверять сложные статистические данные.



Плюс:

- Great Expectations – специальный инструмент для обеспечения качества данных.

Минус:

- Может не предлагать полный набор возможностей интеграции данных.

Выбор инструмента

Помимо перечисленных инструментов есть и коммерческие. Нужно выбирать в зависимости от потребности бизнеса.

Если потребность бизнеса в том, чтобы оценивать качество данных, то подойдут:

- Great Expectations;
- Amazon Deequ;
- Soda.

Если нужно строить ETL-пайплайны, то подойдут:

- Talend;
- NiFi.

Если нужно производить тестирования над данными, то подойдут:

- Cucumber;
- Great Expectations.

Если нужна очистка данных, то подойдет:

- OpenRefine.

Как вам урок?

Изучил, далее