МОДЕЛЬ МНОГОМЕРНОГО ПРЕДСТАВЛЕНИЯ ДАННЫХ В ХРАНИЛИЩАХ ДАННЫХ

Геврасёва И. П.
Кафедра информационных технологий автоматизированных систем, Белорусский государственный университет информатики и радиоэлектороники
Объединённый институт проблем информатики Национальной академии наук Беларуси
Минск, Республика Беларусь
E-mail: irina_gevraseva@mail.ru

Рассматривается многомерная модель данных, ее основные понятия и схемы, используемые при проектировании хранилища данных.

Введение

Типичная структура хранилища данных существенно отличается от структуры обычной операционной базы данных. Как правило, в базах данных используется реляционная модель данных, но для хранилищ данных более эффективным является приминение многомерной, или пространнсвенной, модели данных, основанной на использовании таблиц фактов и измерений.

I. Осовные понятия для многомерной модели данных

Многомерное моделирование - это метод логического проектирования данных в стандартной интуитивно понятной структуре для высокопроизводительного доступа, состоящей из одной таблицы с многокомпонентным ключом, называемой таблицей фактов, и набора меньших таблиц, называемых таблицами измерений.

По сравнению с реляционной моделью многомерная организация данных обладает более высокой наглядностью и информативностью.

Основными понятиями для многомерных моделей данных являются:

- *Агрегируемость* данных означает рассмотрение информации на различных уровнях ее обобщения;
- *Историчность* данных предполагает обеспечение высокого уровня статичности данных и их взаимосвязей, а также обязательность привязки данных ко времени;
- *Прогнозируемость* данных подразумевает задание функций прогнозирования и применение их к различным временным интервалам.

Многомерность модели данных означает не многомерность визуализации цифровых данных, а многомерное логическое представление структуры информации при описании и в операциях манипулирования данными.

Основным достоинством многомерной модели данных является удобство и эффективность аналитической обработки больших объемов данных, связанных со временем.

Недостатком многомерной модели данных является ее громоздкость для простейших задач обычной оперативной обработки информации.

II. Структура многомерной модели данных

Данные в многомерной модели организованны не по третьей нормальной форме, а с использованием двух типов таблиц – таблиц фактов и измерений:

Таблицы Фактов. Таблица фактов – основная таблица хранилища данных. Она содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться.

Таблица фактов, как правило, содержит уникальный составной ключ, объединяющий первичные ключи таблиц измерений. Чаще всего это целочисленные значения либо значения типа «дата/время». Помимо этого, таблица фактов содержит одно или несколько числовых полей-фактов, на основании которых в дальнейшем будут получены агрегатные данные.

Для многомерного анализа пригодны таблицы фактов, содержащие как можно более подробные данные, т.е. данные без агрегации. Обработка, хранящихся в таблицах фактов данных, происходит в дальнейшем.

Таблицы Измерений. Таблицы измерений содержат неизменяемые или редко изменяемые данные. Эти таблицы содержат описательного характера данные, относительно которых будут анализироваться собранные в таблице фактов данные. Скорость роста таблиц измерений должна быть незначительной по сравнению со скоростью роста таблицы фактов.

В зависимости от связей между таблицами фактов и таблицами измерений существует три схемы организации многомерной базы данных – «звезд», «снежинка» и «галактика».

III. Схемы организации многомерной модели данных

Схема «Звезда». На рисунке 1 представлена реализация схемы звезда.

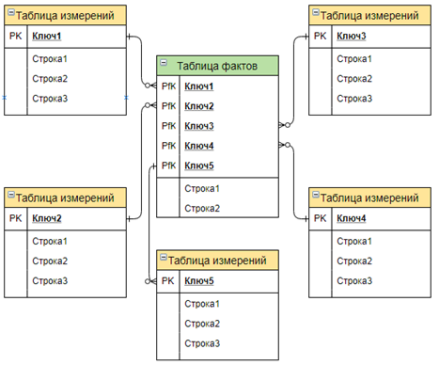

Рис. 1 – Схема «звезда»

В схеме «звезда» таблица фактов, содержащая данные для анализа, является центральной. От таблицы фактов идут связи к таблицам измерений, содержащим описательную информацию. Таблица фактов и таблицы измерений связаны идентифицирующими связями, при этом первичные ключи таблицы измерений мигрируют в таблицу фактов в качестве внешних ключей. Первичный ключ таблицы фактов целиком состоит из первичных ключей всех таблиц измерений. Как правило, таблицы измерений денормализованы и не имеют связей с другими таблицами измерений.

Преимуществом рассматриваемой схемы является то, что благодаря денормализации таблиц измерений упрощается восприятие структуры данных и формулировка запросов, уменьшается количество операций соединения таблиц при обработке запроса, что, в свою очередь, повышает производительность обработки данных.

Недостатком данной схемы в связи с денормализацией таблиц измерений является появление избыточности данных, которое влечет за собой увеличение объема памяти, необходимой для хранения данных.

Схема «Снежинка». Реализация схемы «снежинка» представлена на рисунке 2.

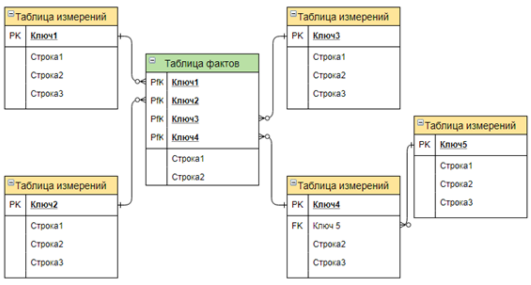

Рис. 2 – Схема «снежинка»

Схема «снежинка» имеет схожие отличительные черты со схемой «звезда». В схеме «снежинка» так же, как и в схеме «звезда», центральной таблицей является таблица фактов, имеющая связи с описательными таблицами измерений. Отличительной чертой рассматриваемой схемы является то, что таблицы измерений имеют более нормализованную структуру и могут быть связаны с другими таблицами измерений, что не характерно для схемы «звезда».

Схема «Галактика». Это более сложная структура, имеющая несколько таблиц фактов, которые могут использовать общую таблицу измерений. В целом ее можно представить, как усложненную версию схемы «звезда». Соответственно, преимущества и недостатки данной схемы такие же, как и у схемы «звезда». Однако более сложная структура данной схемы влечет за собой увеличения сложности запросов для обработки данных.

Нормализация таблиц измерений позволяет минимизировать избыточность данных и более эффективно выполнять запросы, связанные со структурой значений измерений, что является существенным преимуществом схемы ««снежинка». При этом же более нормализованная структура таблиц измерений влечет за собой потерю производительности при выполнении запросов.

Для организации хранилища данных в основном используют схему «звезда», так как производительность является более приоритетным преимуществом.

Список литературы

1. Kimball, R. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition / R. Kimball, M. Ross // Wiley Publishing, Inc. – 2013. – С. 564.
2. Inmon, W. H. Building the Data Warehouse, Fourth Edition / W. H. Inmon // Wiley Publishing, Inc. – 2005. – С. 576.
3. Han, J. Data mining Concepts and Techniques, Third edition / J. Han, M. Kamber, J. Pei // Morgan Kaufmann, Inc. – 2011. – С.744