УДК 681.3.019:378.1

НЕКОТОРЫЕ ПРИЧИНЫ ИСПОЛЬЗОВАНИЯ ХРАНИЛИЩ ДАННЫХ

Эсенбеков А., Турдаалиев Б.Р., Тороев А.А.
КГУСТА им. Н. Исанова

В статье рассмотрены недостатки классических баз данных и СУБД, а также принципы их устранения по технологии построения Хранилищ данных. Показано, что технология построения Хранилищ Данных является универсальным механизмом предоставления доступа к любому типу информации, включая реляционные и нереляционные данные.

Ключевые слова: Базы данных, реляция, хранилище, SQL-запросы, технология.

МААЛЫМАТ САКТООЧУ КОЛДОНУУНУН АЙРЫМ СЕБЕПТЕРИ

Эсенбеков А., Турдаалиев Б.Р., Төрөев А.А.
Н.Исанов ат. КМКТАУ

Макалада классикалык маалымат базаларынын жана МБББнын кемчиликтери, ошондой эле маалымат кампаларын куруу технологиясын колдонуу менен аларды жоюу принциптери талкууланат. Маалыматтар кампаларын куруу технологиясы ар кандай түрдөгү маалыматтын, анын ичинде реляциялык жана реляциялык эмес маалыматтардын жеткиликтүүлүгүн камсыздоонун универсалдуу механизми экени көрсөтүлгөн.

Баштапкы сөздөр: Берилиштер базасы, реляциялык, сактоо, SQL сурамдары, технология.

SOME REASONS TO USE DATA STORAGE

Esenbekov A., Turdaaliev B.R., Toroev A.A.
KSUCTA named of N. Isanov

The article discusses the shortcomings of classical databases and DBMS, as well as the principles of their elimination using the technology of building data warehouses. It is shown that the technology of building Data Warehouses is a

universal mechanism for providing access to any type of information, including relational and non-relational data.
Key words: Databases, relational, storage, SQL queries, technology.

Бурный рост информатизации всех сфер деятельности человечества привело к необходимости хранения и обработки больших объемов данных. Эксперты IDC (International Data Corporation) считают, что сегодняшних хранилищ хватит лишь для 15 % данных /1/. Одним из основных факторов этого роста является увеличение доли автоматически генерируемых данных. Большие объемы полезных данных создаются с систем видеонаблюдения, встроенных в оборудование, медицинских систем, информации с компьютеров, смартфонов, бытовой электроники. По оценкам IDC, количество устройств в мире, которые можно подключить к Интернету, приближается к 200 млрд, из которых 14 млрд, или 7 %, уже подключены и активно передают данные.

По прогнозам /1/, инвестиции в IT-инфраструктуру цифровой вселенной (оборудование, телекоммуникации, хранение и управление информацией и персонал) вырастут на 40 %. Причем инвестиции в хранение и защиту информации, обработку «больших данных» (Big Data) и облачные технологии будут расти значительно быстрее. Большие данные диктуют новые взаимосвязанные принципы обработки информации /2/.

Первый — это способность анализировать все данные, а не довольствоваться их частью или статистическими выборками.

Второй — готовность иметь дело с неупорядоченными данными в ущерб точности.

Третий — изменение образа мыслей: доверять корреляциям, а не гнаться за труднодостижимым поиском причинно-следственных зависимостей.

Существенно и то, что на сегодняшний день используется менее 3 % из 23 % потенциально полезных данных, которые могли бы найти применение с технологиями Big Data.

Беспрецедентный рост информации в мире, необходимость хранения и обработки всей массы накопленных данных требует создания хранилищ, построенных на новых технических средствах, использующих новые модели и методы эффективной обработки данных. Традиционные системы управления базами данных (СУБД) предназначались для создания и использования информационных моделей — корпоративных баз данных (БД) в конкретных сферах деятельности. Корпоративные (закрытые) информационные и автоматизированные системы определили условия эксплуатации и требования к их БД:

а) предопределенный и ограниченный круг пользователей с фиксированными функциями и правами, а следовательно, относительно определенная и устойчивая структура (схема) данных; б) равномерный рост общего объема данных с малоизменяющимся объемом оперативных данных;

в) необходимость независимого совместного доступа (изменения) к данным, обусловившая создание моделей транзакционной обработки БД;

г)эффективная работа в реальном времени.

Средства реализации корпоративных информационных систем (КИС), использующие современные серверы баз данных, обеспечивают) фундаментальные требования к хранению данных:

1. Consistency — согласованность, понимаемая как целостность по ограничениям;
2. Avaliability — доступность данных;
3. Partition Tollerance — распределение БД по физическим узлам (стабильная работа при линейно растущем объеме).

Однако общие тенденции в глобализации производства, электронной коммерции и информатизации общества формулируют новые требования и стимулируют развитие информационных систем:

а) создание новых моделей данных, не требующих строго фиксированной структуры;

б) использование парадигмы объектно-ориентированного программирования в СУБД;

Новые требования к информационным системам выявили недостатки используемых в них реляционных СУБД:

1.Строгая типизация, приводящая к несоответствию структуры БД структуре данных реального объекта. Для хранения в реляционной базе данные одного информационного объекта должны быть декомпозированы и распределены по множеству равноценных нормализованных таблиц.

2.Атомарность (единственность и неделимость) данных не адекватно представляет множественные свойства и групповые данные.

3.Статичность данных. Серверы реляционных баз данных (РБД) не имеют специальных средств для представления истории изменения данных.

4.Отдельное от информационного объекта хранение и выполнение его собственных действий. Поведение объекта в РБД описывается в виде хранимых в базе функций, процедур и триггеров, не принадлежащих информационному объекту.

5.Плохая масштабируемость, вызывающая стремительное падение производительности при росте объема данных и количества используемых в запросах соединений (JOIN) таблиц.

6. Неустойчивость к отказам оборудования.

При наличии существенных недостатков необходимо помнить и учитывать достоинства реляционной модели данных, обусловливающие ее продолжающееся использование в КИС:

· наглядность исходного табличного представления данных и результатов запросов;

· реляционная полнота языка SQL-запросов, расширенная мощными средствами обработки данных;

· независимость запросов от физической структуры данных (наличия указателей и связей) — возможность построить любой новый запрос без изменений и дополнений в структуре БД.

Большие данные диктуют новые взаимосвязанные принципы обработки информации /2/. Первый — это способность анализировать все данные, а не довольствоваться их частью или статистическими выборками. Второй — готовность иметь дело с неупорядоченными данными в ущерб точности. Третий — изменение образа мыслей: доверять корреляциям, а не гнаться за труднодостижимым поиском причинно-следственных зависимостей.

Одним из альтернативных решений является применение теории и практики хранилищ данных. Согласно классическому определению Б. Инмона, хранилище данных представляет предметно-ориентированный, интегрированный, привязанный ко времени и неизменяемый набор данных, предназначенный для поддержки принятия решений /4/.

Перечисленные характеристики принципиально отличают хранилище данных от базы данных OLTP-системы.

Предметная ориентированность предполагает, что данные в хранилище данных объединяются в категории и хранятся в соответствии с областями, которые они описывают (продажи, бюджетные поступления и т.п.).

Реализуется предметная ориентированность через использование особых схем организации данных. Это позволяет упростить создание аналитических запросов, а также увеличить скорость их выполнения.

Предметная ориентированность данных в хранилище данных отличается от организации данных в базе данных OLTP-систем, группирующихся в соответствии с бизнес-операциями (выписка счетов, отгрузка товара и т.п.), оптимизированных на интенсивное оперативное выполнение операций вставки/обновления/удаления небольших порций данных; имеющих, как правило, нормализованную реляционную структуру баз данных; удовлетворяющих жестким требованиям ссылочной целостности данных.

Интегрированность означает, что данные для анализа не берутся напрямую из источников, в частности, баз данных OLTP-систем предприятия. Исходные данные извлекаются, проверяются, очищаются,

унифицируются и т.п. (см. ETL в словаре), чтобы удовлетворять требованиям аналитика, исследующего не отдельную бизнес функцию, а деятельность всего предприятия. Процесс ETL позволяет повысить качество данных аналитической системы и скорость выполнения аналитических запросов к хранилищу данных.

Привязка ко времени подразумевает, что данные в хранилище всегда жестко «привязаны» к определенному периоду времени. Данные, выбранные из баз данных OLTP-систем и других источников, накапливаются в хранилище в виде исторических слоев, каждый из которых относится к конкретному периоду времени. Это позволяет анализировать тенденции в развитии бизнеса.

Неизменяемость означает, что, попав в определенный исторический слой хранилища, данные уже никогда не будут изменены. Стабильность данных способствует достоверности результатов аналитических запросов. Это также отличает хранилища данных от баз данных OLTP систем.

Таким образом, хранилище можно охарактеризовать как особую базу данных, содержащую данные:

а) масштаба организации;

б) собранные из различных источников за длительный период времени, очищенные и консолидированные;

в) привязанные ко времени;

г) структурированные в целях упрощения выполнения аналитических запросов.

Заключение

Актуальность технологии хранилищ данных обусловлена их практической значимостью для анализа больших объемов данных. Исходные данные преобразуются таким образом, чтобы наглядно отразить структуру деятельности предприятия. При этом имеется возможность использовать данные из разных источников, притом источники информации могу быть нереляционными.

ЛИТЕРАТУРА

1. Рагимова С. Большие данные (Big Data) — одна из ключевых технологий будущего // Коммерсант.ru: сайт. Режим доступа: http://www.kommersant.ru/doc/2614791?9f476940
2. Рост объема информации — реалии цифровой вселенной // Технологии и средства связи. 2013. № 1. Режим доступа: http://www.tssonline.ru/articles2/fix-corp/rost-obemainformatsii — realii-tsifrovoy-vselennoy.
3. Майер-Шенбергер В., Кукьер К. Большие данные : Революция, которая изменит то, как мы живем, работаем и мыслим. М. : Манн, Иванов и Фербер, 2013. 240 с.
4. Codd, E.F.; Codd S.B. and Salley C.T. Providing OLAP to UserAnalysts: An IT Mandate (1993).