Компьютерные инструменты в образовании, 2026
№ -: 1–9
УДК: 621.320
http://cte.eltech.ru

ВОСПРОИЗВОДИМОЕ ЭКСПЕРИМЕНТАЛЬНОЕ СРАВНЕНИЕ СРЕДСТВ ВАЛИДАЦИИ ДАННЫХ В ETL-ПАЙПЛАЙНАХ

Метелкин М. А.[1], студент, maxwjiill@gmail.com, orcid.org/0009-0003-6547-1102
Пархоменко В. А.[1], ст. преподаватель, parhomenko.v@gmail.com, orcid.org/0000-0001-7757-377X

[1]СПбПУ, Санкт-Петербургский политехнический университет Петра Великого, ул. Политехническая, дом 29, Санкт-Петербург, 195251, Россия

Аннотация

Рост зависимости организаций от аналитики повышает требования к качеству данных, а ETL-пайплайны становятся ключевой точкой накопления и проявления дефектов на этапах извлечения, трансформации и загрузки. При этом на практике широко используются разные средства валидации данных, но выбор инструмента часто не подкреплен сопоставимыми измерениями качества обнаружения нарушений и эксплуатационной стоимости проверок.

В работе предложен воспроизводимый экспериментальный протокол сравнения средств валидации на этапах ETL-пайплайна и выполнено сопоставление Great Expectations, SodaCL и dbt tests на едином наборе логических правил. Эксперименты проводятся в детерминированной среде с контролируемой инъекцией дефектов, а метрики времени и ресурсов фиксируются в техническом контуре наблюдаемости. Показано, что при функционально эквивалентном обнаружении нарушений инструменты существенно различаются по накладным расходам; также экспериментально подтверждена практическая целесообразность этапной стратегии валидации, комбинирующей проверки уровня пайплайна и ограничения целостности СУБД.

Ключевые слова: *etl-пайплайны, качество данных, валидация данных, воспроизводимые эксперименты, great expectations, sodacl, dbt tests.*

Цитирование: Метелкин М. А., Пархоменко В. А. Воспроизводимое сравнение средств валидации данных в ETL-пайплайнах // Компьютерные инструменты в образовании. 2026. № -. С. 1–9.

1. ВВЕДЕНИЕ

Дефекты данных в ETL-пайплайнах приводят к искажению аналитических результатов и требуют системной постановки проверок качества на разных этапах обработки [1, 3]. Международные стандарты качества данных задают терминологию и модели (в частности, измерение валидности в ISO/IEC 25012), однако для практического применения требования необходимо операционализировать в виде конкретных проверок и измеримых метрик [4, 5]. Одновременно развивается экосистема инструментов контроля качества данных, но сопоставимые и воспроизводимые протоколы их сравнения остаются ограниченными, что затрудняет обоснованный выбор инструмента и стратегии валидации [6].

Цель работы — сопоставить распространенные средства валидации данных в контексте ETL-пайплайна по двум группам метрик:

1. качество обнаружения нарушений (факт наличия нарушений и число проваленных проверок);
2. эксплуатационная стоимость (время выполнения и потребление ресурсов).

Дополнительно проверяется гипотеза о том, что этапная стратегия (комбинация проверок уровня пайплайна и ограничений целостности в реляционном слое) обеспечивает более выгодный компромисс «обнаружение–стоимость» по сравнению с универсальными стратегиями.

2. ОБЗОР РАБОТ

2.1. Качество данных и валидация в ETL

ETL-пайплайны включают неоднородные источники, последовательности преобразований и режимы инкрементальной загрузки, что увеличивает число потенциальных точек возникновения дефектов [2, 9]. В стандарте ISO 8000 качество данных рассматривается как характеристика данных в контексте требований и использования, а ISO/IEC 25012 формализует модель качества данных, где валидность выступает одним из ключевых измерений [4, 5]. Для инженерной практики это означает необходимость трансляции требований качества в набор проверок, привязанных к этапам пайплайна и типам нарушений (схема, полнота, согласованность, уникальность и др.).

2.2. Подходы к контролю качества и инструменты

Современные подходы к обеспечению качества в пайплайнах данных включают автоматизацию проверок, мониторинг метрик качества и интеграцию контроля в процесс поставки (CI/CD для данных) [7, 8]. Обзоры средств измерения и мониторинга качества данных отмечают разнообразие инструментов и отсутствие единых критериев сравнения, особенно при учете эксплуатационных характеристик [6].

На практике распространены фреймворки декларативной валидации и профилирования (например, Great Expectations и SodaCL), а также средства SQL-тестирования в контексте трансформаций (например, dbt tests). Эти инструменты отличаются по области применимости (сырые данные vs реляционные слои), формату правил и механизмам формирования отчетности, что делает важным их сравнение на едином наборе логических проверок и в контролируемой среде [10, 11].

3. АРХИТЕКТУРА ИССЛЕДОВАТЕЛЬСКОГО СТЕНДА

Стенд включает: внешний источник данных; оркестратор Apache Airflow, управляющий жизненным циклом прогонов и обеспечивающий выполнение пайплайна в виде набора DAG; Python-приложение, реализующее извлечение данных и загрузку в STG, построение витрины DDS, а также поддерживающее контуры валидации и контролируемого внесения дефектов; хранилище PostgreSQL со слоями STG (raw JSON), DDS и техническим слоем TECH для аудита и учета статусов батчей; конфигурационные артефакты (.env и YAML), задающие параметры подключения, состав проверок, их критичность и сценарии контролируемого внесения дефектов (рис. 1).

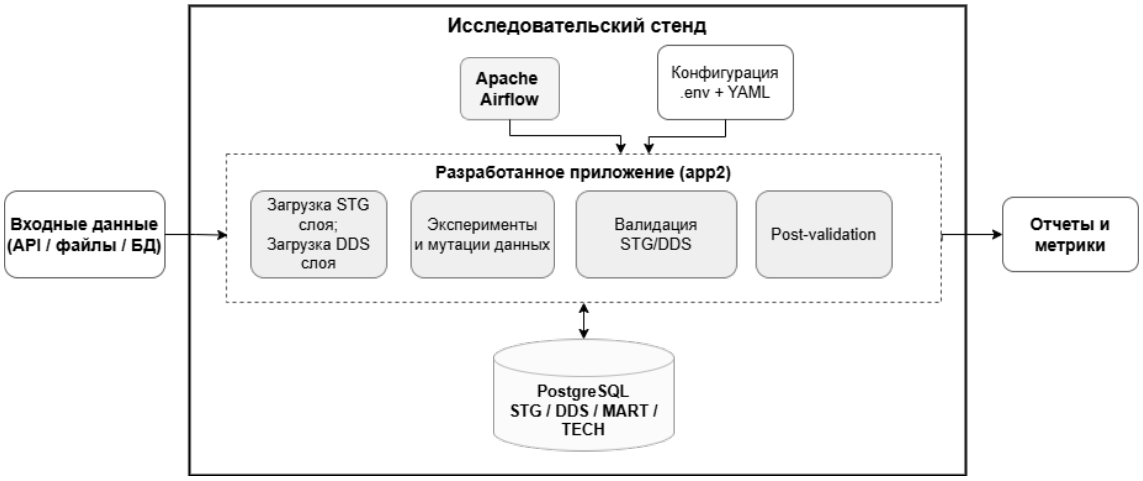

Рис. 1. Архитектура стенда

4. ТЕСТИРОВАНИЕ СТЕНДА

Для подтверждения корректности экспериментальной среды и исключения влияния дефектов стенда на результаты сравнения используется многоуровневое тестирование: модульные, интеграционные и сквозные проверки обеспечивают устойчивое прохождение ключевого потока STG → DDS → MART и корректную фиксацию метрик прогонов в техническом контуре наблюдаемости; дополнительно применяются практики статического анализа и контроля качества, что повышает доверие к воспроизводимости измерений и позволяет интерпретировать различия метрик как свойства сравниваемых средств валидации, а не артефакты инфраструктуры.

5. ЭКСПЕРИМЕНТЫ

5.1. Дизайн эксперимента и метрики

Сравнение выполнено для трех инструментов валидации: Great Expectations (GX), SodaCL и dbt tests. Для обеспечения функциональной сопоставимости используются идентичные логические правила и единый набор данных. Эксперимент включает базовые прогоны и прогоны с контролируемыми изменениями, имитирующими типовые дефекты данных на этапах ETL; результаты агрегируются по повторным запускам. Для каждого прогона фиксируются метрики качества обнаружения и метрики стоимости выполнения, регистрируемые в техническом контуре наблюдаемости.

Тестирование и экспериментальные прогоны выполняются в среде разработчика: Windows 11 Pro (64-bit, версия 10.0.22631, сборка 22631). Аппаратная конфигурация включала центральный процессор AMD Ryzen 5 7600X (6 физических ядер, 12 логических потоков, до 4.7 ГГц) и 32 ГБ оперативной памяти DDR5 (5600 МГц). Для обеспечения воспроизводимости используется контейнеризация Docker Compose и изоляция тестовой инфраструктуры.

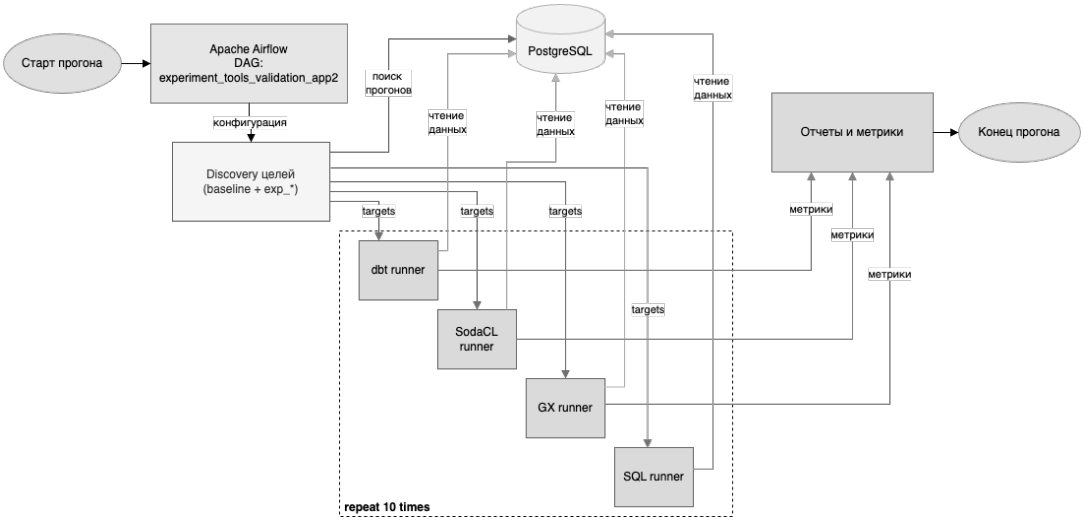

Рис. 2. Архитектура экспериментального контура сравнения средств валидации данных

5.2. Сравнение инструментов по этапам ETL

Для сопоставимых наборов правил инструменты демонстрируют функционально эквивалентное обнаружение нарушений, однако существенно различаются по накладным расходам. На этапе E (Extract/STG, сырые JSON-данные) SodaCL обеспечивает существенно меньшую длительность по сравнению с GX (в среднем 0.15 с против 2.23 с в экспериментальных прогонах). На этапе T (Transform/DDS) минимальные накладные расходы показывает SodaCL (около 0.07 с), GX занимает промежуточное положение (около 1.12 с), а dbt tests демонстрирует наибольшую длительность (около 4.62 с), что связано с особенностями выполнения тестов и накладными расходами запуска. На этапе L (Load/MART) различия усиливаются: GX является наиболее затратным (в среднем 40.71 с), тогда как SodaCL и dbt tests остаются существенно быстрее (2.76 с и 6.31 с соответственно); измерения времени и потребления ресурсов представлены на рис. 3–4.

Таблица 1. Показатели, используемые для построения графиков сравнения производительности (прогоны с контролируемыми изменениями данных; статистика по 10 повторным запускам)

Этап	Инструмент	Длительность, с (среднее ± ст. отклонение)	CPU, % (среднее ± ст. отклонение)	RAM, МБ (среднее ± ст. отклонение)
E (STG)	GX	2.23 ± 0.13	15.76 ± 1.29	365.11 ± 2.69
E (STG)	SodaCL	0.15 ± 0.01	10.90 ± 3.44	356.07 ± 0.45
T (DDS)	GX	1.12 ± 1.85	43.60 ± 9.03	373.85 ± 4.65
T (DDS)	dbt tests	4.62 ± 0.37	0.14 ± 0.12	354.60 ± 1.55
T (DDS)	SodaCL	0.07 ± 0.12	32.18 ± 12.62	377.95 ± 0.00
L (MART)	GX	40.71 ± 26.24	0.71 ± 0.12	363.04 ± 2.68
L (MART)	dbt tests	6.31 ± 1.76	0.09 ± 0.09	356.68 ± 1.57
L (MART)	SodaCL	2.76 ± 1.78	0.52 ± 0.24	358.55 ± 0.98

5.3. Сравнение стратегий: универсальный и этапный подход

Проверена гипотеза о преимуществах этапной стратегии, комбинирующей средства уровня пайплайна и ограничения целостности в реляционном слое: сравнивались универсальные стратегии (GX для всех этапов; SodaCL для всех этапов) и этапные стратегии (SodaCL + SQL constraints + SodaCL; SodaCL + SQL constraints + dbt tests). На прогонах с контролируемыми изменениями этапные стратегии демонстрируют более высокие метрики обнаружения нарушений (доля запусков с нарушениями около 23% против 13% у универсальных стратегий), при этом стратегия SodaCL+SQL+SodaCL сохраняет сопоставимую стоимость по времени относительно универсальной SodaCL (3.01 с против 3.06 с), а стратегия с dbt tests увеличивает длительность (6.81 с). Универсальная стратегия GX является наиболее затратной по времени (43.35 с) и не показывает преимущества по метрикам обнаружения относительно универсальной SodaCL (табл. 2).

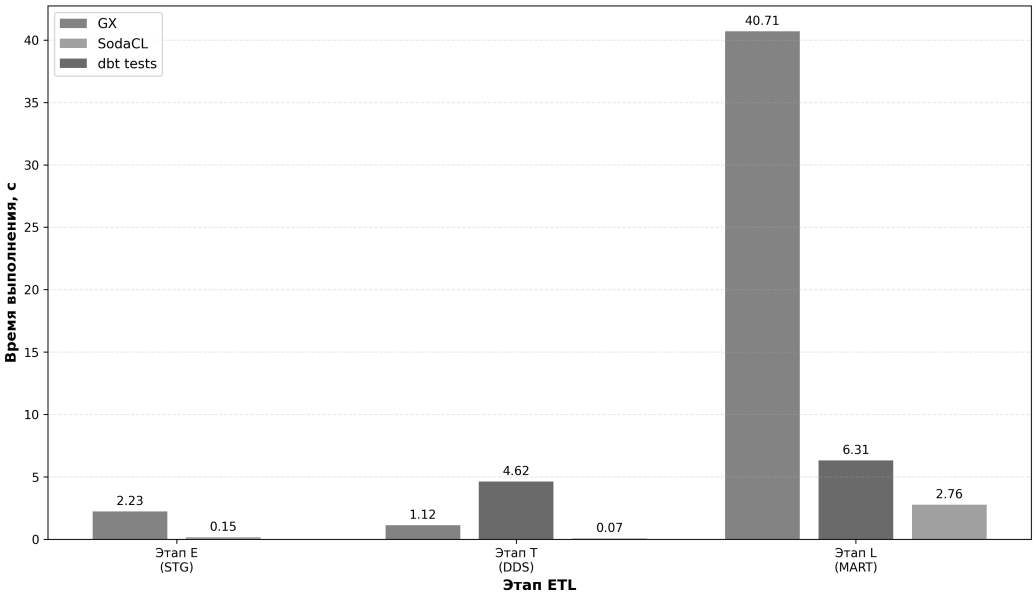

Рис. 3. Сравнение длительности выполнения валидации по этапам ETL (агрегированные результаты экспериментальных прогонов)

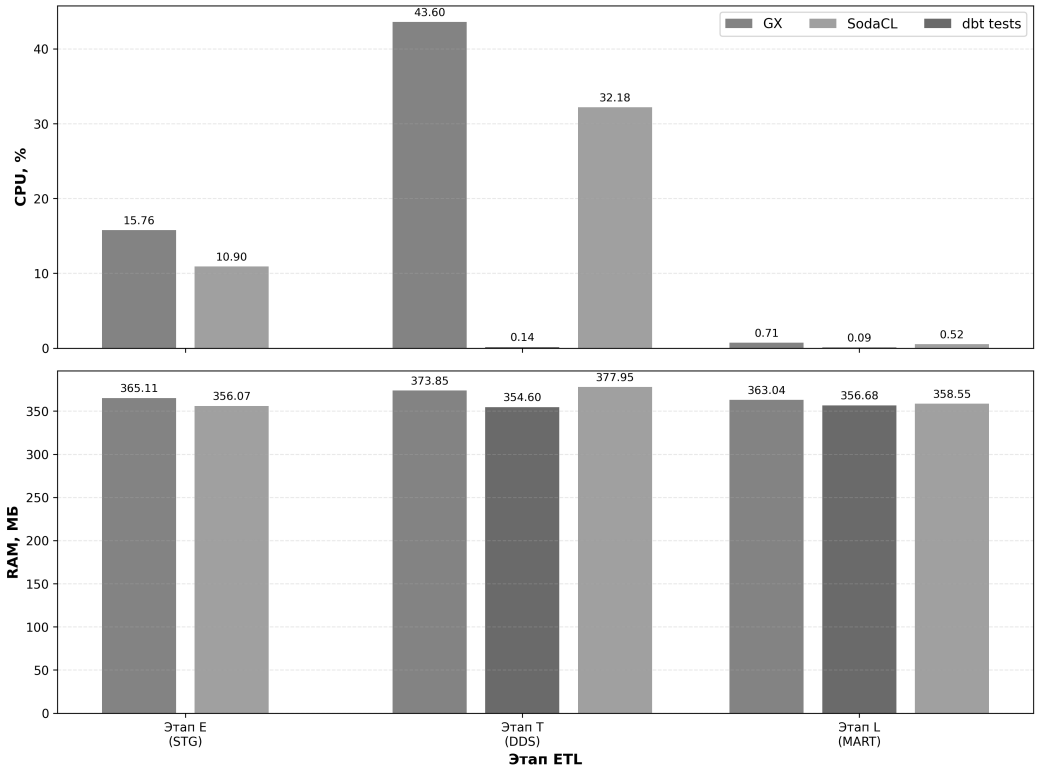

Рис. 4. Сравнение потребления CPU и RAM по инструментам валидации (экспериментальные прогоны)

Таблица 2. Сравнение стратегий валидации на прогонах с контролируемыми изменениями данных (агрегированные метрики)

Стратегия	n	Доля запусков с нарушениями, %	Проваленных проверок, шт. (среднее)	Длительность, с ($mean \pm std$)
Универсальная (GX)	186	12.90	0.226	43.349 ± 24.549
Универсальная (SodaCL)	208	13.94	0.236	3.064 ± 1.894
Этапная (SodaCL + SQL + SodaCL)	208	23.56	0.418	3.012 ± 1.785
Этапная (SodaCL + SQL + dbt)	207	23.19	0.425	6.814 ± 1.777

5.4. Ограничения эксперимента

Эксперимент имеет следующие ограничения:

1. Эксперимент выполнен на репрезентативном, но ограниченном наборе данных предметной области. Выводы о производительности инструментов применимы к данному масштабу данных; поведение при существенном увеличении объема (миллионы строк) требует дополнительного исследования.

2. Эксперимент использует фиксированный набор правил для GX/dbt/SodaCL. Выводы о сопоставимости результатов применимы к указанным наборам правил; расширение набора проверок может выявить дополнительные различия между инструментами.
3. dbt tests не применялся на этапе E (STG) в силу ограничений инструмента по работе с сырыми JSON-данными.
4. Все прогоны выполнены в единой инфраструктуре. Поведение инструментов в распределенной инфраструктуре или при параллельном выполнении проверок не исследовалось.

6. ЗАКЛЮЧЕНИЕ

Предложен воспроизводимый протокол экспериментального сравнения средств валидации данных в ETL-пайплайнах с унифицированным учетом метрик обнаружения и стоимости выполнения в техническом контуре наблюдаемости. Экспериментально показано, что при эквивалентном обнаружении нарушений на сопоставимых наборах правил инструменты существенно различаются по накладным расходам, причем различия зависят от этапа ETL и формата данных. Дополнительно подтверждено, что этапная стратегия, комбинирующая проверки уровня пайплайна и ограничения целостности СУБД, повышает чувствительность к нарушениям при сопоставимой стоимости по времени, что делает ее практичным выбором для реляционных слоев ETL при ограниченном бюджете на проверки.

Список литературы

1. *Kimball R., Caserta J.* The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley, 2004.
2. *Walha A., Ghozzi F., Gargouri F.* Data integration from traditional to big data: main features and comparisons of ETL approaches. The Journal of Supercomputing, 2024, vol. 80, no. 19, pp. 26687–26725.
3. *Foidl H., Golendukhina V., Ramler R., Felderer M.* Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers. Journal of Systems and Software, 2024, vol. 207, 111855.
4. *International Organization for Standardization.* ISO 8000-1:2022 — Data quality — Part 1: Overview. ISO, 2022.
5. *ISO/IEC.* ISO/IEC 25012:2008 — Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. ISO, 2008.
6. *Ehrlinger L., Wöß W.* A Survey of Data Quality Measurement and Monitoring Tools. Frontiers in Big Data, 2022, vol. 5, 850611.
7. *Schelter S. et al.* Automating Large-Scale Data Quality Verification. PVLDB, 2018, vol. 11, no. 12, pp. 1781–1794.
8. *Yang H. et al.* Unlocking the Power of CI/CD for Data Pipelines in Distributed Data Warehouses. PVLDB, 2025, vol. 18, no. 12, pp. 4887–4895.
9. *Ong T. et al.* A Framework for Classification of Electronic Health Data Extraction-Transformation-Loading Challenges in Data Network Participation. eGEMs, 2017, vol. 5, no. 1, p. 10.
10. *Great Expectations.* Great Expectations Documentation, 2025.
11. *Soda.* Soda Documentation, 2025.

Computer tools in education, 2026
№ -: 1–9
http://cte.eltech.ru

Reproducible experimental comparison of data validation tools in ETL pipelines

Metelkin M. A.[1], student, Peter the Great St. Petersburg Polytechnic University,
maxwjiill@gmail.com, orcid.org/0009-0003-6547-1102
Parkhomenko V. A.[1], senior lecturer, Peter the Great St. Petersburg Polytechnic University,
parhomenko.v@gmail.com, orcid.org/0000-0001-7757-377X

[1]SPbPU, Peter the Great St. Petersburg Polytechnic University, Politekhnicheskaya str., house 29, Saint-Petersburg, 195251, Russia

Abstract

Data quality issues in ETL pipelines frequently lead to incorrect analytical results, while the ecosystem of data validation tools keeps expanding. However, tool selection is often not supported by reproducible, comparable measurements of both violation detection quality and operational cost.

This paper proposes a reproducible experimental protocol for comparing ETL-stage data validation tools and reports results for Great Expectations, SodaCL and dbt tests using an identical set of logical rules. Experiments are executed in a deterministic environment with controlled defect injection; runtime and resource metrics are captured in a unified observability layer. The study shows that, while detection results can be functionally equivalent for comparable rule sets, the tools differ significantly in overhead depending on the ETL stage. The paper also confirms the practical benefits of a stage-based validation strategy combining pipeline-level checks with database integrity constraints.

Keywords: *etl pipelines, data quality, data validation, reproducible experiments, great expectations, sodacl, dbt tests.*

Citation: Metelkin M. A., Parkhomenko V. A.. Reproducible experimental comparison of data validation tools in ETL pipelines. Computer tools in education, 2026. № -. P. 1–9. .

References

1. Kimball R., Caserta J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley, 2004.
2. Walha A., Ghozzi F., Gargouri F. Data integration from traditional to big data: main features and comparisons of ETL approaches. The Journal of Supercomputing, 2024, vol. 80, no. 19, pp. 26687–26725.
3. Foidl H., Golendukhina V., Ramler R., Felderer M. Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers. Journal of Systems and Software, 2024, vol. 207, 111855.
4. International Organization for Standardization. ISO 8000-1:2022 — Data quality — Part 1: Overview. ISO, 2022.
5. ISO/IEC. ISO/IEC 25012:2008 — Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. ISO, 2008.
6. Ehrlinger L., Wöß W. A Survey of Data Quality Measurement and Monitoring Tools. Frontiers in Big Data, 2022, vol. 5, 850611.

7. Schelter S. et al. Automating Large-Scale Data Quality Verification. PVLDB, 2018, vol. 11, no. 12, pp. 1781–1794.
8. Yang H. et al. Unlocking the Power of CI/CD for Data Pipelines in Distributed Data Warehouses. PVLDB, 2025, vol. 18, no. 12, pp. 4887–4895.
9. Ong T. et al. A Framework for Classification of Electronic Health Data Extraction-Transformation-Loading Challenges in Data Network Participation. eGEMs, 2017, vol. 5, no. 1, p. 10.
10. Great Expectations. Great Expectations Documentation, 2025.
11. Soda. Soda Documentation, 2025.