УДК 004.622

КАЧЕСТВО ДАННЫХ В BIG DATA

А.И. Лордкипанидзе
Студент 4 курса, кафедры ПИКС
91430072@study.bsuir.by

В.Д. Владымцев
Ассистент кафедры информатики, инженер программист ОИАСУ ЦИИР БГУИР
v.vladymtsev@bsuir.by

А.Н. Марков
Старший преподаватель, заместитель начальника Центра информатизации и инновационных разработок БГУИР
a.n.markov@bsuir.by

А.И. Лордкипанидзе
Студент 4 курса "Информационные системы и технологии в бизнес-менеджменте" БГУИР.

В.Д. Владымцев
Ассистент кафедры информатики, инженер-программист ОИАСУ ЦИИР БГУИР

А.Н. Марков
Старший преподаватель кафедры ПОИТ, заместитель начальника Центра информатизации и инновационных разработок Белорусского государственного университета информатики и радиоэлектроники.

Аннотация. Качество данных является одним из ключевых факторов, влияющих на успешность обработки и использование больших объемов данных. В связи с этим вопросы качества данных в big data становятся все более актуальными. В данной аннотации будут рассмотрены основные проблемы, связанные с качеством данных в big data, а также методы и инструменты, используемые для их решения. Особое внимание будет уделено таким аспектам, как точность, полнота, актуальность, надежность, целостность и безопасность данных. Решение проблем, связанных с качеством данных в big data, требует комплексного подхода и использования различных методов и инструментов, включая методы проверки и очистки данных, автоматизированные процессы обновления данных, аутентификацию и контроль доступа, шифрование и многое другое.

Ключевые слова: big data, качество данных, точность данных, полнота данных, актуальность данных, надежность данных, целостность данных, безопасность данных, проверка данных, очистка данных, обновление данных, контроль доступа, шифрование данных.

Введение.

В последнее время обработка больших объемов данных, или big data [1], стала неотъемлемой частью работы многих компаний и организаций. Однако, обработка и использование таких больших объемов данных не всегда является простой задачей и может сопровождаться рядом проблем, которые могут повлиять на качество данных и результаты анализа. Одной из таких проблем является качество данных в big data. Качество данных играет важную роль в успешности обработки и использования данных, поэтому данной теме уделяется все больше внимания.

В данном тексте будет рассмотрено, что такое качество данных в big data, какие проблемы связаны с ним, а также какие методы и инструменты используются для обеспечения качества данных.

Качество данных в big data.

Качество данных [2] - одна из ключевых проблем, связанных с обработкой больших объемов данных (Big Data).

Неправильные данные могут привести к неправильным выводам и решениям, поэтому важно обеспечить высокое качество данных.

Таблица 1. Источники данных

Внутренние	Внешние
ERP [3]	Социальные сети
Классификаторы	Интернет
CRM	Специализированные DataSet

Ниже представлены основные из проблем, связанных с качеством данных, а также методы и инструменты, которые могут помочь решить эти проблемы.

Недостаточная точность.

Данные могут содержать ошибки, неточности и неточности, которые могут привести к неправильным выводам. Одним из способов решения этой проблемы является метод очистки данных, который позволяет идентифицировать и исправить ошибки в данных.

Этот процесс может включать в себя удаление дубликатов, заполнение отсутствующих значений, а также коррекцию ошибок.

Низкая связность данных.

Данные могут быть разбросаны по разным системам и приложениям, что затрудняет их интеграцию и анализ.

Для решения этой проблемы можно использовать методы интеграции данных, которые позволяют объединить данные из разных источников в единый набор данных. Этот процесс может включать в себя приведение данных к одному формату, обработку дубликатов и определение связей между данными.

Низкая полнота.

Данные могут быть неполными, что может привести к неправильным выводам и решениям. Одним из способов решения этой проблемы является метод заполнения данных, который позволяет заполнить отсутствующие значения на основе имеющихся данных.

Этот процесс может включать в себя использование методов машинного обучения и статистических методов для предсказания значений.

Низкая консистентность.

Данные могут быть несогласованными, что может привести к проблемам при интеграции и анализе данных. Для решения этой проблемы можно использовать методы управления данными, которые позволяют обеспечить согласованность данных в разных системах и приложениях.

Низкая актуальность.

Данные могут устареть, что может привести к неправильным выводам и решениям. Для решения этой проблемы можно использовать методы обновления данных, которые позволяют обновлять данные в реальном времени или периодически, чтобы обеспечить их актуальность.

Для этого могут использоваться различные методы, например, анализ изменений в источниках данных или настройка автоматических процессов обновления данных.

Недостаточная безопасность.

Обработка больших объемов данных может привести к угрозам безопасности, например, к краже данных, взлому системы или потере данных.

Для обеспечения безопасности данных можно использовать методы шифрования, аутентификации, авторизации и контроля доступа, а также использовать соответствующие инструменты и технологии для обнаружения и предотвращения угроз.

Сложность анализа.

Обработка больших объемов данных может быть очень сложной и требовать значительных вычислительных ресурсов. Для решения этой проблемы можно использовать

специализированные инструменты и технологии, такие как Hadoop [4], Spark [5] и другие, которые позволяют обрабатывать большие объемы данных быстро и эффективно. Пример использования технологии Hadoop представлен на рисунке 1.

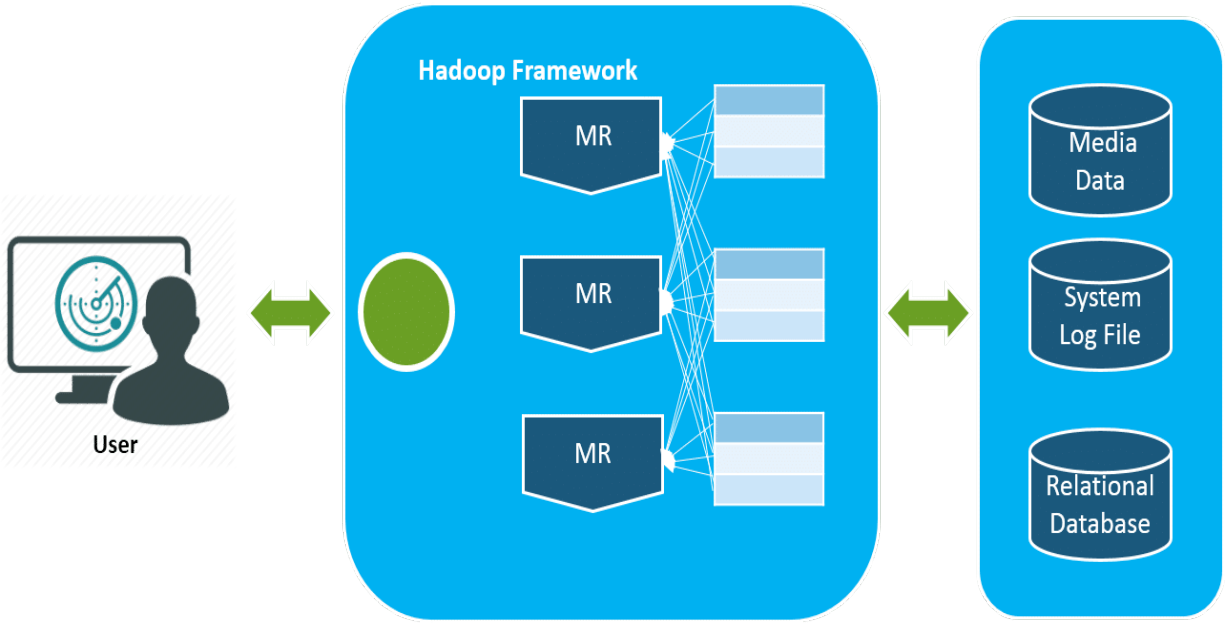

Рисунок 1. Пример использования технологии Hadoop.

Недостаточное понимание данных.

Важно иметь хорошее понимание данных, которые вы обрабатываете, чтобы сделать правильные выводы и принять правильные решения. Для этого необходимо проводить анализ данных, чтобы определить их структуру, связи, распределение и другие важные характеристики.

Кроме того, необходимо учитывать контекст, в котором были собраны данные, чтобы правильно интерпретировать их значения и использовать их для принятия решений.

Заключение.

Качество данных является важным фактором при обработке больших объемов данных. Проблемы, связанные с качеством данных, могут влиять на результаты анализа и приводить к некорректным выводам.

Однако, современные методы и инструменты позволяют решить многие из этих проблем. Кроме того, улучшение качества данных в big data может повысить эффективность бизнес-процессов и улучшить принятие решений.

Основными методами, используемыми для обеспечения качества данных в big data, являются проверка и очистка данных, обновление данных, контроль доступа, шифрование и многое другое. Однако, необходимо помнить, что обеспечение качества данных является процессом постоянного улучшения, и требует непрерывного мониторинга и анализа.

Список литературы

[1] Теоретический минимум по Big Data. Всё, что нужно знать о больших данных. — СПб.: Питер, 2019. — 208 с.: ил. — (Серия «Библиотека программиста»). ISBN 978-5-4461-1040-7.

[2] Редман, Т. К. Качество данных: полевое руководство. Издательство Digital Press., 2016 — 260 с.:

[3] Астапкина, К. С. Анализ и описание современных ERP-систем / Астапкина К. С. // Актуальные научные исследования в современном мире. – 2022. – Вып. 1(81), Ч. 10. – С. 164–166.

[4] Big Data: Concepts, Technology and Architecture / B. Balusamy [и др.]. – Hoboken : John Wiley & Sons, Inc., 2021. – 368 с

[5] Свирновский, А. В. Ключевые концепции Apache Spark / Свирновский А. В. // Новые информационные технологии в научных исследованиях: материалы XXV Юбилейной Всероссийской научно-технической конференции студентов, молодых ученых и специалистов, Рязань, 18-20 ноября 2020 г. / Рязанский государственный радиотехнический университет им. В. Ф. Уткина. – Рязань : ИП Коняхин А. В. (Book Jet), 2020 – С. 192 – 194.

DATA QUALITY IN BIG DATA

A.I. Lordkipanidze
4th year student, Department of PIKS

V.D. Vladymtsev
Assistant of the Department of Computer Science, Software Engineer of DIACS CIIR BSUIR

A.N. Markov
Senior lecturer of the department, Deputy head of the Center for Informatization and Innovative Developments

Center for Informatization and Development of the Belarusian University of State Informatics and Radioelectronics, Republic of Belarus
Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: 91430072@study.bsuir.by, v.vladymtsev@bsuir.by, a.n.markov@bsuir.by

Abstract. Data quality is one of the key factors that affects the success of processing and using large volumes of data, or big data. Therefore, issues related to data quality in big data are becoming increasingly relevant. This annotation examines the main problems related to data quality in big data, as well as the methods and tools used to address them. Special attention is paid to aspects such as accuracy, completeness, timeliness, reliability, integrity, and security of data. Solving problems related to data quality in big data requires a comprehensive approach and the use of various methods and tools, including data validation and cleansing methods, automated data update processes, authentication and access control, encryption, and much more.

Keywords: Data quality, Big data, Accuracy, Completeness, Timeliness, Reliability, Integrity, Security, Data validation, Data cleansing, Automated data update processes, Authentication, Access control, Encryption.