Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _____

КАФЕДРА _____

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Методы работы с озёрами данных

Студент ИУ5-24М _____ Н.В. Журавлев
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ М.В. Виноградова
(Подпись, дата) (И.О.Фамилия)

Консультант _____ _____
(Подпись, дата) (И.О.Фамилия)

2024 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой ИУ5
(Индекс)
_____ В.И. Терехов
(И.О.Фамилия)
« 15 » февраля 20 24 г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме Методы работы с озёрами данных

Студент группы ИУ5-24М

Журавлев Николай Вадимович
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

Учебная

Источник тематики (кафедра, предприятие, НИР) НИР

График выполнения НИР: 25% к ___ нед., 50% к ___ нед., 75% к __ нед., 100% к ___ нед.

Техническое задание Произвести изучение озёр данных и произвести обзор существующих платформ для создания озёр данных

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на 12 листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

Дата выдачи задания « 15 » февраля 2024 г.

Руководитель НИР _____ М.В. Виноградова
(Подпись, дата) (И.О.Фамилия)

Студент _____ Н.В. Журавлев
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Оглавление

Введение

Озеро данных (Data lake)— это система или хранилище данных, которые хранятся в необработанном формате. Data lake обычно представляет собой единое хранилище данных, включающее необработанные копии данных исходной системы. Data lake может включать структурированные данные из реляционных баз данных (строки и столбцы), полуструктурированные данные (CSV, журналы, XML, JSON), неструктурированные данные (электронные письма, документы, PDF-файлы) и двоичные данные (изображения, аудио, видео) [1].

Озёра данных предназначены для того, чтобы собирать, хранить и обрабатывать большое количество информации, поступающей практически непрерывным потоком. Такую информацию называют Big Data, или большими данными. Data Lake полезны всем компаниям, которые планируют анализировать большие данные любой области. Само по себе озеро данных бесполезно, потому что это просто хранилище. Чтобы с ним работать, нужны инструменты для очистки, структурирования, извлечения и анализа данных, и специалисты для работы с этими инструментами [2].

Часто Data lake используют для хранения важной информации, которая пока не используется в аналитике. Или даже для данных, которые кажутся бесполезными, но, вероятно, пригодятся компании в будущем. Data lake позволяет накапливать данные «про запас», а не под конкретный запрос бизнеса. За счет того, что данные всегда «под рукой», компания может быстро проверить любую гипотезу или использовать данные для своих целей [3].

Основная идея построения

Основная идея Data lake заключается в следующем: все данные, отправляемые организацией, будут храниться в единой структуре данных, называемой Data Lake. Данные будут храниться в озере в исходном формате. Будет исключена сложная предварительная обработка и преобразование

данных при загрузке в Data lake. Как только данные помещены в озеро, они доступны для анализа всем сотрудникам организации.

Озеро представляет собой файловое хранилище на нескольких серверах, в котором лежат данные. Как правило данные распределены между серверами, чтобы хранилище можно было быстро масштабировать — подключить новые серверы для расширения места.

К серверам настраивают подключение разных источников данных, доступных компании. Каналы поставки данных называют пайплайнами, а всю схему подключения — ETL-процессом. Обычно всё настроено так, чтобы данные загружались автоматически.

Хотя Data lake и неструктурированное, порядок в нём всё-таки должен быть, иначе спустя время накопится огромное количество данных, в которых невозможно будет разобраться. Поэтому перед добавлением в озеро данные размечают и запоминают, откуда и в каком формате они поступили. В итоге внутри озера данных хранятся не только сами объекты, но и метаданные, то есть информация об объектах. Это облегчает поиск, извлечение и анализ данных в будущем.

В архитектуре озера данных должны быть предусмотрены инструменты резервного копирования, чтобы информация не терялась.

Общий алгоритм работы выглядит следующим образом:

1. В одном из источников формируются данные.
2. По заранее настроенному маршруту данные с серверов отправляются в Data lake.
3. При поступлении данные размечаются: записывается их источник, время поступления, формат и структура.
4. Данные помещаются в озеро и хранятся там. Как правило, срок хранения не ограничен, хотя иногда данные удаляют по мере устаревания или использования в аналитике.

При необходимости данные извлекают из хранилища по определённым критериям и используются [3, 4].

Недостатки Data lake

Озера данных оптимизированы для высокой пропускной способности, но ради этого приходится жертвовать качеством данных [5]:

1. В Data lake не требуется структурировать данные, поэтому их сложнее анализировать.
2. Data lake не имеет инструментов для целостного получения всех данных.
3. Без квалифицированного контроля за озерами данных трудно гарантировать конфиденциальность и безопасность хранилища.
4. Если управление озером организовано плохо, в нем быстро накапливаются большие объемы неконтролируемых, и, возможно, бесполезных данных. Для эффективной фильтрации данных и отсечения недостоверных источников требуется высокая квалификация.

Если в Data lake хранится слишком много данных, которые плохо организованы, без надлежащего управления метаданными и надежного управления данными, найти соответствующие данные становится все труднее. Через определенное время данные теряют свою актуальность и, если данные все еще остаются в хранилище данных, в течение длительных периодов времени накапливается все больше и больше неактуальных данных. Неправильные временные метки набора данных также приводят к тому, что информацию невозможно найти или оценить. И в таком случае образуется то, что называется болотом данных (data swamp).

Существуют типичные характеристики болота данных, на наличие которых вы можете проверить свое озеро данных и затем от них избавиться:

1. Большие данные без какой-либо организации и документации, например, через каталог данных или концепцию ролей.

2. Отсутствует метаинформация структурированных или неструктурированных данных.
3. Устаревшие и неверные данные.
4. Нет директора по данным или владельца продукта, который управляет платформой.
5. Отсутствующие или нарушенные связи между информацией.

Для очистки данных при замусоривании данных могут оказаться полезны такие роли, как владелец продукта или директор по цифровым технологиям, которые организуют и развивают Data lake. Кроме того, необходимо создать каталог данных, который обеспечит ясность данных. Вместе с концепцией ролей это гарантирует, что данные дойдут до нужных людей. Неверные и старые данные должны быть удалены или заархивированы, поскольку это в любом случае часто требуется нормативными актами и может также привести к снижению затрат. Требованиями к записи данных являются, например, маркировка источника данных, маркировка метаданных и содержательная номенклатура [6].

Обзор платформ для создания и управления data lake

Google cloud

GCP предлагает набор услуг автоматического масштабирования, которые позволяют создать озеро данных, которое интегрируется с существующими приложениями. К ним относятся Dataflow и Cloud Data Fusion для поглощения данных, Cloud Storage для хранения, а также Dataproc и BigQuery для обработки данных и аналитики. Google Cloud предоставляет инструменты и рабочие процессы для управления озерами данных на протяжении всего их жизненного цикла. Google структурирует свои услуги озера данных по четырем ключевым этапам жизненного цикла озера данных [7]:

1. Приём — позволяет данным из многочисленных источников, таких как потоки данных о событиях, журналы и устройства IoT, хранилища

исторических данных, данные из транзакционных приложений, поступать в озеро данных.

2. Хранение — хранение данных в надежном и легкодоступном формате.
3. Обработка — преобразование данных из исходного формата в формат, позволяющий использовать и анализировать.
4. Исследование и визуализация — анализ данных и представление их в виде визуализаций или отчетов, предоставляющих ценную информацию бизнес-пользователям.

Так же на этой платформе имеется возможность интеграции уже существующих Data lake из некоторых других платформ.

Hadoop Azure Data Lake

Является платформой, в которой и создавалась концепция Data lake [8]. В озерах данных данные чаще всего хранятся в распределенной файловой системе Hadoop (HDFS). Эта система позволяет осуществлять одновременную обработку данных. Это связано с тем, что при приеме данные разбиваются на сегменты и распределяются по разным узлам кластера.

HDFS обладает рядом отличительных свойств [9]:

1. Большой размер блока по сравнению с другими файловыми системами (>64MB), поскольку HDFS предназначена для хранения большого количества огромных (>10GB) файлов;
2. Ориентация на недорогие и, поэтому не самые надежные сервера – отказоустойчивость всего кластера обеспечивается за счет репликации данных;
3. Зеркалирование и репликация осуществляются на уровне кластера, а не на уровне узлов данных;

4. Репликация происходит в асинхронном режиме – информация распределяется по нескольким серверам прямо во время загрузки, поэтому выход из строя отдельных узлов данных не повлечет за собой полную пропажу данных;
5. HDFS оптимизирована для потоковых считываний файлов, поэтому применять ее для нерегулярных и произвольных считываний нецелесообразно;
6. Клиенты могут считывать и писать файлы HDFS напрямую через программный интерфейс Java;
7. Файлы пишутся однократно, что исключает внесение в них любых произвольных изменений;
8. Принцип WORM (Write-once and read-many, один раз записать – много раз прочитать) полностью освобождает систему от блокировок типа «запись-чтение». Запись в файл в одно время доступен только одному процессу, что исключает конфликты множественной записи.
9. HDFS оптимизирована под потоковую передачу данных;
10. Сжатие данных и рациональное использование дискового пространства позволило снизить нагрузку на каналы передачи данных, которые чаще всего являются узким местом в распределенных средах;
11. Самодиагностика — каждый узел данных через определенные интервалы времени отправляет диагностические сообщения узлу имен, который записывает логи операций над файлами в специальный журнал;
12. Все метаданные сервера имен хранятся в оперативной памяти.

AWS EMR

AWS EMR - это сервис, предоставляемый Amazon Web Services (AWS), который позволяет организациям хранить и анализировать большие объемы данных в облаке. Он объединяет в себе возможности сервиса EMR (Elastic MapReduce) для обработки и анализа больших данных с функциональностью Data Lake, обеспечивая пользователям удобный и масштабируемый способ работы с данными.

Особенность AWS EMR, что она использует все продукту от AWS. Для обработки используется AWS Lake Formation, его особенности [10]:

1. Импортирование данных из существующих баз данных. Данные сканируются, когда пользователь предоставляет AWS Lake Formation местоположение текущих баз данных и свои данные для входа.
2. Организация и маркировка данных. Lake Formation предлагает коллекцию технических метаданных, извлеченных из источников данных, для потребителей, которые ищут наборы данных.
3. Преобразование данных. Такие преобразования, как перезапись форматов дат для обеспечения единообразия, возможны с помощью Lake Formation. Amazon Data Lake Formation создает шаблоны преобразований и организует процессы, которые будут их выполнять.
4. Принудительное шифрование. Пользовательское Data lake зашифровано с помощью шифрования Amazon S3 через Lake Formation. Чтобы предотвратить удаление вредоносных данных при передаче, можно использовать отдельные учетные записи для исходного и целевого регионов при использовании S3.
5. Управление контролем доступа. Lake Formation позволяет управлять разрешениями на доступ к данным в Data lake из одного места. Доступ к данным можно ограничить на уровне базы данных, таблицы, столбца, строки и ячейки с помощью правил безопасности. Эти

политики применяются к пользователям и ролям, а также к пользователям и группам, объединенным через внешнего поставщика удостоверений.

6. Настройте ведение журнала аудита. Мониторинг доступа к данным на платформах аналитики и машинного обучения.
7. Метатеги данных для бизнеса. В Data Lake на Amazon можно определить соответствующие варианты использования и уровни конфиденциальности данных, используя безопасность формирования и ограничения доступа.
8. Поиск данные для анализа. Пользователи Lake Formation имеют доступ к текстовому поиску, выполняемому онлайн, для поиска и фильтрации наборов данных, хранящихся в общей библиотеке данных.

Microsoft Azure Data Lake

В Azure Data Lake представлены все возможности, упрощающие хранение данных любых объема, формата и скорости передачи, а также выполнение любых видов обработки и анализа на разных платформах и языках для разработчиков, специалистов по обработке и анализу данных и аналитиков. Azure Data Lake упрощает получение и хранение данных, одновременно ускоряя работу пакетной, потоковой и интерактивной аналитики.

Особенность заключается в заранее собранном наборе инструментов, таких как [11]:

1. Azure HDInsight — это управляемая комплексная облачная служба аналитики с открытым кодом, предназначенная для предприятий. С помощью HDInsight можно использовать платформы с открытым кодом, такие как Apache Spark, Apache Hive, LLAP, Apache Kafka, Hadoop и т. д. в среде Azure. Azure HDInsight можно применять в различных сценариях обработки больших данных. Это могут быть

исторические данные (данные, которые уже собираются и хранятся) или данные в режиме реального времени (данные, которые передаются непосредственно из источника).

2. Data Lake Store — это высокомасштабируемое облачное озеро данных, предназначенное для предприятий, создано в соответствии с открытыми стандартами HDFS. В нём отсутствии ограничений на размер данных, есть возможностью выполнять огромное количество параллельных аналитических задач и имеется единая платформа хранения данных. Так же имеется проверка подлинности данных с помощью Microsoft Entra ID и управления доступом на основе ролей.
3. Data Lake Analytics — служба заданий аналитики. Это облачная служба аналитики, в которой можно с легкостью разрабатывать и выполнять программы обработки и программы массовых параллельных операций преобразования данных на U-SQL, R, Python и .NET. В ней нет инфраструктуры, так как нет серверов, виртуальных машин или кластеров, которые нужно ждать, настраивать и которыми нужно управлять. Можно масштабировать вычислительную мощность, измеряемую в единицах Azure Data Lake Analytics (AU). Имеются специальные библиотеки на языке .NET, R и Python с помощью, которых обрабатываются данные.

Выводы

В данной работе было проведено изучение озёр данных, их строение и недостатки. Так же был выполнен поиск и проведён обзор найденных платформ, на базе которых можно организовать озёра данных, после чего для каждой было представлено краткое их описание и выделены их ключевые особенности.

Литература

1. [Электронный ресурс] – 2024 г. – Режим доступа: https://en.wikipedia.org/wiki/Data_lake, свободный.
2. [Электронный ресурс] – 2024 г. – Режим доступа: https://practicum.yandex.ru/blog/chto-takoe-ozera-dannyh, свободный.
3. [Электронный ресурс] – 2024 г. – Режим доступа: https://cloud.vk.com/blog/chto-takoe-ozera-dannyh-i-zachem-tam-hranyat-big-data, свободный.
4. Pwint Phyu Khine, Zhao Shun Wang Data Lake: A New Ideology in Big Data Era. – Wuhan, China, 2017
5. [Электронный ресурс] – 2024 г. – Режим доступа: https://yandex.cloud/ru/docs/glossary/datalake, свободный.
6. [Электронный ресурс] – 2024 г. – Режим доступа: https://medium.com/codex/what-is-a-data-swamp-38b1aed54dc6, свободный.
7. [Электронный ресурс] – 2024 г. – Режим доступа: https://bluexp.netapp.com/blog/gcp-cvo-blg-google-cloud-data-lake-4-phases-of-the-data-lake-lifecycle, свободный.
8. [Электронный ресурс] – 2024 г. – Режим доступа: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/, свободный.
9. [Электронный ресурс] – 2024 г. – Режим доступа: https://bigdataschool.ru/wiki/hdfs, свободный.
10. [Электронный ресурс] – 2024 г. – Режим доступа: https://k21academy.com/amazon-web-services/aws-data/aws-lake-formation/, свободный.
11. [Электронный ресурс] – 2024 г. – Режим доступа: https://azure.microsoft.com/ru-ru/solutions/data-lake, свободный.