ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

Казакова И.А.
доцент кафедры «Математическое обеспечение и применение ЭВМ» Государственного образовательного учреждения высшего профессионального образования «Пензенский государственный университет»
Kazakova I.A.

ХРАНИЛИЩА И ВИТРИНЫ ДАННЫХ В СИСТЕМАХ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

Аннотация. *В статье доказывается потребность и возможность использования хранилищ и витрин данных для анализа и обработки данных с целью принятия решений.*

STOREHOUSES AND SHOW-WINDOWS OF THE DATA IN SYSTEMS OF SUPPORT OF DECISION-MAKING

SUMMARY. *In article the requirement and possibility of use of storehouses and show-windows of the data for the analysis and data processing for the purpose of decision-making is proved.*

Ключевые слова: хранилище данных, база данных, витрина данных, таблица фактов, таблица измерений, бизнес-аналитика.

Keywords: data warehouse, database, data mart, the table of the facts, the table of measurements, business analytics.

По мере своего развития любая организация накапливает значительное количество информации [1]. Это, например, масса несовместимых таблиц с данными о клиентах, ни одна из которых не содержит полную информацию и не синхронизирована с другими. Данные в этих таблицах имеют разнообразную и сложную структуру. Со временем объемы этой информации уже не позволяют эффективно использовать данные при принятии управленческих решений. Этому мешают различия в способах хранения информации в транзакционных системах, отвечающих за оперативную обработку данных. Кроме того, в транзакционных базах данных достаточно сложно выполнить

анализ информации за предыдущие годы. Информация несогласованна, рассредоточенна, часто избыточна и не всегда достоверна. При поиске решения этих проблем и возникла идея создания хранилищ данных.

Концепция хранилищ информации появилась в 80-х годах XX века в фирме IBM. Считается, что первой публикацией, посвященной хранилищам данных, была статья Барри Девлина и Пола Мэрфи, опубликованная в журнале IBM Systems Journal в 1988г. Статья называлась «Архитектура деловых и информационных систем» («An Architecture for a Business and Information System»). В 1992 году Уильям Г.Инмон – технический директор компании Prism – в своей монографии «Построение хранилищ данных» («Building the Data Warehouse») дал определение хранилища данных:

Хранилище данных (англ. *Data Warehouse*) – предметно-ориентированная, интегрированная, вариантная во времени, неразрушаемая совокупность данных, предназначенная для поддержки принятия управленческих решений.

В настоящее время чаще используют другое определение, практически мало отличающееся от классического: хранилище данных – это предметно-ориентированная информационная корпоративная база данных, специально разработанная и предназначенная для анализа бизнес-процессов в организации с целью поддержки принятия решений.

Основная цель хранилищ – создание единого логического представления данных, содержащихся в разнотипных базах данных или в единой модели корпоративных данных.

Хранилищам данных присущи следующие черты [2]:

Предметная ориентированность. Информация в хранилище организована в соответствии с основными аспектами деятельности предприятия (заказчики, продажи, склад и т. п)

Интегрированность. В разных базах одни и те же данные могут быть выражены в разных единицах измерения. При загрузке в хранилище данные должны быть проверены, очищены и приведены к единому виду. Анализировать такие интегрированные данные намного проще.

Привязка ко времени. Данные, выбранные из оперативных баз данных, накапливаются в хранилище в виде «исторических архивов», каждый из которых относится к конкретному периоду времени. Это позволяет анализировать тенденции в развитии бизнеса.

Неизменяемость. Попав в хранилище, данные уже никогда не меняются. Стабильность данных облегчает их анализ.

Существуют следующие отличия типичных хранилищ данных от обычной реляционной базы данных:

1. Базы данных предназначены для автоматизации бизнес-процессов, тогда как основной задачей хранилищ данных является содержательный анализ информации для качественного функционирования систем поддержки принятия решений. Например, продажа товара и выписка счета производятся с использованием базы данных, предназначенной для обработки транзакций, а анализ динамики продаж за несколько лет, позволяющий спланировать работу с поставщиками, с помощью хранилища данных. Именно поэтому архитектура построения хранилища и принципы проектирования модели данных отличны от тех, что применяются в оперативных системах.

2. Базы данных постоянно изменяются в процессе работы пользователей, а хранилище данных достаточно стабильно – данные в нем обычно обновляются по определенному графику. Эти данные не удаляются и не обновляются непосредственно, а только косвенно – путем приема в загрузочную секцию новых данных. Данные, поступающие в хранилище данных, становятся доступны только для чтения. Можно утверждать, что данные в хранилище точны и корректны в том случае, если они привязаны к некоторому промежутку или моменту времени. Так как данные загружаются в хранилище с определённой периодичностью, актуальность данных несколько отстает от систем, основанных на базах данных.

3. Базы данных чаще всего являются источником данных для хранилищ.

Идея витрин данных (Data Mart) возникла, когда выяснилось, что разработка корпоративного хранилища – длительный и дорогостоящий процесс, требующий значительных усилий по анализу деятельности организации и переориентации ее на новые технологии. Витрины данных возникли с целью избежать трудностей разработки и внедрения хранилищ.

Витрина данных – это специализированное хранилище, обслуживающее, как правило, единственное направление деятельности организации, например, учет складских запасов. Построение витрин данных является менее затратным процессом, чем построение хранилищ данных, так как бизнес-процессы, происходящие в одном из направлений деятельности организации, лучше изучены и не столь сложны, как процессы в масштабах всей организации. При современном уровне развития информационных технологий витрину подразделения можно организовать за 2-3 месяца. Установлено, что наиболее оптимальный вариант использования витрин – это обслуживание 10-15 человек. Необходимо отметить, что успех небольшого проекта (стоимость которого невелика по сравнению со стоимостью разработки корпоративного хранилища), приводит к быстрой окупаемости затрат и способствует продвижению новых технологий. Достоинствами применения витрин данных является физическое разделение данных между группами аналитиков, а также относительная простота семантики данных в пределах одной витрины.

Успех внедрения витрин привел к появлению концепции замены корпоративного хранилища совокупностью витрин данных. Однако эксплуатация витрин данных показала, что с увеличением количества витрин в организации возрастает сложность их взаимодействия, так как не удается сделать витрины полностью независимыми от хранилищ данных. Поэтому чаще всего разработка корпоративного хранилища идет параллельно с разработкой и внедрением витрин данных.

При построении схемы взаимодействия корпоративного хранилища и витрин данных рекомендуется определить некоторую специальную структуру для хранения исторических данных и дополнительно развернуть ряд витрин, заполняемых данными из этой структуры. Тем самым удается разделить два процесса: накопление исторических данных и их анализ.

Фактическим стандартом структуры данных при разработке хранилищ и витрин данных является «звезда», основанная на единственной таблице фактов и множестве таблиц измерений.

Таблица фактов содержит числовые параметры и имеет, как правило, небольшое количество полей – не более двадцати. Она может состоять из миллионов строк и содержать суммирующие или фактические данные, ко-

торые могут помочь ответить на требуемые вопросы. В этой таблице соединяются данные, которые хранились бы во многих таблицах традиционных реляционных баз данных.

Таблицы измерений содержат описательную информацию о числовых значениях в таблице фактов, т.е. они содержат атрибуты фактов. Это неизменяемые или редко изменяемые данные. Обычно в них содержится значительно меньше строк, чем в таблицах фактов, но значительно большее число полей. Атрибуты таблиц измерений обычно используются при визуализации данных во всевозможных отчетах и запросах.

Таблица фактов и таблицы измерений связаны идентифицирующими связями. Таблица фактов, как правило, содержит уникальный составной ключ, объединяющий первичные ключи таблиц измерений. Чаще всего это целочисленные значения либо значения типа «дата/время». В размерной модели направления связей явно не указываются, они определяются типом таблиц. У таблицы фактов есть внешний ключ к каждой соответствующей таблице измерений. Эти внешние ключи можно использовать двумя способами:

1. Для соединения с таблицей измерений с целью выбора описательной информации об этом измерении.

2. Для выполнения обобщений в таблице фактов. Так, можно суммировать содержащиеся в таблице фактов количественные показатели, которые относятся к конкретному покупателю.

Например, в описании продаж таблица фактов может содержать данные о том, какое количество товара реализовано и на какую сумму, а также внешние ключи к таблицам измерений, которые характеризуют операцию продажи (какой товар, когда и кем был продан, какой способ платежа был выбран) [3].

Отношения между таблицей фактов и таблицами измерений должны быть простыми, должен существовать только один возможный путь соединения любых двух таблиц, а смысл этого соединения должен быть очевиден и хорошо понятен.

ЛИТЕРАТУРА

1. Туманов, В.Е. Проектирование хранилищ данных для систем бизнес-аналитики [Текст] / В.Е. Туманов. – М. : Интернет-университет информационных технологий; БИНОМ; Лаборатория знаний, 2010. – 615 с.

2. Туманов, В.Е., Маклаков, С.В. Проектирование реляционных хранилищ данных [Текст] / В.Е. Туманов, С.В. Маклаков. – М. : Диалог-МИФИ, 2007. – 333 с.

3. Чубукова, И. А. Data Mining [Текст] : учеб. пособие / И.А. Чубукова. – М. : Интернет-университет информационных технологий; БИНОМ; Лаборатория знаний, 2006. – 382 с.