НЕГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ ЧАСТНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ФИНАНСОВО-ПРОМЫШЛЕННЫЙ УНИВЕРСИТЕТ
«СИНЕРГИЯ»

РЕФЕРАТ

На тему «Источники данных и хранение информации на предприятии»

По дисциплине: «Информационно-аналитические системы»

Обучающийся ФИО

Москва 2023 г.

Оглавление:

Введение С.3

Раздел I. Современные подходы к хранению и обработке электронных данных. С.4-6

Раздел II. Современные средства хранения и обработки электронных данных на предприятиях. С.7-12

Заключение С.13

Список использованных источников и литературы С.14

Введение

Ежедневный растущий объём электронных данных ставит сложные задачи перед традиционными способами по организации хранения, обработки и анализа данных. Целесообразность раскрытия данной темы подтверждается наличием высокого спроса на услуги хранения данных и аналитической обработки данных.

В работе использованы различные источники информации, включая научно-методический журнал «Проблемы современной науки и образования»[1] и электронный ресурс «Информационные системы в экономике: практикум»[2].

Цель работы: сформировать представление об основных источниках данных на предприятии и принципах хранения данных на предприятии.

Задачи:

1. Рассказать о современных подходах к хранению и обработке электронных данных.
2. Рассказать о современных средствах хранения и обработки электронных данных на предприятиях.

Реферат состоит из введения, 2 разделов, заключения, списка использованных источников и литературы.

---

[1] Научно-методический журнал «Проблемы современной науки и образования» URL: https://ipi1.ru/ (дата обращения: 18.06.2023)

[2] «Информационные системы в экономике: практикум»: [электронный ресурс]. URL: https://cchgeu.ru/upload/iblock/942/lv0q4bx2r9e2gssw2uwlp9n27982kk1s/Infomatsionnye-sistemy-v-ekonomike-Praktikum.pdf

Раздел I. Современные подходы к хранению и обработке электронных данных.

На сегодняшний день нелегко измерить общий объем электронных данных, хранящихся во всем мире, однако по оценкам IDC размер «цифрового мира» в 2006 г. составлял около 0.18 зеттабайта, а через 5 лет к 2011 году должен был достигнуть около 1.8 зеттабайта, тем самым продемонстрировав десятикратный рост. Согласно данным IDC объем данных к 2020 году должен был достигнуть отметки в 44 зеттабайта.

Источниками таких объемов данных являются такие как:

- Главная фондовая биржа США, генерирует 1 терабайт данных в день.
- Хранилище данных социальной сети Facebook ежедневно увеличивает объем данных на 500 терабайт.
- Internet Archive Stores хранящая данные интернет-сайтов по состоянию на октябрь 2012 уже хранит 10 петабайт данных и ежемесячно прирастает 20 терабайтами в месяц.
- Большой адронный коллайдер, расположенный около Женевы, генерирует около 15 петабайт в год.

Ежедневно растущий объем электронных данных ставит перед нами задачу по организации в хранении, обработке и анализе данных.

Большой объем данных, а также информации хранится в специализированных реляционных базах данных, которые называют хранилищами данных (ХД либо Data Warehouse).

Хранилища данных в отличие от оперативных баз данных OLTP (On-Line Transaction Processing), работающих с приложениями, имеют некоторые функциональные ограничения, что позволяет уменьшить время выполнения запросов. Отличия ХД от обычной базы данных:

Обычные базы данных (БД) предназначены для помощи в выполнении повседневной работе, а ХД для принятия решений;

Обычные БД подвержены постоянному изменению данных, ХД в свою очередь выполняют обновление базы согласно предписанному времени без изменения предыдущих данных;

Обычные БД чаще всего являются источником ХД, а ХД могут также пополняться из других внешних источников;

Зачастую ХД имеет ненормализованную структуру, что позволяет заметно увеличить скорость выполнения запросов.

Ральф Кимбалл, один из авторов концепции хранилищ данных, сформулировал основные требования к хранилищам данных:

- Поддержка высокой скорости получения данных из ХД;
- Поддержка внутренней непротиворечивости данных;
- Возможность получения и сравнения так называемых срезов данных (slice and dice);
- Наличие удобных утилит просмотра данных в ХД;
- Полнота и достоверность хранимых данных;
- Поддержка качественного процесса пополнения данных.

Одним из основных принципов построения ХД является использование единой структуры метаданных: системные таблицы хранилища данных имеют жестко заданную структуру, а содержащаяся в них информация четко описывает модель данных ХД, в соответствии с которой загружаются и обрабатываются классификаторы и данные. Таким образом, это позволяет начать построение универсальных программных компонентов, взаимодействующих с ХД.

На сегодняшний день не все инструменты способны справиться с большими объемами данных. Hadoop является набором инструментов позволяющих работать с большими данными. Средняя производительность жестких дисков около 100 МБ/с, то есть для обработки 1 ТБ данных потребуется примерно 2.5 часа времени. Параллельная обработка данных с нескольких дисков позволяет улучшить показатели в несколько раз. Например, на обработку 1 ТБ данных с дисков потребуется 2 минуты.

Распределенная файловая система HDFS отвечает за организацию и хранение данных в Hadoop кластерах.

Принципы проектирование в Hadoop:

Так как сбои в аппаратной системе неизбежны. HDFS реализует надежные алгоритмы репликации данных, а метаданные файловой системы используют журнал, позволяющий восстановить требуемое состояние.

Система HDFS построена таким образом, что позволяет обработку больших объемов данных с наиболее максимальной производительностью благодаря поточной обработке данных. Система оптимизирована для работы с большим объемом данных.

Вычисления происходят намного эффективнее благодаря программному интерфейсу, который предоставляет HDFS. В Hadoop все вычисления разбиваются на несколько подмножеств, каждое из которых обрабатывается на отдельном узле кластера. Представляется это в виде последовательности map задач и reduce задач. Каждый узел в map задачах получает на вход множество пар.

Вычисления в Hadoop представляются в виде последовательности map и reduce задач. В начале вычислений входное множество данных разбивается на несколько подмножеств. Каждое подмножество обрабатывается на отдельном узле кластера. Map задача на каждом узле получает на вход множество пар ключ-значений и возвращает другое множество. По ключу все пары сортируются, группируются и передаются на вход reduce, которая в свою очередь формирует итоговый результат.

Эффективность использования Hadoop можно заметить в одном из интересных примеров тестирования скорости сортировки данных. Рекордные показатели в 2008 году предоставила компания Google, 1TB данных в Hadoop кластере компании Google удалось отсортировать за 68 с. В 2009 году в отчете компании Yahoo утверждалось, что им удалось это сделать за рекордные 62 с.

Раздел II. Современные средства хранения и обработки электронных данных на предприятиях.

В процессе деятельности предприятия накапливается большое количество информации. Эта информация может быть количественной, т. е. иметь конкретное численное выражение, и качественной, определяющей мнения консультантов, суждения специалистов, экспертные оценки. В свою очередь, количественная информация подразделяется на учетную и неучетную. Источниками учетной информации на предприятии выступают:

- Бухгалтерский учет и отчетность;
- Статистический учет и отчетность;
- Оперативный учет и отчетность;
- Выборочные учетные данные.

К источникам неучетной информации можно отнести:

Результаты аудиторских проверок (внешних и внутренних), различных ревизий (внутриведомственных и вневедомственных);

- Результаты проверок налоговой службы;
- Материалы производственных совещаний;
- Протоколы собраний трудовых коллективов;
- Материалы средств массовой информации;
- Докладные и объяснительные записки сотрудников;
- Переписка с вышестоящими организациями;
- Результаты взаимоотношений с финансовыми и кредитными
- организациями;
- Материалы, получаемые в процессе взаимодействия с
- исполнителями.

Кроме этого, может использоваться различный нормативный материал: ГОСТы, внутренние стандарты, справочники, прейскуранты и т. п.

Вся эта информация должна храниться на предприятии и быть в любой момент доступна для пользователя. Для хранения информации могут быть

использованы различные средства: файловые системы, оперативные базы данных (OLTP) и хранилища данных (DWH).

В современных условиях большинство рабочих мест сотрудников оснащено персональными компьютерами (АРМ – автоматизированное рабочее место). В процессе работы на каждом АРМе накапливается оперативная информация, документы, сопровождающие те или иные бизнес-процессы. Эта информация хранится на компьютере в виде файлов.

По определению файл – это именованная область внешней памяти, в которую можно записывать и из которой можно считывать данные.

Файлы бывают разных типов: обычные файлы, специальные файлы, файлы-каталоги.

Обычные файлы – это файлы различного формата, такие как офисные документы, отсканированные бумажные документы, Webстраницы, графические изображения, чертежи, видеофайлы, которые можно отобразить на экране и распечатать на принтере.

Специальные файлы – это файлы, которые позволяют пользователю выполнять операции ввода-вывода, используя обычные команды записи в файл или чтения из файла.

Каталог – это группа файлов, объединенных пользователем по определенному признаку. В каталоге содержится список файлов, входящих в него, и устанавливается соответствие между файлами и их характеристиками (атрибутами). В качестве атрибутов файлов могут быть использованы разные характеристики, например:

- Владелец или создатель файла;
- Информация о доступе к файлу;
- Пароль для доступа;
- Различные признаки, например: «только для чтения», «системный файл», «архивный файл» и т. п.;
- Время создания или последнего изменения файла;
- Размер файла и т. д.

Для организации хранения и управления файлами на компьютере используется файловая система, представляющая собой функциональную часть операционной системы. Файловая система должна обеспечивать пользователю:

- Контролируемый доступ к файлам;
- Возможность осуществлять различные операции с файлами: создавать, удалять, копировать, изменять;
- Возможность обмена данными между файлами;
- Возможность восстанавливать свои файлы в случае их повреждения. Файловые системы предназначены для обслуживания многих тысяч файлов и обеспечивают хранение слабо структурированной информации.

Оперативные базы данных (OLTP – Online Transaction Processing – обработка транзакций в реальном времени). Оперативные базы данных используются предприятиями для поддержания их повседневной деятельности, для отслеживания информации, с которой они имеют дело в процессе решения оперативных задач. Это может быть информация о произведенных товарах, принятых заказах, оказанных услугах, выплатах, доходах и т. п.

Результатом фиксации указанной информации становятся одна или несколько записей в оперативной базе данных. Сам процесс фиксации называют бизнес-транзакцией, а информацию – данными транзакции. По определению транзакция – это последовательность операторов манипулирования данными, выполняющаяся как единое целое и переводящая базу данных из одного целостного состояния в другое целостное состояние.

Системы оперативной обработки транзакций служат для хранения данных о выполняемых бизнес-транзакциях. Основная функция подобных систем заключается в одновременном выполнении большого количества коротких транзакций от большого числа пользователей. Примером

транзакции может быть следующее действие: «перечислить определенную сумму денег со счета А на счет В».

OLTP-системы призваны сохранять данные бизнес-транзакций по мере их поступления. Они обычно имеют дело с текущими значениями каких-либо параметров. Например, типичное банковское OLTP-приложение имеет дело с текущими остатками денег на клиентском счете.

- OLTP-системы характеризуются:
- Поддержкой большого числа пользователей;
- Короткими транзакциями;
- Относительно короткими запросами;
- Малым временем отклика на запрос.

Для поддержания различных аспектов своей каждодневной деятельности организации обычно используют разные OLTPсистемы. Одна система предназначена для обработки заказов, другая – для ведения бухучета, третья – для обслуживания потребностей производства, четвертая – для управления персоналом. К числу транзакционных систем относятся ERP-системы, автоматизированные банковские системы (АБС), биллинговые системы, учетные системы и некоторые другие.

Данные в OLTP-системы поступают в основном из внутренних источников, причем это текущие данные за период от нескольких месяцев до одного года. Объемы хранимых данных могут составлять сотни мегабайт, гигабайты. Частота обновления данных высокая, обновления происходят маленькими порциями. Основное их назначение – фиксация данных, оперативный поиск и преобразование данных. В основе таких систем лежат оперативные базы данных.

По определению Уильяма Инмона, основоположника хранилищ данных, «Хранилище данных – это предметно-ориентированное, привязанное ко времени и неизменяемое собрание данных для под держки процесса принятия управляющих решений». Задача хранилища – предоставить лицу,

принимающему решения, информацию для анализа в одном месте и в простой, понятной для восприятия структуре.

Данные в хранилище попадают из оперативных систем (OLTPсистем), которые предназначены для автоматизации бизнеспроцессов, и других внутренних источников информации. Хранилище также может пополняться за счет внешних источников информации, например статистических отчетов и т.п.

Концепция хранилищ данных – это концепция подготовки данных для анализа. Она предполагает реализацию единого интегрированного источника данных. Объединяя в себе всю информацию, необходимую для эффективного функционирования организации, хранилища являются основой единого информационного пространства организации, используемого для проведения аналитической работы и подготовки данных для принятия решений.

Единое информационное хранилище обеспечивает:

- Объединение данных и приведение их к единой структуре;
- Повышение производительности получения данных;
- Проведение эффективного анализа данных.

Хранилищу данных свойственна малая частота изменений, изменения производятся большими порциями и обычно по расписанию. Хранилище объединяет внутренние и внешние данные, в составе этих данных – текущие данные и исторические за период до нескольких десятков лет. Объемы хранимых данных – гигабайты и терабайты. Основное назначение хранилищ данных – это хранение детализированных и агрегированных исторических данных, аналитическая обработка, прогнозирование и моделирование.

Можно выделить два типа хранилищ данных: корпоративные хранилища данных (enterprise data warehouses) и витрины, или киоски, данных (data marts).

Корпоративные хранилища данных содержат информацию, относящуюся к деятельности всей корпорации и собранную из множества оперативных источников данных. Такие хранилища обычно консолидируют

информацию по всем аспектам деятельности корпорации и используются для принятия как стратегических, так и тактических решений.

Корпоративное хранилище содержит обобщающую и детальную информацию; его объем может достигать от десятков гигабайт до одного или нескольких терабайт.

Витрины данных (небольшие хранилища данных) содержат подмножество корпоративных данных и создаются для определенной группы пользователей, отделов или подразделений внутри организации. Они охватывают конкретный аспект, интересующий сотрудников данного отдела. Витрина данных может получать данные из корпоративного хранилища (зависимая), или данные могут поступать непосредственно из оперативных источников (независимая витрина).

Заключение

Увеличение общего объёма электронных данных, хранящихся во всём мире, ставит перед нами задачу в увеличении эффективности действий, направленных на организацию хранения, обработки и анализа данных. Для этого создаются концепции хранилищ данных, инструменты обработки больших данных и принципы работы с этими инструментами. Эффективность предпринимаемых действий возможно оценить с помощью примеров тестирования сортировки данных.

Список использованных источников и литературы

1. Информационно аналитические системы. основы проектирования и применения // URL: https://ipi1.ru/images/PDF/2017/87/PMSE-5-87.pdf#page=26&zoom=100,90,160 (дата обращения: 18.06.2023).
2. Информационные системы в экономике: практикум // Воронежский государственный технический университет URL: https://cchgeu.ru/upload/iblock/942/lv0q4bx2r9e2gssw2uwlp9n27982kk1s/Infomatsionnye-sistemy-v-ekonomike-Praktikum.pdf (дата обращения: 18.06.2023).
3. Кэмпбелл Лейн, Черити Мейджорс Базы данных. Инжиниринг надежности. - СПб.: Питер, 2020. - 304 с.
4. Научно-методический журнал «Проблемы современной науки и образования» URL: https://ipi1.ru/ (дата обращения: 18.06.2023).
5. Проектирование и реализация хранилища данных для анализа бизнес деятельности компании // Репозиторий тольяттинского государственного университета URL: https://dspace.tltsu.ru/bitstream/123456789/4135/1/Павлов%20В.В._ПИбд-1202а.pdf (дата обращения: 18.06.2023).