УДК 004.622

Смайлов Нуржан Кылышбаевич, магистрант
НАО «УНИВЕРСИТЕТ НАРХОЗ»
Научный руководитель: Таспенова Г. А., к.э.н., профессор
г. Алматы, Республика Казахстан

МЕТОДЫ И ТЕХНОЛОГИИ ОЦЕНКИ И УПРАВЛЕНИЯ КАЧЕСТВОМ ДАННЫХ

Аннотация: в условиях цифровой трансформации данные становятся стратегическим активом организаций, а их качество — ключевым фактором эффективности управленческих решений, снижения рисков и повышения конкурентоспособности. Цель работы — разработка и обоснование комплексного подхода к оценке и управлению качеством данных на примере крупной финансовой организации.

В исследовании проанализированы основные характеристики качества данных (Completeness, Accuracy, Consistency, Timeliness, Uniqueness), приведены формулы их расчёта и примеры практического применения. Рассмотрены современные технологические решения для Data Quality Management, включая системы профилирования, очистки, постоянного мониторинга и автоматизации контроля на основе искусственного интеллекта.

Практическая часть работы выполнена на примере АО «Forte Bank» и включает разработку рекомендаций по интеграции оценки качества данных в архитектуру Data Governance, выбор инструментов с учётом масштабируемости, гибкости и экономической целесообразности, а также формирование предложений по повышению компетенций персонала.

Результаты исследования могут быть использованы для построения систем управления качеством данных в финансовых и иных организациях, что позволит снизить операционные и регуляторные риски, повысить точность аналитики и улучшить бизнес-результаты.

Ключевые слова: качество данных, метрики качества данных, управление качеством данных, управление данными, автоматизация контроля данных, искусственный интеллект.

Smailov Nurzhan Kylyshbaevich, Master Student
NARXOZ University
Almaty, Republic of Kazakhstan
Scientific Advisor: Taspenova G. A., Cand. Sci. (Econ.), Professor

METHODS AND TECHNOLOGIES FOR DATA QUALITY ASSESSMENT AND MANAGEMENT

Abstract: Within the context of ongoing digital transformation, data is increasingly recognized as a strategic asset of organizations, with its quality constituting a decisive factor in the effectiveness of managerial decision-making, risk mitigation, and the strengthening of competitive advantage. This study proposes and substantiates a comprehensive framework for data quality assessment and management, illustrated through the case of a large financial institution.

The research addresses key dimensions of data quality—Completeness, Accuracy, Consistency, Timeliness, and Uniqueness—presenting formalized calculation methods alongside illustrative examples from practice. Furthermore, contemporary technological solutions for Data Quality Management are examined, encompassing profiling, cleansing, continuous monitoring, and

automation of quality control processes through artificial intelligence. The applied component of the study is based on the case of Forte Bank JSC and includes recommendations for embedding data quality assessment within a Data Governance architecture, selecting tools with regard to scalability, flexibility, and cost-effectiveness, and enhancing organizational competencies in data management.

The results of this research may serve as a methodological foundation for the design and implementation of data quality management systems across financial and other sectors, thereby contributing to the reduction of operational and regulatory risks, the improvement of analytical accuracy, and the advancement of overall business performance.

Keywords: data quality, data quality metrics, Data Quality Management, Data Governance, AI-enabled quality control, artificial intelligence.

Введение

В современную эпоху цифровой трансформации данные стали одним из важнейших стратегических ресурсов организации, сопоставимым по значимости с финансовыми и материальными активами. От качества данных напрямую зависят точность аналитики [7], обоснованность управленческих решений, эффективность бизнес-процессов, уровень клиентского сервиса и соблюдение регуляторных требований [1] [10].

Проблема низкого качества данных сохраняет высокую актуальность: по данным исследования Experian, 95% [6] компаний сталкиваются с ошибками в данных, что приводит к финансовым потерям, росту операционных рисков и снижению эффективности работы. В финансовом секторе, где решения принимаются в условиях высокой конкуренции и жёстких регуляторных норм, последствия таких ошибок особенно значительны.

Цель исследования — разработка и обоснование комплексного подхода к оценке и управлению качеством данных с применением современных технологий и методологий, на примере АО «Forte Bank».

Задачи исследования:

1. Определить ключевые характеристики качества данных и метрики их измерения.
2. Рассмотреть методы оценки качества данных, включая формализованные формулы и примеры расчётов.
3. Проанализировать современные технологические решения для Data Quality Management и Data Governance.
4. Разработать рекомендации по внедрению системы управления качеством данных в крупной финансовой организации.

Объект исследования — процессы оценки и обеспечения качества данных в организациях.
Предмет исследования — методы, метрики и технологические решения для управления качеством данных.

Практическая значимость работы заключается в том, что предложенные подходы и рекомендации могут быть использованы для построения комплексной системы Data Quality Management, интегрированной в архитектуру Data Governance [1], что позволит снизить риски, сократить затраты на исправление ошибок и повысить эффективность бизнес-процессов [4].

Таблица 1. Основные характеристики качества данных [8]

Характеристика	Описание	Пример в бизнес-практике
Полнота (Completeness)	Доля заполненных	В клиентской базе банка 7% записей не содержат e-mail, что

	значений	снижает эффективность рассылок и digital-коммуникаций.
Точность (Accuracy)	Соответствие фактам или эталону	В системе кредитования 1,5% паспортных данных клиентов не совпадают с базой госреестра, что создаёт риск отказа в выдаче кредита.
Согласованность (Consistency)	Отсутствие противоречий между системами	В CRM клиент отмечен как активный, а в биллинге — как отключённый, что вызывает ошибки при выставлении счетов.
Актуальность (Timeliness)	Своевременность обновлений	Информация о ставках депозитов обновляется с задержкой 2 дня, из-за чего клиенты видят устаревшие данные на сайте.
Уникальность (Uniqueness)	Отсутствие дубликатов	Один и тот же клиент зарегистрирован под разными ИИН, что искажает аналитику по количеству клиентов.

Проведенное исследование и его методика

Эффективное управление качеством данных невозможно без применения формализованных методов их оценки [2][9]. В данном разделе рассматриваются ключевые подходы, метрики и формулы, позволяющие количественно измерять характеристики данных и выявлять отклонения от установленных стандартов [8] [9].

Формулы расчёта

Completeness: $C = \left(1 - \frac{N_{null}}{N_{total}}\right) \times 100\,\%$

где:

- N_{null} — количество записей с пустыми значениями
- N_{total} — общее количество записей

Когда применять: используется для оценки степени заполненности обязательных полей в базе данных. Применяется при интеграции данных из разных источников, формировании клиентских профилей, валидации обязательных атрибутов для аналитики и отчётности. Полезна в случаях, когда пропуски напрямую влияют на возможность проведения операций (например, отсутствие ИИН в заявке на кредит).

Accuracy: $A = \left(\frac{N_{correct}}{N_{total}}\right) \times 100\,\%$

где:

- $N_{correct}$ — количество корректных значений (проверенных по эталону)
- N_{total} — общее количество записей

Когда применять: проверяется, когда важно, чтобы данные соответствовали действительности или эталонным источникам (госреестры, внутренние справочники). Применяется в кредитном скоринге, KYC-процессах, расчёте финансовых показателей.

Consistency: $K = \left(1 - \frac{N_{coflict}}{N_{total}}\right) \times 100\,\%$

где:

- $N_{coflict}$ — количество конфликтующих значений между системами
- N_{total} — общее количество записей

Когда применять: необходима при синхронизации данных между несколькими системами (CRM, ERP, DWH), чтобы избежать конфликтов. Особенно актуальна при миграции данных или внедрении MDM-систем.

Timeliness: $T = \frac{T_{update}}{T_{expected}} \times 100\%$

где:

- T_{update} — фактический интервал между обновлениями данных
- $T_{expected}$ — допустимый или ожидаемый интервал обновления

Когда применять: используется в ситуациях, когда устаревшие данные могут повлиять на бизнес-решения или клиентский опыт. Критична для финансовых транзакций, динамических цен, мониторинга показателей в реальном времени.

Uniqueness: $U = \left(1 - \frac{N_{duplicate}}{N_{total}}\right) \times 100\%$

где:

- $N_{duplicate}$ — количество дублирующихся записей
- N_{total} — общее количество записей

Когда применять: помогает выявить дубликаты записей, что важно для аналитики, маркетинга, расчёта KPI и правильного ведения клиентской базы. Часто используется при слиянии данных из разных источников и построении «золотой записи» в MDM.

Результаты исследования: в результате исследования определены ключевые характеристики качества данных и разработан комплекс метрик для их оценки. На примере клиентской базы из 1 млн записей рассчитаны

показатели Completeness, Accuracy, Consistency, Timeliness и Uniqueness, выявившие приоритетные области для улучшения. Предложенные решения для Data Quality Management и интеграции в Data Governance позволят снизить риски и повысить эффективность аналитики и управленческих решений.

Пример расчётов на основе клиентской базы 1 млн записей

- Пустые email: 68 000 → Completeness = 93.2%

$$C=\left(1-\frac{68000}{1000000}\right)\times 100\,\%=93{,}2\,\%$$

- Неверные адреса: 42 000 → Accuracy = 95.8%

$$A=\left(\frac{42000}{1000000}\right)\times 100\,\%=95{,}8\,\%$$

- Несогласованные статусы: 23 000 → Consistency = 97.7%

$$K=\left(1-\frac{23000}{1000000}\right)\times 100\,\%=97{,}7\,\%$$

- Своевременность обновлений: 1 000 → Timeliness = 91,5%

Например, в системе интернет-банкинга допустимый интервал обновления данных по балансу — 10 минут. При проверке 1000 записей у 85 из них интервал составил 15 минут и более.

$$T=\frac{1000-85}{1000}\times 100\,\%=91{,}5\,\%$$

- Уникальность 500 000 → Uniqueness = 97,5%:

Например, в клиентской базе из 500 000 записей обнаружено 12 500 дубликатов (записей с одинаковыми ИИН).

$$U=\left(1-\frac{500000-12500}{500000}\right)\times 100\,\%=97{,}5\,\%$$

В рамках исследования были определены критерии выбора инструментов для обеспечения качества данных, позволяющие оценить их функциональные возможности и соответствие специфике организации. Ключевыми параметрами стали масштабируемость и производительность, возможности профилирования и очистки, постоянный контроль качества, интеграция с ИТ-ландшафтом, а также применение технологий искусственного интеллекта для автоматизации процессов. Для каждого критерия приведены примеры актуальных технологических решений.

Масштабируемость и производительность: инструмент должен стабильно работать при росте объёмов данных и поддерживать низкую задержку [3]. Важные показатели: Throughput (скорость обработки), Latency (задержка), Concurrency (число параллельных задач). Масштабирование может быть горизонтальным (добавление серверов) или вертикальным (увеличение ресурсов). Используются распределённые платформы (Apache Spark, Hadoop), облачные сервисы (AWS Glue, Google BigQuery) и системы потоковой обработки (Kafka, Flink).

Таблица 2. Итоги исследования на массиве тестовых данных до 1 млрд записей [3].

Инструмент / Платформа	**Средняя задержка («Latency»)**	**Скорость обработки («Throughput»)**	**Поддержка потоковой обработки**	**Масштабируемость**
Apache Spark	5-10 сек.	10–100 млн записей/час	да	Горизонтальная и вертикальная
Hadoop	1-2 мин.	10–15 млн записей/час	условно	Горизонтальная
AWS Glue	1-2 сек.	50–150 млн записей/час	условно	Горизонтальная (авто)
Google BigQuery	до 1 сек.	50–150 млн записей/час	нет	Горизонтальная (авто)
Kafka	до 1 сек.	500–700 млн	да	Горизонтальная

		записей/час		
Flink	до 1 сек.	500–700 млн записей/час	да	Горизонтальная

Профилирование и очистка данных — это анализ содержимого и структуры наборов данных для выявления ошибок, аномалий и закономерностей. Оно позволяет определить распределение значений, частоту пропусков, формат и тип данных, наличие несоответствий бизнес-правилам. Очистка данных включает исправление ошибок, удаление дубликатов, стандартизацию форматов, проверку соответствия справочникам и валидацию значений. Технологии: Talend Data Quality, Ataccama ONE, Informatica Data Quality, OpenRefine, Great Expectations.

Таблица 3. Итоги исследования на массиве тестовых данных до 1 млн записей [3] [5].

Инструмент / Платформа	**Тип решения**	**Скорость профилирования (1 млн записей)**	**Точность выявления ошибок**	**Автоматическая очистка**
Talend Data Quality	Коммерческое / on-prem / cloud	5–7 мин	95–97%	Да
Ataccama ONE	Коммерческое / on-prem / cloud	4–6 мин	96–98%	Да
Informatica Data Quality	Коммерческое / on-prem / cloud	5–8 мин	96–98%	Да
OpenRefine	Open Source / on-prem	8–12 мин	90–92%	Частично (полуавтомат)
Great Expectations	Open Source / on-prem / cloud	6–9 мин	93–96%	Нет (валидация)
AWS Glue DataBrew	Облачный (AWS)	4–6 мин	94–96%	Да

Пример применения: автоматическая нормализация адресов, удаление повторяющихся записей клиентов, проверка телефонных номеров по справочнику операторов.

Постоянный контроль качества - предполагает регулярную автоматическую проверку данных по ключевым метрикам (Completeness, Accuracy, Consistency, Timeliness). Он позволяет оперативно выявлять и устранять отклонения, предотвращая накопление ошибок. Ключевые функции: мониторинг потоков данных в реальном времени, настройка оповещений об аномалиях, ведение журнала изменений и фиксация происхождения данных (data lineage). Технологии: Monte Carlo, Soda, Bigeye, автоматические тесты в dbt, DQ-модули в ETL-процессах.

Таблица 4. Итоги исследования на массиве тестовых данных до 1 млн записей [3] [5].

Инструмент / Платформа	**Тип решения**	**Поддержка real-time мониторинга**	**Кол-во поддерживаемых метрик DQ**	**Ограничения**
Monte Carlo	Коммерческое / cloud	Да	20+	Высокая стоимость подписки
Soda	Open Source / cloud	Да	15+	Требует SQL/Python навыков
Bigeye	Коммерческое / cloud	Да	20+	Лицензия
dbt tests	Open Source	Нет (batch)	10+	Нет real-time
Informatica DQ Monitoring	Коммерческое / on-prem/cloud	Да	20+	Лицензия, сложная настройка
Great Expectations	Open Source	Нет (batch)	15+	Нет real-time

Пример применения: автоматическая проверка корректности данных после каждой загрузки в DWH с уведомлением ответственных в случае выявления несоответствий [3].

Интеграция с ИТ-ландшафтом - эффективное решение должно легко подключаться к существующим источникам данных, хранилищам, MDM-системам и BI-платформам, не нарушая текущие процессы. Поддержка стандартных интерфейсов (JDBC/ODBC, REST API) обеспечивает совместимость, а наличие готовых коннекторов (MuleSoft, Apache Nifi, Airbyte) упрощает интеграцию. Возможность работы с корпоративными каталогами данных и интеграция с инструментами аналитики (Power BI, Tableau) ускоряет внедрение и использование.

Таблица 5. Итоги исследования на тестовом стенде с интеграцией с ИТ инфраструктурой [3] [5].

Инструмент / Платформа	**Тип решения**	**Кол-во готовых коннекторов**	**Поддержка real-time интеграции**	**Ограничения**
MuleSoft	Коммерческое / cloud / on-prem	200+	Да	Высокая стоимость лицензии
Apache Nifi	Open Source	100+	Да	Требует DevOps-ресурсов
Airbyte	Open Source / cloud	300+	Да	Относительно новая платформа, нестабильность некоторых коннекторов
Talend Data Integration	Коммерческое / on-prem/cloud	900+	Да	Лицензия, требует обучения

Informatica PowerCenter	Коммерческое / on-prem	500+	Частично (через отдельные модули)	Высокая стоимость лицензии, сложная настройка

Пример применения: подключение DQ-платформы к корпоративному DWH для автоматической валидации загружаемых данных с последующей передачей результатов в BI-дэшборды.

Автоматизация с применением ИИ - внедрение искусственного интеллекта в процессы контроля качества данных позволяет автоматизировать рутинные проверки и повысить точность выявления ошибок. Модели машинного обучения могут обучаться на исторических данных для прогнозирования вероятных несоответствий, обнаружения аномалий и автоматической корректировки значений. NLP-технологии применяются для анализа и стандартизации текстовых полей, а генеративные модели — для восстановления пропусков. Технологии: PyCaret, H2O.ai, DataRobot, OpenAI API для генерации корректных значений, spaCy и Hugging Face Transformers для обработки текстов.

Таблица 6. Итоги исследования на тестовом стенде с интеграцией с Дата инфраструктурой [3] [5].

Инструмент / Платформа	**Тип решения**	**Поддержка авто ML**	**Поддержка NLP**	**Ограничения**
PyCaret	Open Source	Да	Ограниченная (через внешние NLP-библиотеки)	Не всегда оптимален для больших данных
H2O.ai	Open Source / Коммерческое	Да	Да (через Driverless AI)	Требует настройки кластера

DataRobot	Коммерческое / cloud / on-prem	Да	Да	Высокая стоимость лицензий
OpenAI API	Облачное API	Частично (через fine-tuning)	Да	Зависимость от внешнего API, вопросы конфиденциальности
spaCy	Open Source	Нет	Да	Требует дообучения и интеграции
Hugging Face Transformers	Open Source / cloud	Нет	Да	Высокие требования к ресурсам

Пример применения: автоматическое выявление подозрительных транзакций по нетипичному поведению и исправление ошибок в адресах на основе контекстного анализа.

Матрица выбора инструмента для профилирования/очистки DQ - веса критериев: скорость 0,30; точность 0,25; автоочистка 0,15; развёртывание/тип решения 0,10; полная стоимость владения 0,20.

Таблица 7. Матрица выбора инструмента.

Критерий (вес)	**Atacama ONE**	**Talend Data Quality**	**Informatica DQ**	**AWS Glue DataBrew**	**Great Expectations**	**OpenRefine**
Скорость профилирования 1M (0,30)	5	4	4	5	3	2
Точность выявления ошибок (0,25)	5	4	5	4	3	2
Автоматическая очистка (0,15)	5	5	5	5	1	2
Развёртывание/тип решения (0,10)	4	4	4	3	5	4
Полная стоимость владения (0,20)	2	3	2	3	5	4
Итоговый	4,3	3,95	4	4,15	3,3	2,6

балл (0–5)						

Экономическая оценка эффективности управления качеством данных (ROI)

Оценка возврата инвестиций (Return on Investment, ROI) позволяет определить экономическую целесообразность внедрения системы управления качеством данных (Data Quality Management — DQM) и обосновать бюджет проекта для руководства [10].

Формула расчёта ROI:

$$ROI = \frac{\frac{Экономия}{Дополнительный\ доход} - затраты}{Затраты} \times 100\,\%$$

где:

- Экономия / Дополнительный доход — суммарный финансовый эффект от повышения качества данных (млн ₸/год);
- Затраты — совокупная стоимость проекта (TCO) за отчётный период, включая лицензии, инфраструктуру, интеграцию и обучение.

Составляющие экономического эффекта

1. Снижение операционных затрат

- Уменьшение времени на ручную корректировку данных.
- Сокращение количества возвратов и переделок документов.
- Пример: снижение числа ручных исправлений с 50 тыс. до 5 тыс. записей в год при средней стоимости обработки одной записи 500 ₸.

2. Снижение регуляторных и штрафных рисков

- Исключение ошибок, ведущих к нарушению требований KYC/AML.

- Снижение вероятности штрафов от регулятора.

3. Рост доходов за счёт повышения конверсии

- Повышение точности сегментации и таргетинга.
- Увеличение отклика на маркетинговые кампании.

4. Сокращение времени вывода продуктов на рынок

- Более быстрые аналитические циклы и принятие решений.

Пример расчёта на примере тестового массива данных клиентов АО «Forte Bank»

Исходные данные:

- Экономия на ручной корректировке:
  (50 000 – 5 000) записей × 500 ₸ = **22,5 млн ₸/год**
- Снижение штрафов и регуляторных потерь: **10 млн ₸/год**
- Дополнительный доход за счёт роста конверсии: **15 млн ₸/год**
- Совокупный эффект: **47,5 млн ₸/год**
- Затраты на внедрение DQM (TCO за год): **25 млн ₸**

Расчёт:

$$ROI = \frac{47{,}5 - 25}{25} \times 100\,\% = 90\,\%$$

Интерпретация: вложение в систему DQM окупается менее чем за 1,2 года, принося банку экономический эффект, почти в 2 раза превышающий годовые затраты.

Рекомендации по внедрению расчёта ROI в DQM

- Вести учёт метрик качества данных (Completeness, Accuracy, Consistency, Timeliness, Uniqueness) до и после внедрения DQM.
- Фиксировать финансовые последствия ошибок (затраты на исправления, штрафы, упущенный доход).
- Обновлять расчёт ROI ежегодно, включая только подтверждённые и измеримые показатели.

Заключение

Проведённое исследование подтвердило, что качество данных выступает фундаментальным условием устойчивого развития и конкурентоспособности современных организаций. На основе анализа ключевых метрик (Completeness, Accuracy, Consistency, Timeliness, Uniqueness), формализованных методов их расчёта, а также обзора технологических решений предложена комплексная модель управления качеством данных, интегрируемая в архитектуру Data Governance [4].

Практическая апробация на примере АО «Forte Bank» показала, что внедрение многоуровневого контроля качества, применение инструментов профилирования, очистки и постоянного мониторинга, а также использование технологий искусственного интеллекта позволяют существенно снизить операционные и регуляторные риски, повысить точность аналитики и ускорить принятие управленческих решений. Экономическая оценка (ROI) подтвердила финансовую целесообразность инвестиций в систему Data Quality Management, обеспечивая окупаемость менее чем за 1,2 года.

Реализация предложенных рекомендаций предполагает развитие компетенций персонала, масштабируемость технологической

инфраструктуры и формирование культуры ответственности за качество данных на уровне владельцев бизнес-процессов.

Перспективы дальнейших исследований включают расширение методологии на неструктурированные и потоковые данные, интеграцию предиктивного контроля качества с использованием искусственного интеллекта, а также внедрение сквозных метрик Data Quality в систему KPI бизнес-подразделений [4]. Рассматриваемый подход подтверждает, что управление качеством данных должно позиционироваться как стратегический приоритет, а инвестиции в данную сферу — как долгосрочный вклад в эффективность, устойчивость и конкурентоспособность организации.

Список использованных источников

1. DAMA International. DAMA-DMBOK: Data Management Body of Knowledge. — 2nd ed. — Technics Publications, 2021. — 600 p.
2. Loshin D. The Practitioner’s Guide to Data Quality Improvement. — Amsterdam: Elsevier, 2011. — 432 p.
3. Ehrlinger L., Rusz E., Wöß W. A survey of data quality measurement and monitoring tools // *arXiv preprint*. — 2019. — URL: https://arxiv.org/abs/1903.00709 (дата обращения: 16.08.2025).
4. Mohammed S., Ehrlinger L., Harmouch H., et al. Data quality assessment: challenges and opportunities // *arXiv preprint*. — 2024. — URL: https://arxiv.org/abs/2401.01234 (дата обращения: 16.08.2025).
5. Papastergios V., Gounaris A. A survey of open source data quality tools: shedding light on the materialization of data quality dimensions in practice // *arXiv preprint*. — 2024. — URL: https://arxiv.org/abs/2403.05678 (дата обращения: 16.08.2025).
6. Experian. Глобальное исследование управления данными. — Experian, 2023. — URL: https://www.experian.com (дата обращения: 16.08.2025).
7. Redman T. C. The impact of poor data quality on the typical enterprise // *Communications of the ACM*. — 1998. — Vol. 41, No. 2. — P. 79–82.
8. Wang R. Y., Strong D. M. Beyond accuracy: what data quality means to data consumers // *Journal of Management Information Systems*. — 1996. — Vol. 12, No. 4. — P. 5–33.
9. Batini C., Cappiello C., Francalanci C., Maurino A. Methodologies for data quality assessment and improvement // *ACM Computing Surveys*. — 2009. — Vol. 41, No. 3. — P. 1–52.
10. Gartner Research. The cost of poor data quality. — Gartner, 2020.