

[Презентация к уроку 12.2.5](#)

Текстовая 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

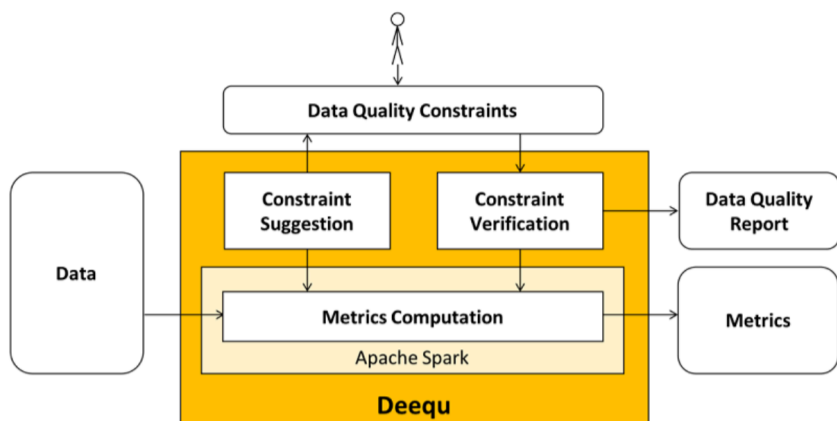
**Очистка данных** – это процесс, в котором выявляются и исправляются ошибки в данных, неточности, несоответствия в датасете.

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:

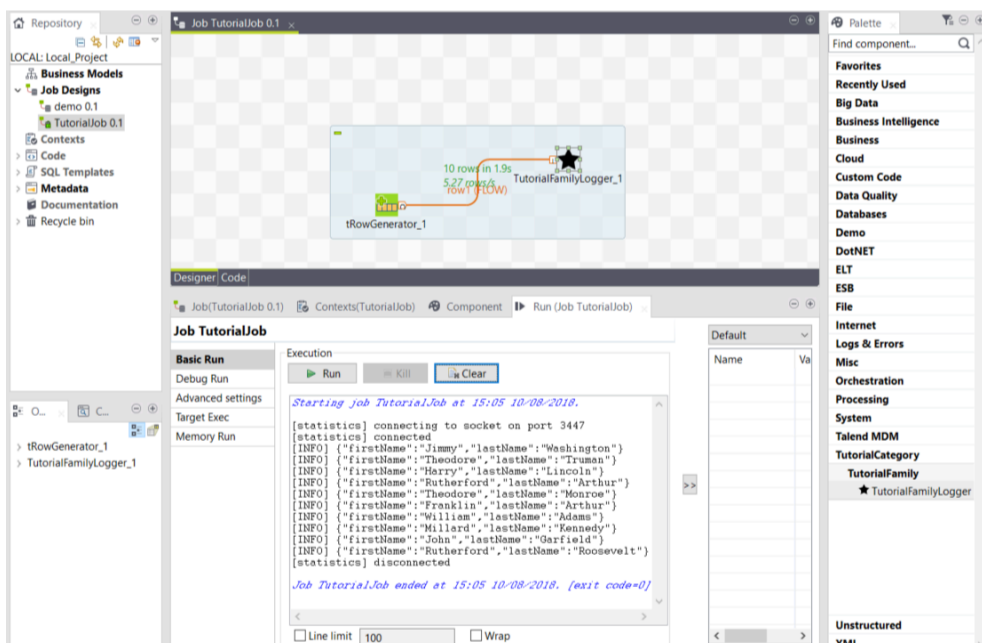
- Deequ специально разработан для оценки и проверки качества данных.
- Хорошо интегрируется с Apache Spark для масштабируемой обработки данных.

### 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.

```

api.us.socrata.com/api/catalog/v1/domains

{
  "results": [
    { "domain": "2010honds.cityofva.org", "count": 4 },
    { "domain": "anopen.mo.on.ca", "count": 59 },
    { "domain": "bchi.bigcitieshealth.org", "count": 84 },
    { "domain": "bes.data.commerce.gov", "count": 3 },
    { "domain": "bis.data.commerce.gov", "count": 2 },
    { "domain": "brigades.opendatametwork.com", "count": 493 },
    { "domain": "brows.lhman.org", "count": 750 },
    { "domain": "bythenumbers.sco.ca.gov", "count": 111 },
    { "domain": "capitalprojects.seattle.gov", "count": 3 },
    { "domain": "census.data.commerce.gov", "count": 212 },
    { "domain": "chhs.data.ca.gov", "count": 438 },
    { "domain": "chronicdata.edu.gov", "count": 379 },
    { "domain": "cip.cityofnovi.org", "count": 6 },
    { "domain": "controllerdata.lacity.org", "count": 1038 },
    { "domain": "dashboard.edmonton.ca", "count": 282 },
    { "domain": "dashboard.hawaii.gov", "count": 942 },
    { "domain": "dashboard.pilano.gov", "count": 112 },
    { "domain": "dashboard.slco.org", "count": 46 },
    { "domain": "data.acgov.org", "count": 256 },
    { "domain": "data.albany.gov", "count": 17 },
    { "domain": "data.stf.gov", "count": 135 },
    { "domain": "data.suburva.gov", "count": 32 },
    { "domain": "data.surtintexas.gov", "count": 1441 },
    { "domain": "data.secret.org", "count": 99 },
    { "domain": "data.baltimorecity.gov", "count": 2026 },
    { "domain": "data.brla.gov", "count": 127 },
    { "domain": "data.burlington.gov", "count": 76 },
    { "domain": "data.caizarv.ca", "count": 285 }
  ]
}

```

233 records			
Show as: rows records Show: 5 10 25 50 records Sort ▾			
All	data_portals	region	item_count
98.	data.nasa.gov	Nasa	31492
38.	data.cityofnewyork.us	Cityofnewyork	9824
212.	reports.data.montgomerycountymd.gov	Montgomerycountymd	5129
76.	data.kcmo.org	Kcmo	4900
224.	www.datos.gov.co	Datos	4588
181.	opendata.utah.gov	Utah	3341
129.	data.seattle.gov	Seattle	2822
21.	data.baltimorecity.gov	Baltimorecity	2026
111.	data.oregon.gov	Oregon	1986
88.	data.medicaid.gov	Medicaid	1756
202.	performance.smcgov.org	Smcgov	1716
86.	data.maryland.gov	Maryland	1559
150.	data.wa.gov	Wa	1490
19.	data.austintexas.gov	Austintexas	1441
69.	data.hawaii.gov	Hawaii	1433
124.	data.results.wa.gov	Wa	1278
44.	data.colorado.gov	Colorado	1157
74.	data.iowa.gov	Iowa	1113

**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:**

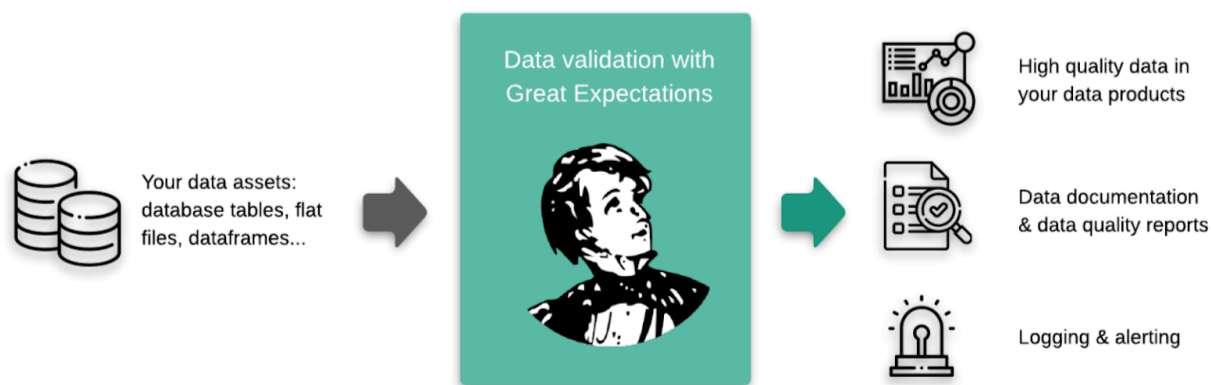
- Soda для оценки и проверки качества данных особенно хорошо работает с открытыми API.

**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 >

