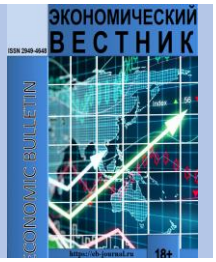
Научно-исследовательский журнал ***«Экономический вестник / Economic Bulletin»***
https://eb-journal.ru
2025, Том 4 № 5 2025, Vol. 4. Iss. 5 https://eb-journal.ru/archives/category/publications
Научная статья / Original article
УДК 659.118



[1] Хачатурова С.С.,
[1] Российский экономический университет имени Г.В. Плеханова

Большие данные: основные концепции и современные практики

Аннотация: настоящее время можно назвать эрой Больших данных, так как их потоки постоянно увеличиваются и влекут за собой возможности, влияющие на процесс сбора, хранения, обработки и анализа данных. Автор статьи отмечает, что объемы Больших данных на данный момент измеряются преимущественно в терабайтах (ТБ), петабайтах (ПБ), зеттабайтах (ЗБ) и этот показатель продолжает активно расти. В современном мире в большинстве сфер деятельности нереально избежать работы с *Большими данными. Большие данные* помогают улучшить качество услуг, оптимизировать производственные процессы, предсказать поведение потребителей и даже предотвратить катастрофы. Данные всегда играли важную роль [1]. Мы встречаемся с ними как пользователи интернета, а компании изучают с их помощью статистику продаж, поставок, деятельность конкурентов. Получается, что все вокруг используют *Большие данные.* Вместе с тем, в условиях активного развития цифровых технологий правовое регулирование *Больших данных* и проблемы признания их приобретают практическую актуальность [6]. Использование *Больших данных* ставит перед обществом важные вопросы, касающиеся защиты персональных данных, конфиденциальности и этики.

Автором отмечается, что термин *Большие данные* (Big data) сегодня продвигается как тренд в информационных технологиях нового века и определяет не только размер наборов данных, который превосходит возможности обычных баз данных (БД). Выделяются в статье, что вместе с потоком информации увеличивается потребность в ее хранении, структуризации, обработки и анализа [2]. Если ранее данные в сеть вносили люди, сейчас же для этого существует огромное количество программ, искусственный интеллект. Облачные платформы занимают все более важное место в стратегии работы с *Большими Данными*.

Ключевые слова: Большие данные, анализ данных, структуризация, обработка данных, искусственный интеллект; структурированные таблицы, поток информации; распределенные вычисления

Для цитирования: Хачатурова С.С. Большие данные: основные концепции и современные практики // Экономический вестник. 2025. Том 4. № 5. С. 105 – 109.

Поступила в редакцию: 2 августа 2025 г.; Одобрена после рецензирования: 3 октября 2025 г.; Принята к публикации: 15 ноября 2025 г.

[1] Khachaturova S.S.,
[1] Plekhanov Russian University of Economics

Big data: key concepts and modern practices

Abstract: the present era can be called the era of Big Data, as its volumes are constantly increasing, bringing with them opportunities that impact the process of collecting, storing, processing, and analyzing data. The author of the article notes that Big Data volumes are currently measured primarily in terabytes (TB), petabytes (PB), and zettabytes (ZB), and this figure continues to grow rapidly. In the modern world, it is impossible to avoid working with Big Data in most fields. Big Data helps improve the quality of services, optimize production processes, predict consumer behavior, and even prevent disasters. We encounter it as internet users, and companies use it to study sales statistics, deliveries, and competitor activities. It turns out that everyone is using Big Data. At the same time, the use of Big Data raises important questions for society regarding the protection of personal data, privacy, and ethics.

The author notes that the term *Big Data* is currently being promoted as a trend in new-age information technologies and defines not only the size of data sets, which exceeds the capabilities of conventional databases. The article notes that the flow of information increases the need for its storage, structuring, processing, and analysis [1]. While humans previously contributed data to the network, a vast array of programs and artificial intelligence now exist for this purpose. Cloud platforms are increasingly playing a key role in Big Data strategies.

Keywords: Big data, data analysis, structuring, data processing, artificial intelligence; structured tables, information flow, distributed computing

For citation: Khachaturova S.S. Big data: key concepts and modern practices. Economic Bulletin. 2025. 4 (5). P. 105 – 109.

The article was submitted: August 2, 2025; Approved after reviewing: October 3, 2025; Accepted for publication: November 15, 2025.

Введение

Термин *Большие данные* обычно относится к массивам данных, настолько обширным и сложным, что традиционные средства и методы обработки данных оказываются недостаточными. Они характеризуются набором параметров, часто называемых пятью V: *Объем, Скорость, Разнообразие, Достоверность и Ценность*.

Объем относится к огромному количеству данных, часто измеряемому в терабайтах, петабайтах или экзабайтах.

Скорость описывает скорость, с которой данные генерируются и требуют обработки в основном в режиме реального времени.

Разнообразие отражает различные формы данных: структурированные таблицы, полуструктурированные журналы, неструктурированный текст, изображения, аудиозаписи и видео.

Достоверность указывает на неопределенность и противоречивость данных и связана с качеством и её точностью.

Ценность подчеркивает конечную цель - полезность этих данных для принятия решений, преобразование необработанных данных в выводы и обоснованные действия.

Мир *Больших данных* становится только еще больше [4]. В цифровую эпоху данные превратились в главный актив, имеющий как экономическую, так и технологическую ценность. Каждый момент времени огромные объемы информации поступают из самых разных источников: социальных сетей, финансовых транзакций, онлайн-поиска, мобильных приложений, датчиков Интернета вещей (IoT), спутниковых систем и корпоративных систем, и это лишь некоторые из них.

Активное развитие информационных технологий имеет побочный эффект: объем данных в мире растет экспоненциально, а сами данные становятся все менее структурированными [3]. Проблема, которую она представляет, заключается не просто в хранении больших объемов данных, а в способности извлекать из них ценность путем своевременной, масштабируемой и эффективной обработки. Эта задача является технической по своей сути и привела к разработке различных архитектурных подходов, разработанных специально для удовлетворения уникальных требований *Больших Данных*.

Традиционные централизованные системы баз данных, которые доминировали в эпоху *Больших данных*, не были рассчитаны на работу в таких условиях. Реляционные базы данных, хотя и остаются основополагающими во многих контекстах, испытывают трудности с горизонтальным масштабированием, обработкой разнородных форматов и получением высокоскоростных потоков данных. Это привело к переходу к архитектурам распределенных вычислений, в которых хранение и обработка данных распределены по кластерам машин. Такие архитектуры не только обеспечивают горизонтальное масштабирование, но и обеспечивают отказоустойчивость, улучшенный параллелизм и модульную интеграцию компонентов, предназначенных для определенных этапов жизненного цикла данных.

Материалы и методы исследований

Большие данные влияют на процесс принятия решений [9]. Большинство современных архитектур построены по принципу многоуровневой или модульной структуры, что обеспечивает гибкость и масштабируемость.

На начальном этапе происходит захват данных – процесс получения и импорта данных из различных источников. Этими источниками могут быть потоки в реальном времени, такие как потоки кликов и телеметрические данные, или пакетные данные из исторических баз и хранилищ данных. Такие инструменты, как Apache Kafka, Apache NiFi и Flume, широко используются для получения высокопроизводительных потоков данных, обеспечивая долговечность, разделение и отказоустойчивость. Для пакетного ввода данных из структури-

рованных систем, таких как реляционные базы данных, обычно используются такие инструменты, как Apache Sqoop.

После ввода данных следующим уровнем является их хранение [5]. Учитывая разнообразие и объем *Больших Данных*, обычные файловые системы оказываются неадекватными. Для решения этой проблемы были разработаны распределенные файловые системы, в первую очередь Hadoop Distributed File System (HDFS). HDFS разбивает большие файлы на более мелкие блоки и хранит их с избыточностью на нескольких узлах, обеспечивая надежность даже в условиях аппаратных сбоев [2]. В облачных средах такие платформы, как Amazon S3, Google Cloud Storage и Azure Data Lake, предлагают масштабируемые объектные хранилища с высокой доступностью. Дополнением к ним являются базы данных NoSQL, такие как Apache Cassandra, MongoDB, HBase и Amazon DynamoDB, которые отказываются от жестких определений схем в пользу гибких, оптимизированных по производительности моделей данных. Эти базы данных особенно хорошо подходят для хранения неструктурированных или полуструктурированных данных в масштабе [2].

После хранения данных главной задачей становится их обработка [7]. На этом этапе данные очищаются, преобразуются, агрегируются и подготавливаются к анализу. На этом уровне преобладают две основные парадигмы: пакетная и потоковая обработка.

Пакетная обработка работает с большими объемами данных, накопленных за определенное время. Она идеально подходит для сложных преобразований и масштабной аналитики, не требующей немедленных результатов. Первоначальная модель MapReduce в Apache Hadoop стала пионером этого подхода, но в итоге ее опередил Apache Spark, который выполняет вычисления в памяти и значительно сокращает время ожидания. Механизм Spark, основанный на Directed Acyclic Graph (DAG), позволяет оптимизировать выполнение заданий, что делает его предпочтительным выбором для итеративных алгоритмов и рабочих процессов машинного обучения.

Потоковая обработка, напротив, предназначена для обработки данных в реальном или близком к реальному времени. Некоторые системы, такие как Apache Storm, Flink и Kafka Streams, специально созданы для обработки непрерывных потоков событий сразу после их возникновения. Это критично для приложений, где важна каждая миллисекунда – например, в борьбе с мошенничеством, системах персональных рекомендаций, мониторинге телеметрии и финансовых торгах. Такие решения позволяют не просто обрабатывать поток данных, но и анализировать временные зависимости, выстраивать динамичные окна событий и выявлять сложные закономерности, что делает их ключевыми элементами в архитектурах, опирающихся на события.

Однако обработка потока – это лишь часть картины. Современные системы анализа *Больших данных* тесно взаимодействуют с инструментами для машинного обучения. Библиотеки вроде Spark MLlib, Scikit-learn и TensorFlow дают возможность строить модели прямо на масштабных данных. Эти инструменты охватывают весь путь – от обучения и тестирования до внедрения систем предсказания, классификации, кластеризации и обработки текста.

После того как получены аналитические выводы, важно представить их в понятной и наглядной форме. Здесь на сцену выходят BI-инструменты: Tableau, Power BI, Apache Superset, Google Data Studio. Они позволяют создавать визуальные дашборды, следить за метриками в реальном времени и строить удобные отчёты. Эти решения делают данные доступными для широкого круга пользователей – от аналитиков до руководства, помогая принимать решения на основе фактов, а не предположений.

Чтобы эффективно связать все эти компоненты, была разработана целая экосистема архитектурных подходов. Одним из самых известных стал подход Lambda, объединяющий два уровня: пакетную обработку для глубокого анализа накопленных данных и потоковую – для оперативного реагирования. Результаты объединяются и подаются конечным пользователям в виде единого источника информации. Однако такая модель требует дублирования логики для разных слоёв, что усложняет сопровождение.

В качестве альтернативы появилась архитектура Kappa, которая рассматривает данные как единый поток, независимо от их возраста. В ней используется единый конвейер обработки как для новых, так и для старых данных, что упрощает реализацию и снижает издержки. Такой подход особенно удобен там, где важна скорость реакции, а историческая аналитика не требует отдельной обработки.

Результаты и обсуждения

Эти архитектурные подходы поддерживаются различными технологическими экосистемами. Например, экосистема Hadoop сегодня выходит далеко за рамки своей первоначальной модели MapReduce и включает в себя широкий спектр инструментов для хранения, обработки и управления данными. В нее входят Hive для SQL-подобных

запросов к большим наборам данных, Pig для создания сценариев потоков данных, Oozie для планирования рабочих процессов и Zookeeper для координации работы распределенных систем. Тем временем Apache Spark вырос в собственную экосистему, поддерживающую потоковую обработку, структурированные запросы (Spark SQL), машинное обучение и графовую аналитику.

Amazon Web Services (AWS), Google Cloud Platform (GCP) и Microsoft Azure предлагают сервисы, адаптированные к каждому уровню архитектуры. Такие сервисы, как EMR, Kinesis и Redshift от AWS, BigQuery и Dataflow от GCP, HDInsight и Synapse Analytics от Azure, позволяют организациям легко масштабироваться, интегрировать передовую аналитику и снижать операционную сложность. Эти платформы часто поддерживают гибридное развертывание, позволяя компаниям сочетать локальные ресурсы с облачными сервисами для важных рабочих нагрузок или устаревших систем.

Архитектурный ландшафт продолжает развиваться. Озера данных – гибрид озер данных и хранилищ данных – призваны обеспечить производительность и возможности управления хранилищами, а также масштабируемость и гибкость озер. Такие платформы, как Databricks и Snowflake, стимулируют этот сдвиг, позволяя проводить аналитику в стиле SQL над сырыми и полуструктурированными данными в распределенных хранилищах. Кроме того, все большее распространение бессерверных вычислений и контейнеризации (с помощью таких инструментов, как Kubernetes и Docker) меняет стратегии развертывания и масштабирования.

Выводы

Таким образом, под *Большими данными* понимаются разнообразные данные, поступающие с высокой скоростью, объем которых постоянно растет [8]. Однако, несмотря на все достижения, остаются серьезные проблемы. Обеспечение качества данных, защита доступа, управление данными и контроль операционных расходов являются постоянными проблемами. Более того, этические и нормативные соображения, такие как GDPR и законы о суверенитете данных, добавляют новые уровни сложности. Архитектуры теперь должны разрабатываться не только для обеспечения производительности и надежности, но и для обеспечения прозрачности, соответствия и возможности аудита.

В заключение следует отметить, что архитектура систем *Больших Данных* – это динамичная область, формирующаяся под влиянием требований масштаба, скорости и разнообразия. Эволюция этих архитектур – от ранних распределенных систем до облачных нативных конвейеров реального времени – отражает более широкие преобразования в том, как организации понимают и используют информацию. Выбор архитектуры зависит от конкретных потребностей приложения: в одних случаях требуется аналитика с низкой задержкой, в других приоритет отдается долгосрочному хранению и сложным вычислениям. Но основные принципы остаются неизменными: предоставление организациям возможности извлекать своевременные, достоверные и действенные сведения из постоянно растущих потоков данных.

Большие данные определяют не только размер наборов данных, превосходящий возможности обычных баз данных, но и неструктурированную информацию, перед обработкой и анализом которой бессильны традиционные методы [2].

Большие данные помогают улучшить качество услуг, оптимизировать производственные процессы, предсказать поведение потребителей и даже предотвратить катастрофы. БД могут влиять на процесс законодательной деятельности по различным направлениям [10]. Вместе с тем, использование *Больших данных* ставит перед обществом важные вопросы, касающиеся защиты персональных данных, конфиденциальности и этики. Необходимо развивать технологии и законы, чтобы обеспечить баланс между инновациями и правами личности. Хранилища служат неким фундаментом для обработки информации, а значит, исследователям стоит обратить внимание на разработку качественных баз, которые смогут выдерживать потоки, структурировать их, а главное, делать это как можно с большей скоростью. На данный момент, ни одна из систем хранения и обработки баз данных не дает разрешения всех возможных проблем, но все же, развитие продолжается, а значит, возможно, в скором времени человечество ждет значительный прорыв в этой сфере.

Список источников

1. Безпалов В.В., Лочан С.А., Федюнин Д.В. и др. Большие данные и возможности их использования при разработке коммуникативной стратегии предприятий регионального промышленного комплекса // Вестник Алтайской академии экономики и права. 2020. № 1-2. С. 28 – 34.

2. Денисова О.Ю., Мухутдинов Э.А. Большие данные – это не только размер данных [Электронный ресурс]. Режим доступа: (дата обращения: 09.07.2025) https://cyberleninka.ru/article/n/bolshie-dannye-eto-ne-tolko-razmer-dannyh

3. Как извлечь большую выгоду из «Больших данных»? [Электронный ресурс]. Режим доступа: https://bosfera.ru/bo/kak-izvlech-bolshuyu-vygodu-iz-bolshih-dannyh (дата обращения: 08.07.2025)

4. Hadoop Distributed File System. [Электронный ресурс]. Режим доступа: https://habr.com/ru/articles/42858/ (дата обращения: 08.07.2025)

5. 17 лучших инструментов и технологий для работы с большими данными. [Электронный ресурс]. Режим доступа: https://habr.com/ru/companies/otus/articles/659657/ (дата обращения: 08.07.2025)

6. Тагаева С.Н., Гатиятуллина Э.М. Большие данные и персональные данные: правовая природа и вопросы регулирования // Цифровое право. 2024. Т. 5. № 2. С. 40 – 52.

7. Смелова А.А. Большие данные и/или малые данные: методы науки о социальных данных для прикладных социально-экономических исследований // Городской социологический семинар: материалы заседаний 2023 года. Научный и технологический суверенитет современных обществ. Санкт-Петербург: ООО "Медиапапир", 2024. С. 22 – 24.

8. Тыщенко Д.Э., Хорожев Г.О. Анализ больших данных: статистические подходы к обработке и анализу больших данных // Аллея науки. 2024. Т. 1. № 11 (98). С. 3 – 14.

9. Rahel D.R. Features of searching for associations in large volumes of marketing data // 17-18 мая 2023 года, 2023. P. 18 – 20.

10. Tukhtasinov A.I. V.S. Sodikov Fractal modeling of big data // 17-18 мая 2023 года, 2023. P. 65 – 67.

References

1. Bezpalov V.V., Lochan S.A., Fedyunin D.V., et al. Big Data and the Possibilities of Their Use in Developing a Communication Strategy for Enterprises of the Regional Industrial Complex. Bulletin of the Altai Academy of Economics and Law. 2020. No. 1-2. P. 28 – 34.

2. Denisova O.Yu., Mukhutdinov E.A. Big Data Is Not Just About the Size of the Data [Electronic resource]. Access mode: (date of access: 07.09.2025) https://cyberleninka.ru/article/n/bolshie-dannye-eto-ne-tolko-razmer-dannyh

3. How to Get More Benefit from Big Data? [Electronic resource]. Access mode: https://bosfera.ru/bo/kak-izvlech-bolshuyu-vygodu-iz-bolshih-dannyh (date of access: 07.09.2025)

4. Hadoop Distributed File System [Electronic resource]. Access mode: https://habr.com/ru/articles/42858/ (date of access: 07.09.2025 (date of access: 07.08.2025)

5. 17 Best Tools and Technologies for Working with Big Data. [Electronic resource]. Access mode: https://habr.com/ru/companies/otus/articles/659657/ (date of access: 07.08.2025)

6. Tagaeva S.N., Gatiyatullina E.M. Big Data and Personal Data: Legal Nature and Regulation Issues. Digital Law. 2024. Vol. 5. No. 2. P. 40 – 52.

7. Smelova A.A. Big Data and/or Small Data: Methods of Social Data Science for Applied Socioeconomic Research. City Sociological Seminar: Proceedings of the 2023 Meetings. Scientific and Technological Sovereignty of Modern Societies. St. Petersburg: OOO "Mediapapir", 2024. P. 22 – 24.

8. Tyshchenko D.E., Khorozhev G.O. Big Data Analysis: Statistical Approaches to Processing and Analysis of Big Data. Alley of Science. 2024. Vol. 1. No. 11 (98). P. 3 – 14.

9. Rahel D.R. Features of searching for associations in large volumes of marketing data. May 17-18, 2023, 2023. P. 18 – 20.

10. Tukhtasinov A.I. V.S. Sodikov Fractal modeling of big data. May 17-18, 2023, 2023. P. 65 – 67.

Информация об авторе

Хачатурова С.С., кандидат экономических наук, доцент, Российский экономический университет имени Г.В. Плеханова