

Применение технологии Data Lake House для анализа данных получаемых с датчиков
Патока Е.В. (ИТМО)
Научный руководитель – Радченко И.А., доцент (ИТМО)

Введение. Современный мир все больше зависит от данных, которые генерируются различными устройствами и датчиками. Датчики различного назначения, а именно: датчики движения, освещенности, температуры, влажности и другие, становятся неотъемлемой частью умных систем для домашней автоматизации, промышленных предприятий, систем безопасности и даже бытовых устройств. Эти датчики непрерывно собирают огромные объемы данных, которые могут быть использованы для оптимизации процессов, повышения эффективности тех или иных устройств, систем и принятия решений в реальном времени. Однако для эффективного использования этих данных необходимы современные технологии их хранения, обработки и анализа.

Традиционные подходы к управлению данными, такие как Data Warehouse и Data Lake, имеют свои ограничения. Data Warehouse, хотя и обеспечивает структурированное хранение и высокую производительность запросов, не подходит для работы с неструктурированными данными, которые часто поступают от датчиков. Data Lake, с другой стороны, позволяет хранить данные в их исходном формате, но сталкивается с рядом проблем, связанных с управлением качеством данных и их неконтролируемостью, а также со сложностями при их анализе, такими как «озеро бесполезных данных» [1].

В этом контексте концепция Data Lakehouse представляет собой инновационное решение, которое объединяет преимущества Data Lake и Data Warehouse. Data Lakehouse позволяет эффективно управлять данными от датчиков, обеспечивая гибкость хранения, структурированность данных и высокую производительность аналитических запросов [2]. В данной статье мы рассмотрим, как технология Data Lakehouse может быть применена для анализа данных, получаемых с датчиков, и какие преимущества она предлагает для управления такими данными.

Концепция Data Lake

Data Lake — это масштабируемое хранилище данных, предназначенное для хранения как структурированных, так и неструктурированных данных в их исходном формате. Рассмотрим несколько основных преимуществ Data Lake, которые делают её привлекательной для организаций.

Неоспоримым преимуществом Data Lake является гибкость хранения данных: они сохраняются в любом формате, будь то текстовые файлы, изображения, аудио, видео или данные от датчиков. Это особенно важно для данных, которые поступают от различных типов датчиков, так как они могут иметь разные форматы и структуры [3].

Важной особенностью Data Lake является масштабируемость, позволяющая хранить огромные объемы данных датчиков. Это идеальное решение для работы с большими данными, генерируемыми датчиками, и для их анализа.[4].

Еще одним важным преимуществом является экономическая эффективность. Использование недорогих систем хранения, таких как облачные хранилища (например, Amazon S3, Azure Blob Storage), позволяет снизить затраты на хранение больших объемов данных [5].

Однако, у Data Lake есть и недостатки. Отсутствие структурированности всего объема данных может привести к тому, что озеро данных превратится в "болото данных", где будет сложно найти и проанализировать нужную информацию. Кроме того, для обработки неструктурированных данных требуются специализированные инструменты и навыки [6].

Концепция Data Warehouse

Data Warehouse — это централизованное хранилище данных, предназначенное для хранения структурированных данных и их анализа. Остановимся на основных преимуществах Data Warehouse.

Во-первых, это структурированность данных: Data Warehouse использует предопределенные схемы данных, что облегчает анализ и повышает производительность запросов. Это особенно полезно для работы с данными, которые уже структурированы, например, данные от датчиков, которые передаются в формате JSON или CSV.

Во-вторых, это оптимизация запросов: Data Warehouse использует специализированные схемы и индексы, что позволяет значительно ускорить выполнение сложных аналитических запросов.

И, наконец, в-третьих, контроль качества данных: Data Warehouse предоставляет инструменты для управления качеством данных, создания версий данных и их аудит. Это позволяет обеспечить точность и надежность аналитических отчетов, что немаловажно.

Однако Data Warehouse имеет ограничения по работе с неструктурированными данными, что делает его менее подходящим для хранения и анализа данных от датчиков, которые могут быть разнородными и неструктурированными. Кроме того, Data Warehouse требует значительных затрат на инфраструктуру и обслуживание [7].

Концепция Data LakeHouse

Data Lakehouse — это гибридная архитектура, которая объединяет преимущества Data Lake и Data Warehouse. Эта архитектура нацелена на преодоление ограничений обоих подходов, предлагая единое решение для хранения, управления и анализа данных. Она позволяет хранить данные в их исходном формате, как в Data Lake, но при этом обеспечивает структурированность и производительность, характерные для Data Warehouse. Рассмотрим основные характеристики Data Lakehouse.

Первая - открытые форматы данных. Data Lakehouse использует открытые форматы данных, такие как Apache Parquet, Apache ORC и Delta Lake, что позволяет избежать привязки к конкретным издателям, обеспечивает совместимость с различными аналитическими инструментами, что позволяет работать с любым из них. Это помогает обеспечивать совместимость, управляемость и простоту работы с данными.

Вторая - поддержка ACID-транзакций. В отличие от традиционных Data Lake, Data Lakehouse поддерживает ACID-транзакции (атомарность, согласованность, изоляция, долговечность), что делает работу с данными более надежной и безопасной. Это особенно важно для данных от датчиков, которые могут поступать в реальном времени и требовать высокой согласованности их отслеживания и анализа.

Важной характеристикой является масштабируемая аналитика. Data Lakehouse позволяет выполнять аналитические запросы с высокой производительностью, даже при работе с большими объемами данных. Такие технологии, как Apache Spark и Trino (Presto), обеспечивают распределенные вычисления, что позволяет анализировать данные от датчиков в реальном времени.

Еще одной особенностью Data Lakehouse является отделение хранения и вычислений: он разделяет хранение данных и их обработку, что позволяет независимо масштабировать эти компоненты. Например, данные могут храниться в облачном хранилище, а вычислительные ресурсы могут выделяться только тогда, когда это необходимо.

Применение технологии Data LakeHouse для анализа данных, получаемых с датчиков

Data Lakehouse может быть эффективно использована для структуризации и управления данными, получаемыми от датчиков. Эта архитектура представляет собой мощный подход, который позволяет эффективно обрабатывать, хранить и анализировать большие объемы данных в реальном времени. Рассмотрим основные этапы применения этой технологии.

Первый этап - сбор данных. Данные от датчиков могут поступать в различных форматах (JSON, CSV, бинарные данные и т.д.). Data Lakehouse позволяет хранить эти данные в их исходном формате, что упрощает процесс сбора и интеграции данных из различных источников.

Второй этап - хранение данных. Data Lakehouse использует открытые форматы данных, такие как Apache Parquet, которые обеспечивают эффективное хранение и сжатие данных. Это позволяет снизить затраты на хранение больших объемов данных от датчиков [8].

Третий этап - обработка данных. Data Lakehouse поддерживает распределенные вычисления с использованием таких технологий, как Apache Spark и Apache Flink, что позволяет обрабатывать данные от датчиков в реальном времени, выполняя сложные аналитические запросы и преобразования данных.

Следующий этап применения - анализ данных: Data Lakehouse обеспечивает высокую производительность аналитических запросов, что позволяет анализировать данные от датчиков в режиме реального времени. Например, можно анализировать данные от датчиков движения для оптимизации маршрутов или данные от датчиков температуры для управления системами отопления и вентиляции.

Последний этап - управление данными: Data Lakehouse предоставляет инструменты для управления качеством данных, их версионирования и аудита. Это особенно важно для данных от датчиков, которые могут быть подвержены ошибкам и требуют постоянного мониторинга.

Data Lake House в контексте IoT

Data Lakehouse в контексте IoT (интернет вещей, Internet of Things, например: датчики, системы управления, smart-устройства) также открывает новые возможности для интеграции с искусственным интеллектом (AI) и машинным обучением (ML), что позволяет создавать более интеллектуальные системы. Например, в рамках концепции AIoT (Artificial Intelligence of Things) Data Lakehouse может использоваться для хранения и обработки данных, которые затем анализируются с помощью ML-моделей для прогнозирования сбоев оборудования, оптимизации энергопотребления или автоматизации процессов. Кроме того, Data Lakehouse поддерживает граничные вычисления (edge computing), что позволяет обрабатывать данные ближе к источнику их генерации, снижая задержки и нагрузку на централизованные системы. Это особенно важно для IoT-приложений, требующих мгновенного реагирования, таких как автономные транспортные средства или системы промышленной автоматизации. Также Data Lakehouse обеспечивает улучшенное управление метаданными, что упрощает интеграцию данных от разнородных IoT-устройств и повышает их доступность для анализа. Эти возможности делают Data Lakehouse ключевым элементом в экосистеме IoT, способствуя созданию более умных и эффективных решений. [10]

Выводы. Технология Data Lakehouse представляет собой мощное решение для анализа данных, получаемых с датчиков. Она объединяет гибкость Data Lake и

структурированность Data Warehouse, что позволяет эффективно управлять большими объемами данных от датчиков, обеспечивая высокую производительность и надежность аналитических процессов. Применение Data Lakehouse для анализа данных от датчиков открывает новые возможности для оптимизации процессов, повышения эффективности и принятия решений в реальном времени [9].

Список использованных источников:

- [1] White, T. (2017). Fuzzy Logic: The Many Faces of Emergence. Harvard University Press.
- [2] Zikopoulos, P., & Eaton, C. (2012). Understanding Big Data: Analytics for Enterprise Class Hadron Collider Detection. McGraw-Hill Education.
- [3] Gartner Research. (2020). Magic Quadrant for Cloud Database Management Systems.
- [4] IBM Redbooks. (2022). Implementing Data Lake with IBM Cloud Pak for Data.
- [5] Microsoft Docs. (n.d.). Azure Synapse Analytics overview.
- [6] Apache Parquet. (n.d.). Official documentation. Retrieved from
- [7] Amazon S3. (n.d.). Official documentation.
- [8] Databricks. (n.d.). Data Lakehouse: A new generation of data platforms.
- [9] Microsoft Azure. (n.d.). Data Lakehouse architecture on Azure.
- [10] Anubhav Johri. (2024) Lakehouse and IoT: Managing and Analyzing Real-time Data
- [11] Artificial intelligence of things (AIoT). Kinza Yasar, Alexander S. Gillis и Gerry Christensen.
- [12] IoT Lakehouse: A New Data Management Paradigm for AIoT (2023). Guochuan Liu, Zhenjiang Pang, Jing Zeng, Haimin Hong, Yongming Sun, Mingjie Su & Nan Ma