> Конспект > 3 урок > Data Quality

> Оглавление

> Data quality

Data quality (Качество данных) - это практика измерения состояния данных, основанная на таких факторах, как достоверность, полнота, согласованность, надежность и а актуальность.

Измерение уровней качества данных может помочь организациям выявить ошибки данных, которые необходимо устранить, и оценить, подходят ли данные в их ИТ-системах для использования по назначению.

> Факторы качества данных

> Ценность (Value)

Ценность (Value) - критерии оценки необходимости использования данных.

Проверки:

- Использование данных - чем больше объектов или субъектов используют наши данные, тем они ценнее
- Выявление застойных данных - поиск данных, которые в настоящее время потеряли актуальность и не используются вообще (например были заменены другим датасетом).

> Актуальность (Relevance)

Актуальность (Relevance) - совокупность характеристик относящихся к соблюдению сроков, синхронизаций или обновления.

Проверки:

- Время задержки.
- Время последней синхронизации.
- Время последнего обновления данных в хранилище.

> Полнота (Completeness)

Полнота (Completeness) - мера измерения доли пробелов в данных.

Проверки:

- Наличие обязательных полей
- Наличие необязательных полей
- Неполное множество - отсутствие части набора данных по неизвестным причинам.

> Согласованность (Consistency)

Согласованность (Consistency) - мера измерения связанности данных. Пример: данные о пользователях содержат не все данные об их покупках. Датасет покупок частично не связан с датасетом пользователей.

Проверки:

- Отсутствие расхождения в данных
- Корректность связей

> Доступность (Availability)

Доступность (Availability) - процессы и инструменты доступа к данным (юридическая, техническая, операционная).

Проверки:

- Анализ метрик изменения данных
- Анализ метрик чтения данных

> Достоверность (Veracity)

Достоверность (Veracity) - набор свойств для обеспечения однозначности и релевантности данных. Пример: возраст должен быть в адекватных пределах - от 0 до 100.

Проверки:

- Значения однозначны
- Значения действительно возможны и в допустимых пределах.

> Бизнес практики

> Следствия плохого качества

- Неправильные выводы бизнеса
- Репутационные потери
- Непредвиденные расходы (издержки)

- Низкое качество выполняемых задач
- Внутриорганизационные конфликты
- Скрытые недоработки/кастомизация
- Низкая эффективность
- Упущенные возможности

> Улучшение качества

- Решения на стороне бизнеса
- Оценка и подбор источников
- Внедрение практик Data Quality

Каждый бизнес определяет свои нормы к качеству данных. Практики с точки зрения бизнеса:

- Культура работы с данными в компании (Data Governance).
- У каждых данных должен быть владелец (Data owner).
- У данных должен быть потребитель.
- У данных должен быть единый источник (Master Data).
- У данных должны быть метрики качества.

> Практики Data Quality

- Проверка данных из внешних источников
- Проверка данных от пользователей
- Проверка состояний всех внутренних хранилищ (DL, DWH, Datamart ...)
- Разработка и встраивание своих библиотек
- Построение инфраструктуры мониторинга и логирования
- Сепарация данных

- Фиксирование ошибок и реакция на них

> Data quality ≠ Monitoring

	Data quality	Monitoring
Применение	Встраивание	Наблюдения
Реакция	Сепарирование данных	Уведомление/Алертинг
Правила	Конфигурируются/Программируются	Задаются
Отслеживание	Значения/Структура/Статистика данных	Показатели состояния

> Компоненты Data Quality

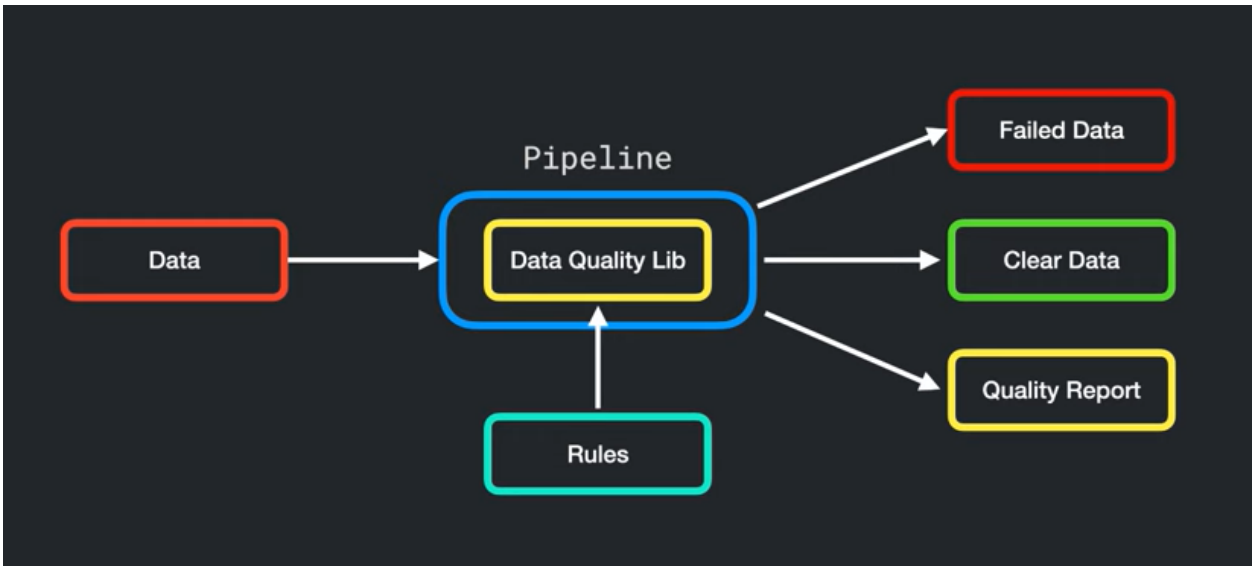
Quality report - Итоговый отчет о проверке данных (итог проверок, статистика данных, объем ошибок/чистых данных, общая оценка качества.

Failed data - Датасет с отобранными данными не удовлетворяющие требованиям качества.

Clear data - Датасет с отобранными данными прошедшими проверки качества.

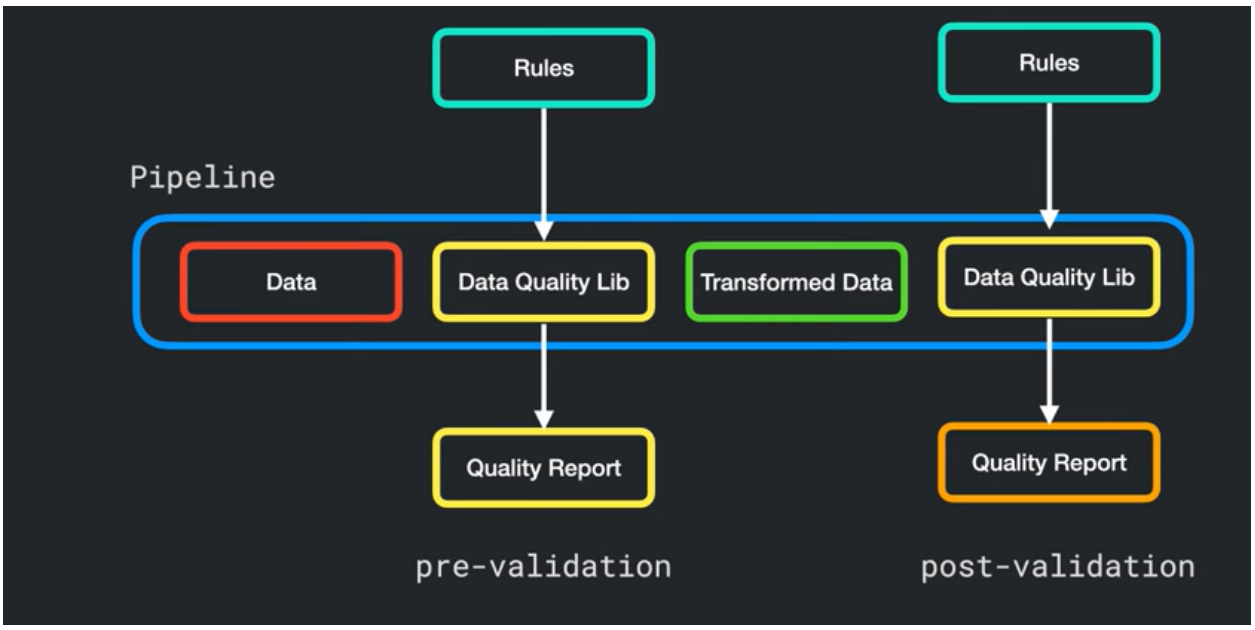
Rules - Установленные правила для проверки данных и оценки итогового качества:

- Описание правил для значения колонок
- Описание правил по набору колонок
- Описание границы разделения качественных данных от некачественных
- Описание вычисляемых статистик по данным
- Описание функции измерения меры качества



При этом, есть 2 способа проверки данных:

- Pre validation
- Post validation



> Использование результатов

Quality Report:

- дальнейший анализ

- Журналирование качества
- Реакция на результат

Failed data:

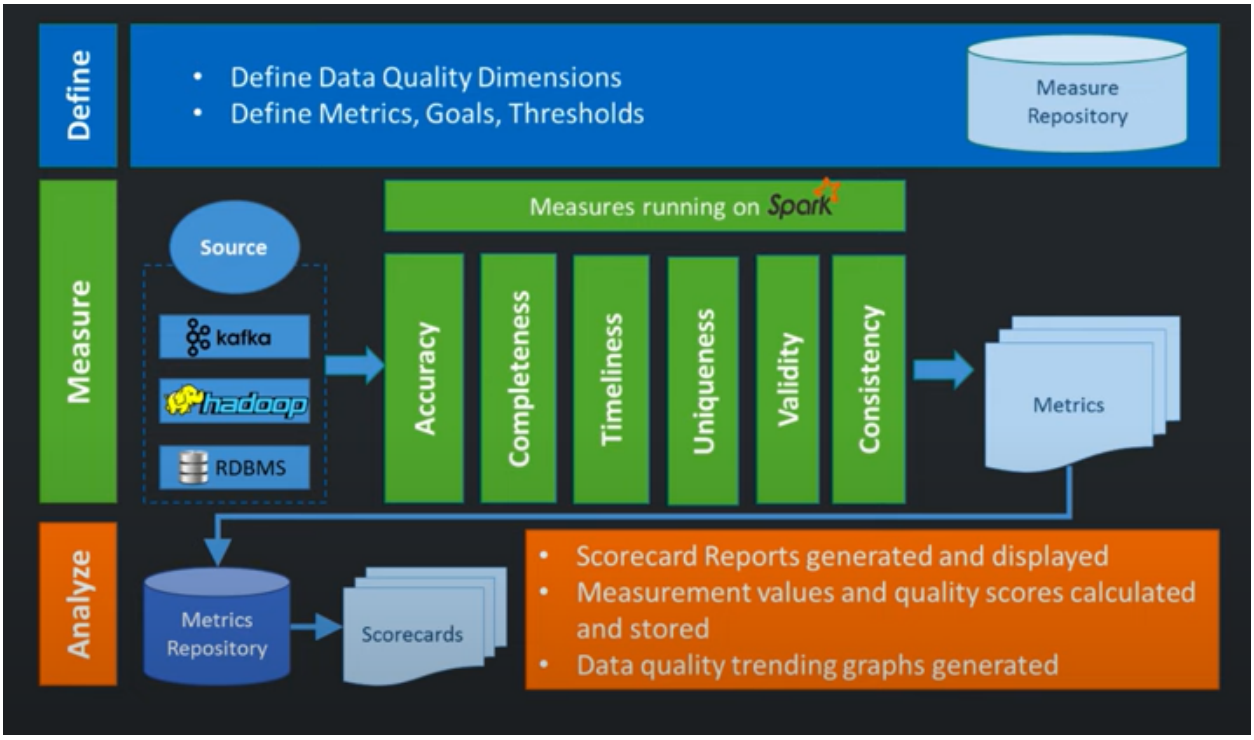
- Сохранение
- Дополнительная проверка
- Возврат источнику

Clear data:

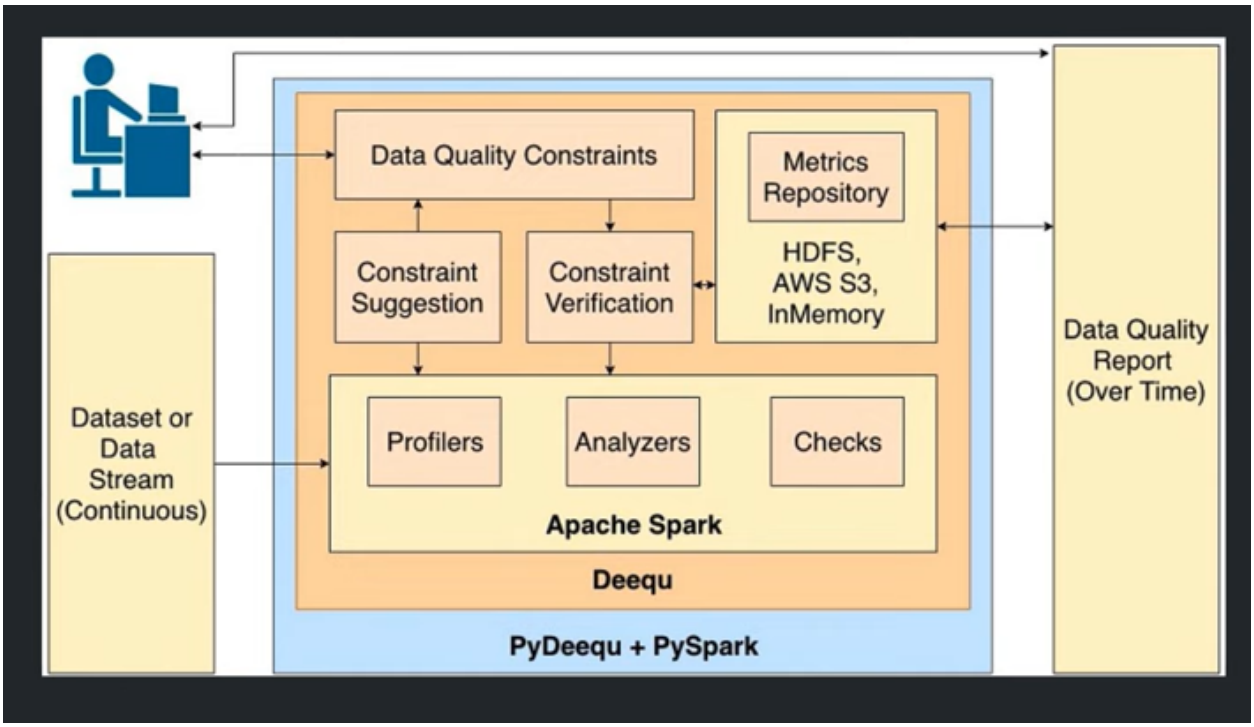
- Дальнейшая обработка согласно пайплайну

> Фреймворки

> Apache Griffin



> PYDEEQU


Data Quality Constraints
Metrics Repository
Constraint Suggestion
Constraint Verification
HDFS, AWS S3, InMemory
Profilers
Analyzers
Checks
Apache Spark
Deequ
PyDeequ + PySpark
Dataset or Data Stream (Continuous)
Data Quality Report (Over Time)