Секция «Математика, информационные технологии и их приложения»

Анализ методов обработки больших данных в информационных системах

Тавторкин Никита Олегович
Аспирант
Мордовский государственный университет им. Н.П. Огарёва, Факультет математики и информационных технологий, Саранск, Россия
E-mail: tavnick@mail.ru

В условиях цифровой трансформации экономики объемы генерируемых данных возрастают экспоненциально, что требует разработки и применения эффективных методов их обработки в информационных системах. Большие данные характеризуются тремя ключевыми свойствами: объемом, скоростью поступления и разнообразием структур [1]. Эти особенности обуславливают необходимость использования специализированных подходов к хранению, обработке и анализу данных.

Одним из базовых методов обработки больших данных является пакетная обработка, предполагающая накопление и последующую обработку данных за определённые интервалы времени [4]. Данный подход широко применяется для выполнения сложных аналитических задач и построения отчётности, обеспечивая высокую эффективность при работе с большими объёмами данных.

Альтернативой выступают потоковые методы обработки данных, ориентированные на работу с данными в реальном времени. Такие технологии, как Apache Kafka и Apache Flink, обеспечивают обработку непрерывных потоков данных с минимальными задержками, что особенно актуально для задач мониторинга, финансовых транзакций и анализа пользовательского поведения [3].

Значительное распространение получили методы обработки данных в распределенных хранилищах, включая колоночные базы данных и Data Lake-архитектуры. Колоночные СУБД, такие как ClickHouse, обеспечивают высокую производительность аналитических запросов за счет оптимизации хранения и сжатия данных. В свою очередь, Data Lake позволяет хранить данные в их исходном виде, обеспечивая гибкость последующего анализа.

Отдельное внимание уделяется методам машинного обучения, применяемым для извлечения знаний из больших данных. Алгоритмы классификации, кластеризации и регрессии позволяют выявлять скрытые закономерности и прогнозировать поведение систем [2]. Однако их эффективность напрямую зависит от качества предварительной обработки данных, включая очистку, нормализацию и агрегацию.

В результате исследования были систематизированы основные методы обработки больших данных: пакетная, потоковая и гибридная обработка, а также подходы с использованием распределённых хранилищ и методов машинного обучения. Определены их преимущества и ограничения в зависимости от требований к скорости, масштабируемости и типу данных. Установлено, что пакетные методы эффективны для аналитики, потоковые — для задач реального времени, а наилучшие результаты достигаются при использовании гибридных архитектур. Полученные выводы могут быть применены при проектировании информационных систем для работы с большими данными.

Источники и литература

1) Алетдинова А.А. Интеллектуальный анализ больших данных: учебное пособие. Новосибирск: НГТУ, 2023.
2) Баланов А.Н. Анализ данных: учебное пособие для СПО. Санкт-Петербург: Лань, 2026.

3) Косников С.Н., Золкин А.Л., Потехина Е.В. [и др.] Современные методы анализа данных в бизнес-аналитике: учебное пособие для вузов. Санкт-Петербург: Лань, 2026.

4) Парамонов А.А., Юрченков И.А., Крынецкий Б.А., Есипов И.В. Системы искусственного интеллекта и большие данные. Раздел «Большие данные»: учебное пособие. Москва: РТУ МИРЭА, 2025.