DOI: 10.65322/j.srd.04182026.07

АНАЛИЗ ИНСТРУМЕНТАЛЬНЫХ СРЕДСТВ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

Имамназаров Эркин Дехканалевич, *, к.п.н. доцент,*
Кафедра Цифровых технологий,
Инамова Гулигавхар Абдуллаевна, *старший преподаватель,*
Кафедра Цифровых технологий,
Наманганский Государственный технический университет

Аннотация: *В статье рассматриваются современные инструментальные средства обработки больших данных в информационных системах. Проведен анализ технологических платформ и программных решений, применяемых для хранения, обработки, анализа и визуализации больших массивов данных. Особое внимание уделено распределенным вычислительным системам, облачным технологиям, средствам интеллектуального анализа данных и инструментам машинного обучения. Выполнено сравнительное исследование наиболее распространенных платформ обработки Big Data, включая Hadoop, Apache Spark, NoSQL-системы и облачные сервисы. Выявлены преимущества, ограничения и особенности применения различных инструментальных средств в корпоративных информационных системах. Результаты исследования могут быть использованы при проектировании цифровой инфраструктуры организаций и выборе программно-технических решений для обработки больших данных.*

Abstract: *The article examines modern tools and technologies for big data processing in information systems. An analysis of technological platforms and software solutions used for storing, processing, analyzing, and visualizing large volumes of data is conducted. Particular attention is paid to distributed computing systems, cloud technologies, data mining tools, and machine learning instruments. A comparative study of the most widely used Big Data processing platforms, including Hadoop, Apache Spark, NoSQL systems, and cloud services, is carried out. The advantages, limitations, and application features of various tools in corporate information systems are identified. The research results can be used in designing the digital infrastructure of organizations and selecting software and hardware solutions for big data processing.*

Ключевые слова: *большие данные, Big Data, информационные системы, Hadoop, Apache Spark, NoSQL, машинное обучение, облачные технологии, аналитика данных.*

Keywords: *big data, Big Data, information systems, Hadoop, Apache Spark, NoSQL, machine learning, cloud technologies, data analytics.*

Введение. В условиях ускоренной цифровизации экономики и глобального роста объемов информации обработка больших данных становится одним из ключевых факторов повышения эффективности современных информационных систем. Ежедневно организации различных отраслей генерируют значительные объемы структурированных, полуструктурированных и неструктурированных данных, требующих высокопроизводительных методов хранения, обработки и анализа. Развитие технологий Интернета вещей, электронных платформ, мобильных приложений, облачных вычислений и искусственного интеллекта способствует дальнейшему увеличению объемов цифровой информации [1-3].

Традиционные методы управления базами данных оказываются недостаточно эффективными при работе с массивами данных высокой размерности и сложности. В связи с этим особую актуальность приобретают инструментальные средства обработки больших данных, обеспечивающие распределенное хранение информации, параллельную обработку данных и интеллектуальный анализ больших потоков информации [4-7].

DOI: 10.65322/j.srd.04182026.07

Современные платформы Big Data позволяют организациям осуществлять прогнозирование, оптимизацию бизнес-процессов, мониторинг поведения пользователей, автоматизацию управленческих решений и повышение конкурентоспособности. Внедрение инструментов обработки больших данных способствует формированию цифровой экономики и развитию интеллектуальных информационных систем [8-12].

Несмотря на значительное количество программных решений в области Big Data, проблема выбора оптимальных инструментальных средств остается актуальной. Различные платформы обладают неодинаковыми характеристиками производительности, масштабируемости, отказоустойчивости и стоимости внедрения. Это требует проведения комплексного анализа существующих технологий и оценки их применимости в информационных системах различного назначения.

Материалы и методы. Целью исследования является анализ современных инструментальных средств обработки больших данных в информационных системах, а также выявление их функциональных особенностей, преимуществ и ограничений. Для достижения поставленной цели в работе предусматривается исследование сущности и ключевых характеристик технологий Big Data, анализ современных платформ обработки данных, проведение сравнительной оценки инструментальных средств, а также определение перспектив развития технологий обработки больших данных в условиях цифровой трансформации информационных систем.

Методологической основой исследования выступили научные труды отечественных и зарубежных ученых в области цифровых технологий, распределенных вычислений и интеллектуального анализа данных. В процессе исследования использовались методы системного анализа, сравнительного анализа, классификации и обобщения.

Для анализа инструментальных средств обработки больших данных были рассмотрены наиболее распространенные программные платформы и технологии, применяемые в современных корпоративных информационных системах. В рамках исследования изучены функциональные возможности платформы Hadoop, вычислительной среды Apache Spark, NoSQL-систем управления базами данных, облачных сервисов обработки информации, а также инструментов машинного обучения и интеллектуальной аналитики данных.

В качестве критериев сравнительного анализа инструментальных средств обработки больших данных использовались показатели производительности обработки информации, масштабируемости и отказоустойчивости систем, скорости обработки потоковых данных, удобства интеграции с корпоративными информационными системами, стоимости внедрения и сопровождения, а также уровня поддержки аналитических и интеллектуальных инструментов обработки данных.

В ходе исследования также был проведен анализ архитектурных особенностей современных информационных систем, ориентированных на работу с большими данными.

По результатом исследования сущность и особенности технологий Big Data характеризует массивы данных, объем, скорость генерации и разнообразие которых превышают возможности традиционных средств обработки информации. Основными характеристиками больших данных являются так называемые «5V»:

Volume — большой объем данных;

Velocity — высокая скорость поступления информации;

Variety — разнообразие форматов данных;

Veracity — достоверность данных;

Value — ценность получаемой информации.

Большие данные формируются из различных источников: социальных сетей, интернет-платформ, мобильных устройств, банковских операций, датчиков Интернета

DOI: 10.65322/j.srd.04182026.07

вещей и корпоративных информационных систем. Эффективная обработка таких данных требует применения распределенных вычислительных архитектур и специализированных программных инструментов.

Одним из наиболее распространенных инструментальных средств обработки больших данных является платформа Hadoop. Данная технология представляет собой программный комплекс для распределенного хранения и обработки данных на кластерах серверов.

Архитектура Hadoop включает следующие основные компоненты Hadoop Distributed File System (HDFS),MapReduce,YARN, Hadoop Common.

Система HDFS обеспечивает распределенное хранение данных с высокой степенью отказоустойчивости. Технология MapReduce используется для параллельной обработки больших массивов информации путем разделения задач на отдельные вычислительные операции.

Преимуществами Hadoop являются высокая масштабируемость и возможность обработки петабайтных объемов данных, отказоустойчивость, поддержка распределенных вычислений, использование недорогого серверного оборудования.

Однако платформа Hadoop обладает и рядом недостатков. К ним относятся сравнительно низкая скорость обработки данных в режиме реального времени, сложность администрирования и высокие требования к квалификации специалистов.

Apache Spark представляет собой современную распределенную платформу обработки данных, ориентированную на высокопроизводительные вычисления. В отличие от Hadoop MapReduce, Spark выполняет значительную часть операций в оперативной памяти, что существенно увеличивает скорость обработки информации.

Платформа Spark включает следующие модули: Spark Core, Spark SQL, Spark Streaming, MLlib, GraphX.

Одним из ключевых преимуществ Apache Spark является поддержка обработки потоковых данных в режиме реального времени. Это особенно важно для современных информационных систем, функционирующих в условиях непрерывного поступления информации.

Сравнительный анализ показывает, что Spark превосходит Hadoop по скорости выполнения аналитических операций и обработке сложных вычислительных задач. Кроме того, Spark обеспечивает более удобную интеграцию с инструментами машинного обучения и искусственного интеллекта.

Вместе с тем использование Spark требует значительных объемов оперативной памяти и вычислительных ресурсов, что может увеличивать стоимость инфраструктуры.

Традиционные реляционные базы данных не всегда способны эффективно работать с большими объемами неструктурированной информации. В связи с этим широкое распространение получили NoSQL-системы.

К основным типам NoSQL-баз данных относятся: документоориентированные базы данных, колоночные базы данных, графовые базы данных, хранилища типа «ключ-значение».

Наиболее популярными решениями являются MongoDB, Cassandra, Redis и Neo4j.

MongoDB применяется для хранения полуструктурированных данных в формате JSON-подобных документов. Cassandra ориентирована на распределенное хранение данных с высокой отказоустойчивостью. Redis обеспечивает высокоскоростную обработку данных в оперативной памяти. Neo4j используется для анализа графовых структур и связей между объектами.

DOI: 10.65322/j.srd.04182026.07

Преимуществами NoSQL-систем являются высокая производительность, горизонтальное масштабирование и гибкость структуры данных, а также возможность работы с неструктурированной информацией.

Недостатками выступают ограниченная поддержка сложных транзакций и отсутствие единых стандартов реализации. Облачные технологии обработки больших данных.

Современные организации все активнее используют облачные платформы для обработки и хранения больших данных. Наиболее распространенными облачными сервисами являются: Amazon Web Services (AWS), Google Cloud Platform, Microsoft Azure.

Облачные технологии предоставляют организациям возможность масштабирования вычислительных ресурсов без необходимости приобретения дорогостоящего оборудования.

Основными преимуществами облачных решений являются: гибкость инфраструктуры, снижение капитальных затрат, высокая доступность сервисов, поддержка распределенных вычислений и интеграция с инструментами искусственного интеллекта.

В то же время использование облачных платформ связано с рисками информационной безопасности, зависимостью от поставщика услуг и необходимостью обеспечения защиты конфиденциальных данных.

Инструменты машинного обучения и аналитики данных является. Одним из наиболее перспективных направлений развития технологий Big Data является интеграция инструментов машинного обучения. Аналитические системы позволяют выявлять скрытые закономерности, прогнозировать события и автоматизировать принятие решений.

Наиболее распространенными инструментами аналитики являются: TensorFlow, PyTorch, Scikit-learn, RapidMiner, KNIME.

Использование алгоритмов машинного обучения обеспечивает: прогнозирование спроса, анализ поведения пользователей, выявление мошеннических операций, интеллектуальную обработку изображений и текстов, автоматизацию бизнес-процессов.

Интеграция аналитических инструментов с платформами обработки больших данных формирует основу современных интеллектуальных информационных систем.

Результаты. Результаты сравнительного анализа показывают, что выбор инструментальных средств зависит от особенностей информационной системы, объемов данных и целей обработки информации.

Hadoop эффективно применяется при обработке сверхбольших массивов данных и организации распределенного хранения информации. Apache Spark обеспечивает высокую скорость аналитической обработки и подходит для задач реального времени. NoSQL-системы ориентированы на хранение неструктурированных данных и высоконагруженные приложения. Облачные платформы позволяют быстро масштабировать инфраструктуру и снижать затраты на внедрение.

Для современных корпоративных информационных систем наиболее эффективным является комбинированный подход, предусматривающий интеграцию нескольких технологических решений.

Результаты исследования подтверждают, что развитие технологий Big Data оказывает существенное влияние на трансформацию информационных систем и цифровизацию экономики. Современные инструментальные средства позволяют организациям эффективно управлять информационными потоками, повышать качество аналитики и оптимизировать процессы принятия решений.

Внедрение технологий обработки больших данных сопровождается рядом организационно-технических и экономических проблем, связанных с высокими затратами на создание инфраструктуры, дефицитом квалифицированных специалистов и сложностью

DOI: 10.65322/j.srd.04182026.07

интеграции различных программных платформ. Кроме того, существенное значение приобретают вопросы обеспечения информационной безопасности и необходимость стандартизации процессов хранения, обработки и передачи данных в современных информационных системах.

Рост объемов цифровой информации повышает риски утечки данных и несанкционированного доступа к корпоративным ресурсам.

Перспективы развития технологий Big Data связаны с дальнейшим внедрением искусственного интеллекта, облачных вычислений и квантовых технологий. Ожидается, что интеллектуальные системы обработки данных будут играть ключевую роль в управлении цифровой экономикой, промышленностью, финансовым сектором и государственным управлением.

В условиях цифровой трансформации предприятий особую актуальность приобретает создание интегрированных аналитических платформ, обеспечивающих обработку больших данных в режиме реального времени. Это позволит повысить адаптивность организаций к изменениям внешней среды и укрепить их конкурентоспособность.

В ходе исследования был проведен анализ современных инструментальных средств обработки больших данных в информационных системах. Установлено, что технологии Big Data являются важнейшим элементом цифровой экономики и обеспечивают эффективное управление информационными ресурсами.

Выводы. Исследование показало, что наиболее распространенными платформами обработки больших данных выступают Hadoop, Apache Spark, NoSQL-системы и облачные сервисы. Каждое из рассмотренных решений обладает определенными преимуществами и ограничениями.

Платформа Hadoop обеспечивает высокую масштабируемость и распределенное хранение данных, Apache Spark отличается высокой скоростью обработки информации, NoSQL-системы обеспечивают гибкость работы с неструктурированными данными, а облачные технологии способствуют снижению затрат на инфраструктуру.

Установлено, что для повышения эффективности современных информационных систем целесообразно использовать комплексный подход, основанный на интеграции различных инструментальных средств обработки данных.

Перспективными направлениями дальнейших исследований являются разработка интеллектуальных методов анализа больших данных, совершенствование алгоритмов машинного обучения и повышение уровня информационной безопасности распределенных вычислительных систем.

Список литературы

1. Dean J., Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters // Communications of the ACM. – 2008. – Vol. 51. – №1. – P. 107–113.

2. White T. Hadoop: The Definitive Guide. – Sebastopol: O'Reilly Media, 2015. – 756 p.

3. Zaharia M., Chowdhury M., Franklin M. Spark: Cluster Computing with Working Sets // Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing. – 2010. – P. 1–7.

4. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. – Amsterdam: Elsevier, 2012. – 703 p.

5. Chen M., Mao S., Liu Y. Big Data: A Survey // Mobile Networks and Applications. – 2014. – Vol. 19. – №2. – P. 171–209.

6. Katal A., Wazid M., Goudar R. Big Data: Issues, Challenges, Tools and Good Practices // International Conference on Contemporary Computing. – 2013. – P. 404–409.

DOI: 10.65322/j.srd.04182026.07

7. Gandomi A., Haider M. Beyond the Hype: Big Data Concepts, Methods and Analytics // International Journal of Information Management. – 2015. – Vol. 35. – №2. – P. 137–144.

8. Marz N., Warren J. Big Data: Principles and Best Practices of Scalable Realtime Data Systems. – New York: Manning Publications, 2015. – 328 p.

9. Stonebraker M. SQL Databases v. NoSQL Databases // Communications of the ACM. – 2010. – Vol. 53. – №4. – P. 10–11.

10. Davenport T. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. – Boston: Harvard Business Review Press, 2014. – 240 p.

11. Инамова, Г. А., Кодиров, З. З., & Дадаханова, Д. (2022). Роль информационно-социальных технологий в образовании. In *Инновации в сельскохозяйственном машиностроении, энергосберегающие технологии и повышение эффективности использования ресурсов* (pp. 173-176).

12. Имамназаров Э., & Дехканов А. (2021). Использование и роль информационных технологий в нашей жизни. Экономика и социум, (6-1 (85)).