УДК 004.6:004.05

А. Р. Магденко
А. В. Грицак

ПРОБЛЕМИ ЯКОСТІ ДАНИХ У BIG DATA ТА МЕТОДИ ЇХ ПОДОЛАННЯ

Вінницький національний технічний університет

Анотація

Це дослідження присвячене критичному аналізу зростаючих проблем якості даних (Data Quality, DQ) в архітектурах Big Data та розробці комплексних методів їх подолання. Встановлено, що традиційні підходи до DQ не справляються з експоненціальним зростанням обсягів, швидкості та різноманітності даних (3V), що призводить до динамічних та багатовимірних дефектів, які прямо впливають на точність аналітичних моделей та надійність прийняття рішень. Основний акцент зроблено на необхідності переходу від реактивного очищення до проактивного забезпечення якості «у джерелі». Запропоновано інтеграцію методів машинного навчання (ML) та штучного інтелекту (AI) для автоматизованого виявлення аномалій та інтелектуальної корекції помилок. Підкреслюється, що технологічні рішення повинні бути доповнені організаційним підходом, а саме – впровадженням політик Data Governance та інституту Data Stewardship. Результати роботи формують основу для розробки гібридних, адаптивних систем управління якістю даних, здатних підтримувати високу надійність Big Data конвеєрів.

Ключові слова: Big Data, Якість даних (Data Quality, DQ), Управління даними (Data Governance), Машинне навчання (ML), Автоматизоване очищення даних, Аналіз аномалій, Повнота даних, Послідовність даних, Data Lake.

Abstracts

This study is devoted to a critical analysis of the growing problems of data quality (DQ) in Big Data architectures and the development of comprehensive methods to overcome them. It has been established that traditional approaches to DQ cannot cope with the exponential growth in data volume, velocity, and variety (3V), which leads to dynamic and multidimensional defects that directly affect the accuracy of analytical models and the reliability of decision-making. The main emphasis is on the need to move from reactive cleansing to proactive quality assurance "at the source." The integration of machine learning (ML) and artificial intelligence (AI) methods for automated anomaly detection and intelligent error correction is proposed. It is emphasized that technological solutions must be complemented by an organizational approach, namely the implementation of Data Governance policies and the institution of Data Stewardship. The results of the work form the basis for the development of hybrid, adaptive data quality management systems capable of maintaining the high reliability of Big Data pipelines.

Keywords: Big Data, Data Quality (DQ), Data Governance, Machine Learning (ML), Automated Data Cleaning, Anomaly Detection, Data Completeness, Data Consistency, Data Lake.

Вступ

У сучасну цифрову еру дані є ключовим стратегічним активом, а технології Big Data (що характеризуються експоненціальним зростанням Обсягу, Швидкості та Різноманітності) стали основою для прийняття рішень та інновацій. Проте, саме ці характеристики Big Data створюють критичні виклики для якості даних (Data Quality, DQ). Традиційні, статичні методи контролю DQ виявляються неефективними в динамічному середовищі, де проблеми якості охоплюють не лише точність і повноту, а й такі аспекти як темпоральність (актуальність) та послідовність між несумісними джерелами. Низька якість даних прямо призводить до спотворення аналітичних результатів, зниження ефективності систем машинного навчання та хибних стратегічних рішень, спричиняючи значні фінансові та репутаційні втрати [1]. Метою даного дослідження є систематизація ключових проблем DQ в екосистемі Big Data та розробка комплексних, гібридних методів їх подолання. Основний акцент робиться на необхідності переходу від реактивного очищення до проактивного забезпечення якості «у джерелі» та інтеграції методів машинного навчання (ML) та штучного інтелекту (AI) для автоматизованого виявлення аномалій і корекції помилок. Водночас, підкреслюється, що технологічні рішення мають бути нерозривно пов'язані з організаційними механізмами Data Governance та Data Stewardship, які забезпечують прозорість та відповідальність за якість даних протягом усього їхнього життєвого циклу.

Результати дослідження

Зростаюча складність та багатовимірність проблем якості даних. Проблематика якості даних (DQ) в екосистемі Big Data радикально відрізняється від тих викликів, з якими стикалися фахівці у світі традиційних реляційних баз даних. Якщо раніше контроль якості був сфокусований переважно на статичних атрибутах, таких як точність, повнота, унікальність і валідність структурованих даних, то тепер він охоплює динамічні, контекстуальні та міжсистемні аспекти. Основні атрибути Big Data – Обсяг (Volume), Швидкість (Velocity) та Різноманітність (Variety), які спільно відомі як "3V", – експоненціально ускладнюють традиційні детерміновані підходи. Некерований Обсяг даних, що вимірюється петабайтами та екзабайтами, робить фізично неможливим застосування ручних методів очищення або навіть вибіркового аудиту, вимагаючи розробки абсолютно нових, масштабованих та розподілених фреймворків DQ, які можуть оперувати з петабайтними сховищами, такими як Data Lakes. Водночас, надзвичайна Швидкість генерації даних, характерна для потоків IoT, фінансових ринків та соціальних медіа, створює критичну проблему актуальності (Timeliness): дані, навіть якщо вони технічно коректні, стають абсолютно марними для систем прийняття рішень у реальному часі, якщо вони не валідуються і не обробляються в межах мілісекундного часового вікна. Це вимагає вбудовування DQ-логіки безпосередньо в архітектури потокової обробки (наприклад, з використанням Apache Flink або Kafka Streams). Найбільшої складності додає Різноманітність, що включає не лише структуровані формати, а й великі масиви напівструктурованих (JSON, XML) та неструктурованих даних (текст, зображення, лог-файли). Ця різнорідність призводить до невідповідностей і проблем послідовності (Consistency), оскільки одне й те саме бізнес-поняття може інтерпретуватися по-різному у різних джерелах. Для подолання цього потрібні складні методи семантичного інтегрування та зіставлення схем, які виходять далеко за межі простих перетворень даних. Таким чином, у Big Data проблеми DQ стають не статичними, а багатовимірними, динамічними та контекстуально залежними [1].

Вплив низької якості даних на аналітичні результати та прийняття рішень. Неякісні дані в екосистемі Big Data становлять значний бізнес-ризик, оскільки вони прямо та неминуче призводять до спотворення аналітичних інсайтів і, як наслідок, до хибних стратегічних рішень. Системи машинного навчання (ML) та штучного інтелекту (AI), які є основою для використання Big Data, є особливо чутливими до якості вхідних даних. Якщо навчальні набори містять велику кількість пропусків, викидів або систематичних невідповідностей, моделі неминуче набувають зміщення (Model Bias) і починають узагальнювати помилки, що різко знижує їхню передбачувальну точність та надійність. Наприклад, модель, призначена для прогнозування поведінки клієнтів, але навчена на даних із суттєвими прогалинами в демографічних або історичних покупках, буде систематично давати невірні рекомендації, що призведе до втрати клієнтів і зниження продажів. Крім того, низька якість даних безпосередньо підриває основну мету інвестицій у Big Data – отримання надійних стратегічних інсайтів. Коли управлінські рішення, чи то про вихід на новий ринок, чи про оптимізацію ланцюжка поставок, ґрунтуються на забруднених даних, компанія несе фінансові втрати, репутаційні збитки і стикається з неефективним розподілом капіталу. Ще один критичний наслідок – це ризики порушення регуляторних норм (Compliance Risks). У багатьох галузях (фінанси, охорона здоров'я) існують суворі вимоги щодо точності та походження даних (Data Provenance), які потрібні для аудиту (наприклад, GDPR або Basel III). Нездатність довести точність і послідовність даних через їхню низьку якість може призвести до значних штрафів та правових наслідків. Таким чином, забезпечення високої DQ є не просто технічним завданням, а критичною передумовою для отримання позитивної рентабельності інвестицій і підтримки конкурентоспроможності [2].

Застосування машинного навчання та штучного інтелекту для автоматизованого управління якістю даних. Розв'язання проблеми DQ у масштабах Big Data вимагає переходу до інтелектуального очищення даних (Intelligent Data Cleansing) та автоматизації. Методи машинного навчання (ML) та штучного інтелекту (AI) є ключовими для цього переходу, оскільки вони здатні ідентифікувати складні патерни помилок, які неможливо виявити за допомогою традиційних, жорстко закодованих правил. Наприклад, неконтрольоване навчання, зокрема алгоритми кластеризації (як-от DBSCAN або Isolation Forest), є надзвичайно ефективним для виявлення аномалій (Outlier Detection). Ці моделі навчаються на "нормальній" поведінці даних і автоматично позначають ті записи, які значно відхиляються від встановлених патернів. Це дозволяє швидко ідентифікувати шахрайські транзакції, несправні сенсори або незвичайні операційні збої. З іншого боку, навчання з учителем використовується для класифікації якості: модель навчається на історично позначених як правильні або неправильні записах, що дозволяє їй автоматично класифікувати нові вхідні дані як "прийнятні", "помилкові" або "вимагають ручного

перегляду". Для роботи з масивами неструктурованих даних застосовується Обробка природної мови (NLP), що дозволяє витягувати, стандартизувати та нормалізувати сутності (наприклад, перетворювати безліч варіантів написання назви організації до єдиного еталонного формату). Навіть для такої складної задачі, як імп'ютація пропущених даних (Imputation), використовуються методи глибокого навчання, які моделюють складні нелінійні взаємозв'язки між атрибутами для більш точного заповнення прогалин, ніж прості статистичні методи. Таким чином, ML/AI перетворює DQ з рутинного процесу на динамічну, самонавчальну систему, яка є єдиним життєздатним способом підтримання високої якості даних у масштабі та зі швидкістю Big Data [3].

Необхідність впровадження культури "Data Governance" та "Data Stewardship". Технологічні рішення для управління якістю даних, якими б досконалими вони не були, залишаться неефективними без належної організаційної підтримки та чітко визначеної відповідальності. У середовищі Big Data, де дані розподілені між численними підрозділами та системами, вкрай необхідне впровадження формальної структури Data Governance (Управління даними) та інституту Data Stewardship (Відповідальність за дані). Data Governance – це стратегічний рівень, який встановлює "що" і "чому" в управлінні даними. Він відповідає за визначення загальних політик, стандартів, процесів та метрик якості даних (наприклад, встановлення порогів допустимої повноти або точності для критичних даних), а також за розподіл повноважень, забезпечення дотримання регуляторних норм і управління ризиками, пов'язаними з даними. DG гарантує, що всі рішення щодо даних узгоджуються зі стратегічними цілями підприємства [4]. На противагу цьому, Data Stewardship є операційним рівнем, який відповідає за виконання цих політик на щоденній основі. Data Stewards є ключовими фігурами, які слугують мостом між бізнес-користувачами та технічними командами. Їхні обов'язки включають вирішення конфліктів якості даних, ведення бізнес-глосарію, керування метаданими та забезпечення практичного дотримання стандартів DQ у своїх предметних областях. Впровадження цієї двокомпонентної системи створює культуру власності та відповідальності за дані, що є критично важливим для підтримання довіри до інформації. Це також єдиний спосіб ефективно керувати походженням даних (Data Provenance), що є необхідним для аудиту, прозорості та дотримання складних регуляторних вимог [5].

Фокус на "Якість у джерелі" та Архітектурах потокової обробки. Найбільш ефективний і економічно вигідний метод подолання проблем DQ у Big Data полягає у переході від реактивного очищення до проактивного забезпечення якості, закладеного в концепції "Якість у джерелі" (Quality at Source). Історично дані збиралися, завантажувалися у сховища, а лише потім проводилося очищення. Однак у масштабах Big Data, де дефекти накопичуються з величезною швидкістю, такий реактивний підхід є надто повільним, дорогим і призводить до того, що Data Lakes перетворюються на "болота даних" (Data Swamps). Принцип "Якість у джерелі" вимагає, щоб валідаційні шлюзи, фільтри та бізнес-логіка впроваджувалися максимально близько до моменту створення даних, тобто на етапі первинного збору (Ingestion). Цей проактивний підхід реалізується завдяки сучасним архітектурам потокової обробки (Stream Processing), таким як Apache Kafka, Flink або Spark Streaming. Ці технології дозволяють здійснювати безперервну (in-stream) валідацію, трансформацію та фільтрацію даних у міру їхнього надходження. Якщо помилка виявляється (наприклад, невалідний формат або відсутнє критичне поле), дані можуть бути миттєво виправлені, відхилені або поміщені у карантин для подальшого ручного втручання. Ключова перевага полягає в тому, що дефекти не мають часу на поширення та "забруднення" основних сховищ, що забезпечує не лише високу якість, а й актуальність даних. Створення таких "Clean Data Pipelines" (чистих конвеєрів даних) є архітектурною необхідністю, яка підтримує надійність та ефективність усіх подальших аналітичних процесів та систем AI, забезпечуючи, що дані завжди "придатні до використання" (fit for use) [6].

Висновки

Проблеми якості даних (DQ) у Big Data є критичною перешкодою, спричиненою безпрецедентним масштабом, високою швидкістю та екстремальною різноманітністю даних, що робить традиційні методи контролю DQ застарілими. Як наслідок, низька якість даних прямо призводить до спотворення аналітичних результатів, неефективності моделей машинного навчання та значних бізнес-ризиків. Ефективне подолання цих викликів вимагає гібридної стратегії, яка охоплює кілька ключових напрямків. По-перше, необхідна технологічна інтеграція шляхом впровадження методів машинного навчання та штучного інтелекту для автоматизованого виявлення аномалій та інтелектуального очищення даних, масштабуючи управління якістю. По-друге, слід перейти до проактивної архітектури, орієнтованої на забезпечення "Якості у джерелі", інтегруючи валідаційні шлюзи в архітектури

потокової обробки для виявлення та корекції дефектів у реальному часі до їхнього накопичення. По-третє, ці кроки мають бути підкріплені організаційним контролем, а саме: впровадженням Data Governance для стратегічного визначення стандартів якості та Data Stewardship для забезпечення операційної відповідальності. Таким чином, управління DQ у Big Data є безперервною, інтегрованою дисципліною, що критично важлива для перетворення сирих даних на надійний стратегічний актив і успішного функціонування систем AI.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Зростаюча складність та багатовимірність проблем якості даних. URL: https://ceur-ws.org/Vol-2725/paper5.pdf (дата звернення: 25.10.2025).
2. Вплив низької якості даних на аналітичні результати та прийняття рішень. URL: menedzhment121-84-97.pdf (дата звернення: 25.10.2025).
3. Застосування машинного навчання та штучного інтелекту для автоматизованого управління якістю даних. URL: https://integralsolutions.pl/uk/ai-automatyzacja-zarzadzanie-danymi/ (дата звернення: 27.10.2025).
4. Data Governance. URL: https://cloud.google.com/learn/what-is-data-governance (дата звернення: 27.10.2025).
5. Data Stewards. URL: https://www.ibm.com/think/topics/data-stewardship (дата звернення: 28.10.2025).
6. Фокус на "Якість у джерелі" та Архітектурах потокової обробки. URL: https://www.researchgate.net/publication/326985824_Real-time_Data_Stream_Processing_-_Challenges_and_Perspectives (дата звернення: 29.10.2025).

Магденко Анастасія Романівна – студентка групи 1КІТС-22б, Факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, м. Вінниця, e-mail: anastasiiamahdenko@gmail.com

Науковий керівник: ***Грицак Анатолій Васильович*** – кандидат технічних наук, доцент, доцент кафедри менеджменту та безпеки інформаційних систем, Вінницький національний технічний університет, м. Вінниця, e-mail: grytsak.a.v@gmail.com

Mahdenko Anastasiia R. – student of group 1KITS-22b, Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia, e-mail: anastasiiamahdenko@gmail.com

Supervisor: ***Hrytsak Anatoly V.*** – Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Management and Security of Information Systems Vinnytsia National Technical University, Vinnytsia, e-mail: grytsak.a.v@gmail.com