СЛЁРМ Обучение для инженеров и технических лидеров

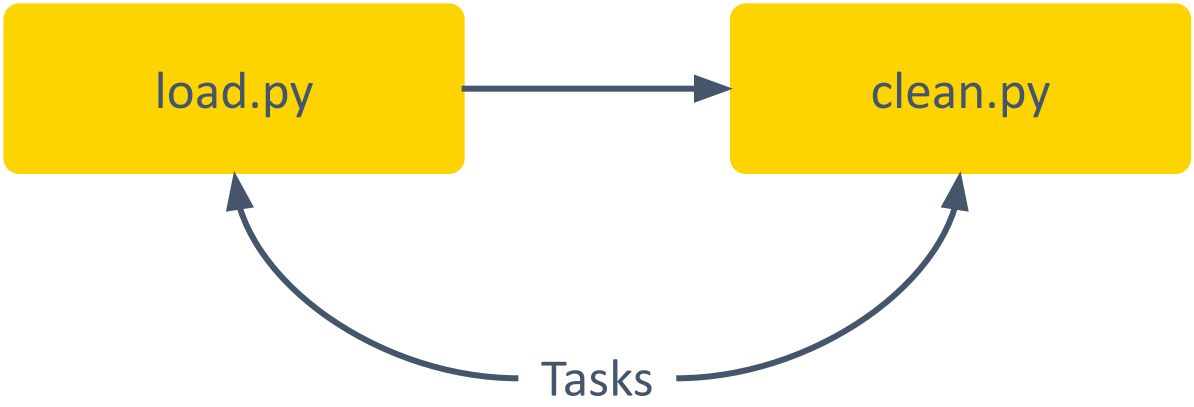
slurm.io

Дата-инженер

Инструменты оценки качества данных

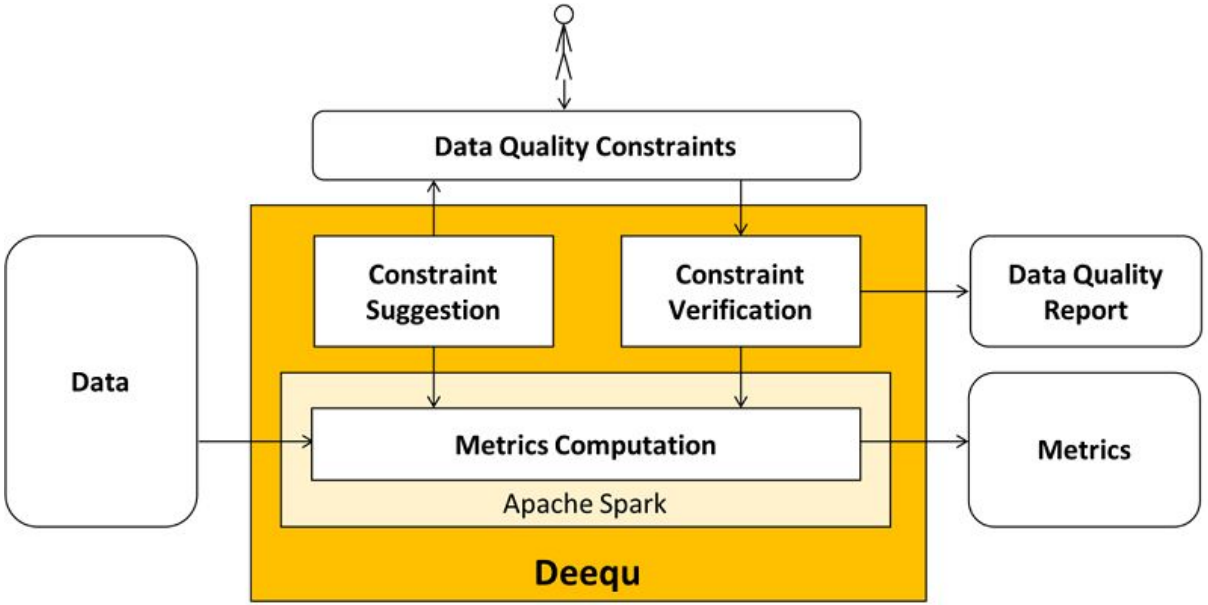
Ася Гайламазян

Практический кейс очистки данных



slurm.io

Amazon Deequ



\+

Deequ специально разработан для оценки и проверки качества данных.

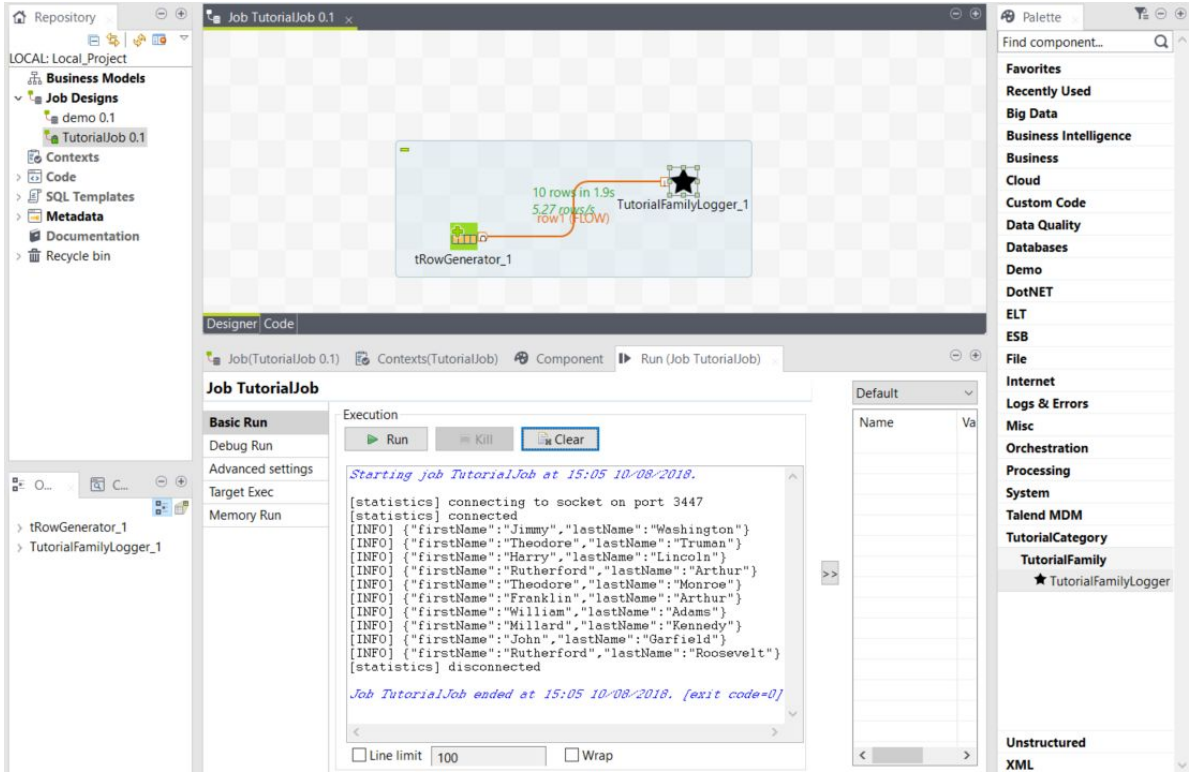
Хорошо интегрируется с apache spark для масштабируемой обработки данных.

\-

Недостаток функций подготовки данных/

slurm.io

Talend



\+

Возможности профилирования, очистки, стандартизации и проверки данных.

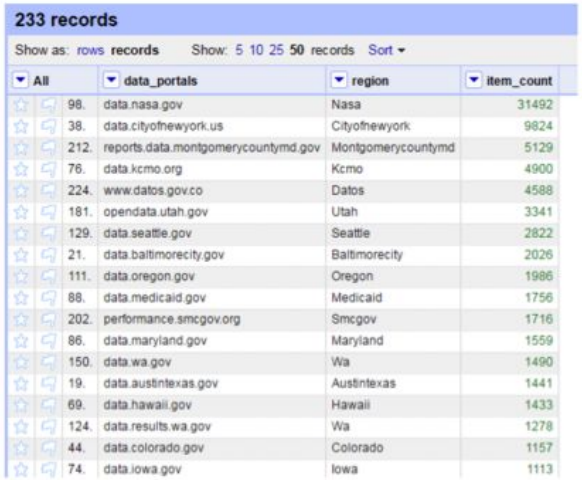
\-

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

slurm.io

OpenRefine

OpenRefine



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

\+

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

\-

Нет расширенных возможностей проверки данных и обнаружения аномалий.

slurm.io

Cucumber

https://cucumber.io/

cucumber

\+

Подходит для проведения тестов ПО.

\-

Может не обеспечивать всю глубину оценки качества данных.

slurm.io

Soda

SODA
Socrata Open Data API

\+

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

\-

Он может быть более специализированным для конкретных случаев.

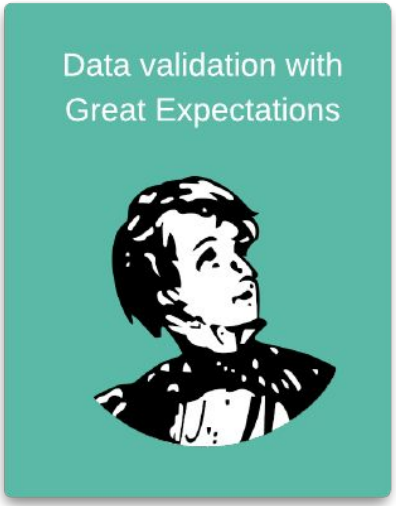
slurm.io

Great Expectations

\+ Great Expectations специальный инструмент для обеспечения качества данных.

\- Может не предлагать полный набор возможностей интеграции данных.

Your data assets: database tables, flat files, dataframes...



High quality data in your data products

Data documentation & data quality reports

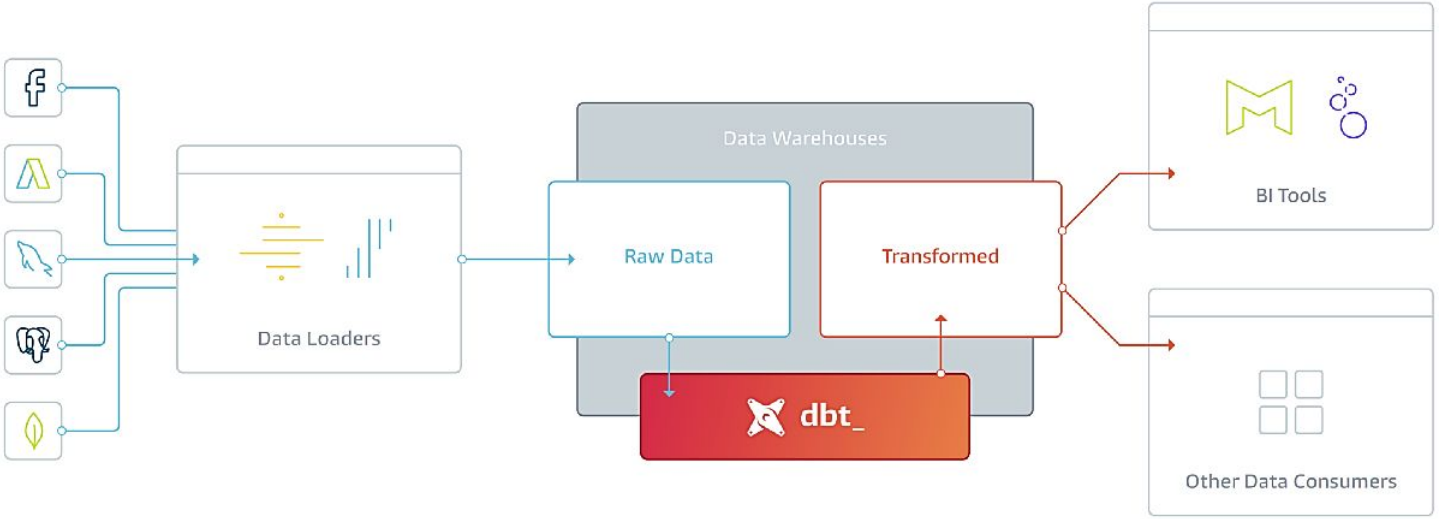
Logging & alerting

slurm.io

dbt

\+ dbt позволяет проверять и тестировать данные с помощью команды dbt test.

\- Может не охватывать весь спектр функций качества данных.



slurm.io

Выбор инструмента

Figure 1. Magic Quadrant for Data Quality Tools

slurm.io