

УДК 004.042

5.2.2. Математические, статистические и инструментальные методы в экономике (физико-математические науки, экономические науки)

АНАЛИЗ ДАННЫХ В РЕАЛЬНОМ ВРЕМЕНИ: ПРЕИМУЩЕСТВА И ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ STREAM PROCESSING

Кушнир Надежда Владимировна
старший преподаватель кафедры информационных систем и программирования
РИНЦ-SCIENCE INDEX. SPIN-код=6951-4012
kushnir.06@mail.ru
*ФГБОУ ВО “Кубанский государственный технологический университет”, Краснодар, Россия
350020, улица Московская, 2, Краснодар, Россия*

Вонарх Юлия Сергеевна
старший преподаватель кафедры информационных систем и программирования
Jul_nosova@mail.ru
*ФГБОУ ВО “Кубанский государственный технологический университет”, Краснодар, Россия
350020, улица Московская, 2, Краснодар, Россия*

Деев Илья Максимович
студент бакалавр
ilya.deyev@gmail.com
*ФГБОУ ВО “Кубанский государственный технологический университет”, Краснодар, Россия
350020, улица Московская, 2, Краснодар, Россия*

Долунц Анастасия Станиславовна
студент бакалавр
11naaastiaaa@gmail.com
*ФГБОУ ВО “Кубанский государственный технологический университет”, Краснодар, Россия
350020, улица Московская, 2, Краснодар, Россия*

Карпенко Никита Андреевич
студент бакалавр
1knyazzz1@gmail.com
*ФГБОУ ВО “Кубанский государственный технологический университет”, Краснодар, Россия
350020, улица Московская, 2, Краснодар, Россия*

Данная статья представляет собой исчерпывающее исследование современных технологий и подходов к обработке данных. Автор акцентирует внимание на важности потоковой обработки в условиях, когда объемы информации стремительно растут. Преимущества, такие как высокая скорость анализа и снижение затрат на хранение, подчеркивают необходимость перехода на новые методы работы с данными. Примеры применения в различных отраслях, включая финансы и Интернет вещей, наглядно демонстрируют актуальность и универсальность технологий потоковой обработки.

UDC 004.042

5.2.2. "Mathematical, statistical and instrumental methods in economics" (physical and mathematical sciences, economic sciences)

REAL-TIME DATA ANALYSIS: ADVANTAGES AND EXAMPLES OF USING STREAM PROCESSING

Kushnir Nadezhda Vladimirovna
Senior Lecturer at the Department of Information Systems and Programming
RSCI-SCIENCE INDEX. SPIN code=6951-4012
kushnir.06@mail.ru
*Kuban State Technological University, Krasnodar, Russia
350020, Moskovskaya, 2, Krasnodar, Russia*

Vonarch Yulia Sergeevna
Senior Lecturer at the Department of Information Systems and Programming
Jul_nosova@mail.ru
*Kuban State Technological University, Krasnodar, Russia
350020, Moskovskaya, 2, Krasnodar, Russia*

Deev Ilya Maksimovich
bachelor's student
ilya.deyev@gmail.com
*Kuban State Technological University, Krasnodar, Russia
350020, Moskovskaya, 2, Krasnodar, Russia*

Dolunts Anastasia Stanislavovna
bachelor's student
11naaastiaaa@gmail.com
*Kuban State Technological University, Krasnodar, Russia
350020, Moskovskaya, 2, Krasnodar, Russia*

Karpenko Nikita Andreevich
bachelor's student
1knyazzz1@gmail.com
*Kuban State Technological University, Krasnodar, Russia
350020, Moskovskaya, 2, Krasnodar, Russia*

This article is an exhaustive study of modern technologies and approaches to data processing. The author focuses on the importance of streaming processing in an environment where the volume of information is growing rapidly. Advantages such as high-speed analysis and lower storage costs emphasize the need to switch to new methods of working with data. Examples of applications in various industries, including finance and the Internet of Things, clearly demonstrate the relevance and versatility of streaming technologies. These technologies, such as Apache Kafka and Amazon Kinesis, together with a detailed

Приведенные технологии, такие как Apache Kafka и Amazon Kinesis, вместе с подробным анализом их возможностей и областей применения, делают материал полезным для всех типов аудитории

analysis of their capabilities and applications, make the material useful for all types of audiences

Ключевые слова: ПОТОКОВАЯ ОБРАБОТКА, ДАННЫЕ В РЕАЛЬНОМ ВРЕМЕНИ, АНАЛИЗ ДАННЫХ, ИНТЕРНЕТ ВЕЩЕЙ, ФИНАНСОВЫЕ УСЛУГИ, АЛГОРИТМИЧЕСКАЯ ТОРГОВЛЯ, ОБНАРУЖЕНИЕ МОШЕННИЧЕСТВА, АРАСНЕ КАФКА, АРАСНЕ FLINK, АРАСНЕ SPARK STREAMING, AMAZON KINESIS, СИСТЕМЫ МОНИТОРИНГА, ХРАНЕНИЕ ДАННЫХ, ТЕХНОЛОГИИ ПОТОКОВОЙ ОБРАБОТКИ, ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ, МАСШТАБИРУЕМОСТЬ, ГИБКОСТЬ, ПРОИЗВОДИТЕЛЬНОСТЬ, АНОМАЛИИ ДАННЫХ, ТЕЛЕМЕТРИ

Keywords: STREAMING, REAL-TIME DATA, DATA ANALYSIS, INTERNET OF THINGS, FINANCIAL SERVICES, ALGORITHMIC TRADING, FRAUD DETECTION, APACHE KAFKA, APACHE FLINK, APACHE SPARK STREAMING, AMAZON KINESIS, MONITORING SYSTEMS, DATA STORAGE, STREAMING TECHNOLOGIES, USE CASES, SCALABILITY, FLEXIBILITY, PERFORMANCE, DATA ANOMALIES, TELEMETRY

<http://dx.doi.org/10.21515/1990-4665-207-048>

Введение.

В современном мире данные превращаются в один из самых ценных ресурсов. С каждым годом объем информации, генерируемой различными источниками, такими как веб-приложения, сенсоры устройства Интернета вещей (IoT), становятся огромными. В условиях, когда данные требуют немедленной обработки, традиционные методы пакетной обработки становятся неэффективными. Метод потоковой обработки (stream processing) предоставляет возможность анализа данных в реальном времени, обеспечивая быстрые и точные результаты. В этой статье мы рассмотрим основные преимущества потоковой обработки данных, примеры её применения, а также технологии и инструменты, используемые в этой области.

Потоковая обработка данных – это подход, при котором данные анализируются и обрабатываются по мере их поступления в систему [1]. В отличие от пакетной обработки, где данные собираются и обрабатываются большими объемами, потоковая обработка обеспечивает моментальное получение результатов. Такой подход стал особенно актуален в таких

сферах, как финансы, телеметрия и Интернет вещей, где требуется быстрая реакция на изменения.

Принципы потоковой обработки [2]:

- непрерывность: потоковая обработка работает с данными, поступающими непрерывно, что позволяет получать результаты в режиме реального времени;

- обработка поэтапно: данные обрабатываются поэтапно, что подразумевает возможность анализа различных аспектов данных на каждом этапе;

- масштабируемость: системы потоковой обработки могут легко масштабироваться для обработки больших объёмов данных.

Производительность и метрики потоковой обработки

Одним из ключевых параметров потоковой обработки является время задержки (Latency), которое можно рассчитать по следующей формуле (1):

$$L = \frac{T_p + T_q}{N} \quad (1)$$

где

L – время необходимое для обработки одного события;

T_p – время, затраченное на фактическую обработку каждого события;

T_q – время, которое событие проводит в очереди перед началом обработки;

N – общее количество обработанных событий.

Пример использования: если система обрабатывает 100 событий, общее время обработки составляет 200 секунд, а суммарное время ожидания в очереди - 50 секунд, то время задержки будет рассчитано следующим образом:

$$L = \frac{200 + 50}{100} = 2,5$$

Это означает, что каждое событие в среднем подвергается задержке в 2,5 секунды, что может быть критически важно для многих приложений.

Преимущества потоковой обработки данных включают

1. Быстрота анализа данных. Возможность анализа данных по мере их поступления позволяет организациям мгновенно реагировать на изменения и принимать решения в реальном времени. Это особенно критически важно в финансовом секторе, где торговцы используют потоковую обработку для анализа цен на инструменты мгновенно.

2. Увеличение производительности. Системы потоковой обработки способны обрабатывать данные параллельно. Операции, такие как фильтрация, агрегация и сложные вычисления, могут выполняться одновременно, что значительно ускоряет общий процесс анализа.

3. Снижение затрат на хранение. Потоковая обработка обеспечивает удаление данных по мере их анализа, что значительно снижает необходимость в хранении больших объемов неактуальной информации. Таким образом, уменьшаются затраты на хранение данных.

4. Повышенная точность и качество данных. Возможность анализа больших объемов данных в реальном времени позволяет моментально выявлять аномалии и ошибки. Это критически важно в таких отраслях, как здравоохранение, где любое замедление может привести к нежелательным последствиям.

5. Гибкость и адаптируемость. Потоковые обработки могут быстро адаптироваться к изменениям условий, что делает их особенно подходящими для динамичных сред, таких как ритейл или интернет-маркетинг. Организации могут быстро реагировать на текущие события и изменять свои стратегии.

Потоковая обработка данных используется в различных отраслях, и можно выделить несколько примеров, где эта технология продемонстрировала свои преимущества.

Вызовы и сложности потоковой обработки

Несмотря на значительные преимущества потоковой обработки данных, этот подход также сталкивается с рядом вызовов и сложностей. Рассмотрим наиболее важные аспекты:

1. Обеспечение надежности:

- Проблемы с потерей данных: в условиях высокой скорости обработки данных существует риск потери информации. Необходимо использовать продуманные механизмы резервного копирования и восстановления.

- Отказоустойчивость: системы потоковой обработки должны быть способны продолжать работу даже в случае частичных сбоев или ошибок. Это требует разработки избыточности и устойчивости к сбоям.

2. Управление потоком данных:

- Переполнение системы: при резком увеличении потока данных система может оказаться перегруженной, что приведет к задержкам и падению производительности. Необходимо заранее предусмотреть возможность масштабирования и настройки системы в зависимости от нагрузки.

- Распределение нагрузки: эффективное распределение данных по различным компонентам системы является критически важным для стабильной работы. Необходимо разработать алгоритмы, способные равномерно распределять загрузку.

3. Безопасность данных:

- Угрозы на уровне сети: потоковые системы подвержены различным угрозам безопасности, таким как атаки «отказ в обслуживании» (DoS),

которые могут нарушить работу системы. Необходимы комплексные меры по защите данных и инфраструктуры.

- Конфиденциальность данных: обработка потоковых данных требует обеспечения защиты конфиденциальной информации. Это накладывает дополнительные требования к шифрованию и контролю доступа к данным.

4. Сложность разработки и интеграции:

- Нужда в специализированных навыках: разработка и поддержка систем потоковой обработки требуют наличия высококвалифицированных специалистов, знакомых с архитектурой и технологиями обработки данных.

- Интеграция с существующими системами: интеграция новых потоковых платформ с уже существующими решениями может быть сложной задачей. Необходимо учитывать совместимость различных технологий и форматов данных.

5. Сложность надежного тестирования:

- Динамическое тестирование: тестирование потоковых приложений требует от разработчиков использования специализированных инструментов, поскольку поведение систем может изменяться в зависимости от текущей нагрузки и состояния сети.

- Моделирование в реальном времени: создание тестовых сред, имитирующих реальную работу системы, может быть сложной задачей, так как требует учёта множества факторов, в том числе распределенных потоковых данных и Variable Change-90.

6. Управление состоянием:

- Сохраняемое состояние: потоковые приложения могут потребовать хранения состояния для обработки событий. Правильное управление

состоянием и его синхронизация между разными компонентами системы могут стать вызовом.

7. Обработка и анализ в реальном времени:

- Проблемы с задержкой: в реальных условиях сложно контролировать время задержки (latency) обработки данных. Необходимы стратегии оптимизации, чтобы минимизировать время от получения данных до получения результата.

- Интеграция с аналитическими инструментами: для полной пользы от потоковой обработки необходимо эффективно интегрировать её с инструментами анализа данных и визуализации, что может потребовать дополнительных усилий.

Преодоление этих вызовов требует комплексного подхода, включая разработку надежной архитектуры, адекватные механизмы защиты данных, активное управление нагрузками и инвестирование в обучение сотрудников. И только тогда потоковая обработка сможет полностью раскрыть свой потенциал, предлагая организациям уникальные возможности для быстрого реагирования на изменения и принятия решений на основе данных в реальном времени.

Примеры применения потоковой обработки

Финансовые услуги.

В финансовом секторе потоковая обработка данных помогает создавать торговые алгоритмы, которые реагируют на мгновенные изменения цен на финансовых рынках [3]. В таблице 1 показано, что данные системы используют потоковые данные для обнаружения мошенничества и мониторинга транзакций.

Таблица 1 – Примеры использования потоковой обработки данных в сфере финансовых услуг.

Применение	Пример
Алгоритмическая торговля	Реакция на изменение цен акций в режиме реального времени
Обнаружение мошенничества	Автоматическое выявление подозрительных транзакций

Интернет вещей (IoT).

Устройства IoT генерируют постоянный поток данных, который требует быстрого анализа [4]. Потоковая обработка позволяет получать данные от сенсоров в реальном времени, что может использоваться для мониторинга состояния оборудования, контроля температуры и повышения безопасности, что показано в таблице 2.

Таблица 2 – Примеры использования потоковой обработки данных с использованием IoT.

Применение	Пример
Умные дома	Автоматизация управления освещением и климатом на основе данных от сенсоров
Промышленность	Мониторинг состояния оборудования в реальном времени для предотвращения аварий

Социальные медиа.

Потоковая обработка активно используется в социальных медиа для анализа пользовательского контента и выявления трендов. Эта информация может быть использована для адаптации маркетинговых стратегий и улучшения взаимодействия с клиентами, как показано в таблице 3.

Таблица 3 – Примеры использования потоковой обработки данных в социальных медиа

Применение	Пример
Анализ трендов	Выявление популярных хештегов и обслуживание на Twitter
Прогнозирование настроений	Анализ тональности постов пользователей

Телеметрия и мониторинг.

В таблице 4 показано как потоковая обработка данных в телеметрии позволяет осуществлять мониторинг систем в реальном времени и обнаруживать неисправности.

Таблица 4 – Примеры использования потоковой обработки данных в телеметрии с использованием мониторинга систем в реальном времени.

Применение	Пример
Авиаперевозки	Мониторинг состояния самолётов и обеспечения безопасности полётов
Автомобили	Сбор данных о состоянии двигателей и системы безопасности

Технологии потоковой обработки

Существует множество инструментов и технологий для реализации потоковой обработки данных [5]. Некоторые из самых популярных включают:

1. Apache Kafka — это распределенная платформа потоковой обработки, позволяющая эффективное взаимодействие между различными источниками и системами потоковой обработки. Kafka может обрабатывать миллионы событий в секунду и крайне надёжен.

Преимущества:

- высокая пропускная способность;
- надёжность и отказоустойчивость;
- подходит для масштабируемых систем.

Области применения Apache Kafka:

- обработка потоковых данных в реальном времени;
- создание систем мониторинга с уведомлениями;
- системы передачи данных между различными компонентами (ETL).

2. Apache Flink — это мощный фреймворк для потоковой обработки, который поддерживает как потоковую, так и пакетную обработку данных. Flink предлагает различные высокоуровневые API, что делает его удобным для реализации сложных приложений.

Преимущества:

- поддержка сложных потоковых приложений;
- высокая производительность;
- гибкие API для обработки данных.

Области применения Apache Flink:

- реализация аналитических систем, которые требуют обработки событий в реальном времени;
- обработка данных IoT, таких как данные от сенсоров и устройств;
- реализация систем мониторинга для обработки данных о производительности и ошибках.

3. Apache Spark Streaming — позволяет обрабатывать потоковые данные в режиме близком к реальному времени, интегрируясь в экосистему Spark для анализа больших данных.

Преимущества:

- удобство использования с существующими инструментами Apache Spark;
- высокая скорость обработки данных;
- поддержка распределенной обработки.

Области применения Apache Spark Streaming:

- обработка с использованием алгоритмов машинного обучения на потоковых данных;
- веб-аналитика для анализа взаимодействия пользователей с веб-сайтами в режиме реального времени;
- реализация систем мониторинга и предсказательной аналитики.

4. Amazon Kinesis — это облачный сервис от AWS, который обеспечивает потоковую обработку данных из любых источников.

Преимущества:

- полностью управляемый сервис;
- лёгкая интеграция с другими сервисами AWS;
- масштабируемость и надёжность.

Области применения Amazon Kinesis:

- анализ потоковых данных в реальном времени;
- системы безопасности, использующие данные с камер видеонаблюдения;
- обработки запасов и магазинов в реальном времени для улучшения логистики.

В таблице 5 приведены сравнения различных технологий потоковой обработки.

Таблица 5 – Сравнение различных технологий потоковой обработки

Технология	Тип обработки	Скорость	Потоковая или пакетная	Масштабируемость	Использование
Apache Kafka	Потоковая	Высокая	Потоковая	Да	Обмен сообщениями, аналитика
Apache Flink	Потоковая и пакетная	Высокая	Потоковая и пакетная	Да	Реальные приложения с анализом
Apache Spark Streaming	Потоковая	Средняя	Потоковая	Да	Большие данные, ETL
Amazon Kinesis	Потоковая	Высокая	Потоковая	Да	Облачные приложения, мониторинг

На рисунке 1 представлен график производительности инструментов потоковой обработки.



Рисунок 1 – График производительности

Заключение

Потоковая обработка данных представляет собой важную технологию, обеспечивающую анализ информации в реальном времени и позволяющую организациям эффективно реагировать на изменения. Системы потоковой

обработки используются в различных отраслях, включая финансы, интернет вещей и социальные медиа. Они позволяют быстрее обрабатывать данные, повышают точность анализа и снижают затраты на хранение информации.

Рост объёмов данных и потребность в их быстрой обработке подчеркивают необходимость интеграции потоковой обработки в бизнес-процессы.

Её популярность будет продолжать расти по мере увеличения объёмов данных и необходимости их быстрой обработки. Применение потоковой обработки открывает новые горизонты для анализа данных, позволяя организациям разрабатывать инновационные решения и оптимизировать свои процессы.

ЛИТЕРАТУРА:

1. Дьяков А. И. Потоковая обработка данных: теоретические и практические аспекты. - М.: Научный мир, 2020, 320 с.
2. Иванов С. П. Современные системы обработки данных в реальном времени. - СПб.: Питер, 2021, 256 с.
3. Орлов В. А. Технологии потокового анализа: от теории к практике. - Казань: Казанский университет, 2022, 180 с.
4. Сидоров К. Р. Интернет вещей: возможности и вызовы. - М.: ФИЗМАТЛИТ, 2021, 400 с.
5. Федоров А. В. Машинное обучение и анализ данных: основы и применение. - Екатеринбург: УралГУФК, 2019, 450 с.

REFERENCES

1. D'jakov A. I. Potokovaja obrabotka dannyh: teoreticheskie i prakticheskie aspekty. - M.: Nauchnyj mir, 2020, 320 s.
2. Ivanov S. P. Sovremennye sistemy obrabotki dannyh v real'nom vremeni. - SPb.: Piter, 2021, 256 s.
3. Orlov V. A. Tehnologii potokovogo analiza: ot teorii k praktike. - Kazan': Kazanskij universitet, 2022, 180 s.
4. Sidorov K. R. Internet veshhej: vozmozhnosti i vyzovy. - M.: FIZMATLIT, 2021, 400 s.
5. Fedorov A. V. Mashinnoe obuchenie i analiz dannyh: osnovy i primenenie. - Ekaterinburg: UralGUFK, 2019, 450 s.