ХРАНИЛИЩА ДАННЫХ

МЕТОДИЧЕСКИЕ УКАЗАНИЯ
для выполнения лабораторных работ и организации
самостоятельной работы студентов
направления подготовки
«Бизнес-информатика»

Томск - 2013

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ» (ТУСУР)

Кафедра автоматизации обработки информации (АОИ)

УТВЕРЖДАЮ
Зав. Кафедрой АОИ
Д.т.н., профессор
_____ Ю. П. Ехлаков
«___» _____ 2013 г.

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

для выполнения лабораторных работ
и организации самостоятельной работы
по дисциплине «**Хранилища данных**»

для студентов направления
«Бизнес-информатика»

Разработчик
доцент каф. АОИ
канд.техн.наук, с.н.с.
_____О.И. Жуковский

.

Томск – 2013

Содержание

1. Проектирование структуры и функционального наполнения OLTP систем

Создание структуры OLTP системы, необходимой для поддержки принятия решений в процессе управления одним из подразделений, решающим задачи ведения недвижимого имущества (земельный комитет, бюро технической инвентаризации, учреждение регистрации прав на недвижимое имущество, риэлтерская контора, комитет по налогам и сборам).

Разработка требований к Киоску Данных, работающему на основе данных OLTP-системы.

Для выполнения данной работы используйте «Методические указания по выполнению лабораторных работ (часть 2) ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ для студентов специальности 230102 – Автоматизированные системы обработки информации и управления» (МУ).

Разработка и создание функциональной модели процесса

Выполнение задания рекомендуется проводить согласно следующим этапам:

Перед началом моделирования, на основе данных сети Интернет, познакомьтесь с особенностями функционирования выбранного вами подразделения сферы управления недвижимым имуществом и выделите основные его задачи. Сформируйте реферативную записку на основе найденных материалов и представьте ее преподавателю.

Начало моделирования. Создайте диаграммы A0 и A-0 и отрецензируйте их у преподавателя. Обратите внимание, что эти две диаграммы полностью рассказывают все о моделируемой системе с минимальной степенью детализации. Диаграмма A-0, часто называемая контекстной диаграммой, определяет все необходимые связи моделируемого процесса с окружающим миром. В первую очередь создайте диаграмму A0 и обобщив ее создайте диаграмму A-0.

Особое внимание обратите на то, что модель должна иметь определенную цель и создаваться с конкретной точки зрения. Выбор цели осуществляется с учетом вопросов, на которые должна ответить модель, а выбор точки зрения – в соответствии с выбором позиции, с которой описывается система.

Подготовьте список основных типов данных и функций, необходимый для дальнейшей декомпозиции системы. Начните заполнять словарь данных. Помните, что несколько различных типов данных могут использоваться одной функцией.

При создании диаграммы А0 строго следуйте рекомендациям раздела 1.2. МУ Правильное расположение блоков является самым важным этапом построения диаграммы.

Построенные вами на данном этапе диаграммы А-0 и А0 должны представлять законченную картину, поскольку они отражают все основные входы, управления, выходы и функции системы.

Продолжение моделирования. Продолжение моделирования основывается на тех же методах, что и начальный этап и выводит модель на следующий уровень детализации. Создайте отдельную диаграмму для, возможно, каждого блока диаграммы верхнего уровня, затем постройте для всех блоков новые диаграммы и так до тех пор, пока модель не будет описывать объект с нужной для достижения вашей цели степенью детализации.

Попытайтесь создать модель не менее чем с четырьмя уровнями детализации.

Не забывайте пополнять и корректировать словарь данных.

Завершение моделирования. Для определения момента завершения моделирования воспользуйтесь указаниями раздела 1.4. МУ.

По завершении моделирования подготовьте отчеты, предусмотренные программой Design/IDEF: отчет о функциях, отчет о дугах, отчет о ссылках и IDEF-отчет.

Разработка и создание концептуальной модели данных процесса

Постройте концептуальную модель данных процесса. Выполнение задания рекомендуется проводить согласно следующим этапам:

Начало работы над моделью. Определите цель моделирования.

Определение сущностей. Целью данного этапа является выявление и определение сущностей, находящихся в пределах моделируемой проблемной области.

В первую очередь проведите идентификацию сущностей и создайте пул (список) сущностей.

Определение отношений. Отношение определяется как ассоциация или связь между двумя сущностями. Для определения отношений используйте матрицу отношений. В ходе определения отношений укажите зависимость между сущностями, присвойте отношениям имя и заполните комментарии к отношениям.

Постройте диаграммы уровней сущностей. Для наиболее важных сущностей постройте отдельные диаграммы, позволяющие наиболее полно представить все их отношения с другими сущностями.

Постарайтесь использовать отношение категоризации.

Определение ключей. В первую очередь разрешите все неспецифические отношения. Проведите идентификацию возможных ключей каждой сущности, выбрав один из них в качестве первичного ключа сущности. Обратите внимание на выполнение правила неповторяемоти и правила необращения в ноль. Проведите определение ключевых атрибутов.

Определение атрибутов. Данная стадия является завершающей стадией разработки модели. В ходе ее выполнения вы должны разработать пул атрибутов, установить принадлежность атрибутов, определить неключевые атрибуты, проверить правильность и детализацию структуры данных.

Анализ созданной модели IDEF1x. Представьте преподавателю построенную вами модель для анализа, используя ее пул сущностей и глоссарий термином.

Опираясь на полученные модели, сформулируйте требования к OLTP-системе, реализующей оперативную информационную поддержку исследуемого вами процесса управления недвижимым имуществом.

Определите набор запросов аналитика к киоску данных, построенному на основе базы данных предложенной вами OLTP-системы.

2. Проектирование структуры хранилища данных.

Разработайте структуры реляционного Хранилища Данных, ориентированного на поддержку принятия решений в области управления недвижимым имуществом территории. Хранилище основывается на данных OLTP-систем, разработанных в рамках предыдущей темы.

Разработка структуры многомерного хранилища данных.

Многомерное моделирование является методом моделирования и визуализации данных как множества числовых или лингвистических показателей или параметров (measures), которые описывают общие

аспекты деятельности организации. Как правило, при многомерном моделировании, основное внимание фокусируется на числовых данных, таких, как число продаж, баланс, прибыль, вес или объекты, которые можно пересчитать - статьи, патенты, книги.

Моделирование Dimensional сходно с моделированием связей и сущностей для реляционной модели, но отличается целями. Реляционная модель акцентируется на целостности и эффективности ввода данных. Размерная модель (Dimentional) ориентирована в первую очередь на выполнение сложных запросов к базе данных.

Метод многомерного моделирования базируется на трех основных понятиях: фактах, измерениях и параметрах (метриках).

Факт (fact) - это набор связанных элементов данных, содержащих метрики и описательные данные. Каждый факт обычно представляет элемент данных, численно описывающий деятельность организации, бизнес-операцию или событие, которое может быть использовано для анализа деятельности организации или бизнес-процессов. В ХД факты сохраняются в базовых таблицах реляционной БД.

Измерение или размерность (dimension) - это интерпретация факта с некоторой точки зрения в реальном мире. Обычно измерения представляются как оси многомерного пространства, точками которого являются связанные с ними факты. В многомерной модели каждый факт связан с одной или несколькими осями. Измерения обычно представляют нечисловые, лингвистические переменные, такие, как филиалы организации, сотрудники организации, покупатели и т. д.

Например, при анализе продаж продукции, производимой или продаваемой организацией, такими измерениями обычно выступают время, покупатели, продавцы, место продажи или складирования товара. Измерения задаются перечислением своих элементов. Элемент измерения (dimensional member) - уникальное имя или идентификатор (лингвистическая переменная), используемая для определения позиции элемента. Например, измерение время может содержать следующие элементы: все месяцы, кварталы, годы.

Часто элементы измерения находятся в отношении "часть-целое" или "родитель-потомок", что позволяет ввести на измерении одну или несколько иерархий. Каждая иерархия может иметь несколько уровней иерархии (hierarchy levels). Каждый элемент измерения должен принадлежать только к одному уровню иерархии, порождая, таким образом, разбиение на непересекающиеся подмножества. Примером может служить иерархия на измерении время, год, полугодия, кварталы, месяцы и дни. Элемент измерения неделя может

принадлежать двум месяцам, поэтому для некого следует определить другую иерархию.

Параметр или показатель (measure) - это числовой атрибут факта, который характеризует эффективность деятельности или бизнес-действия организации с точки зрения измерения. Конкретные значения показателя описываются с помощью переменных. Например, пусть параметрами является численное выражение продаж товара в деньгах, количество проданных единиц товара и т. д. Параметр определяется с помощью комбинации элементов измерения и представляется как факт.

Многомерная модель визуально представляется с помощью куба (или, в случае более трех измерений, гиперкуба), рис.1.

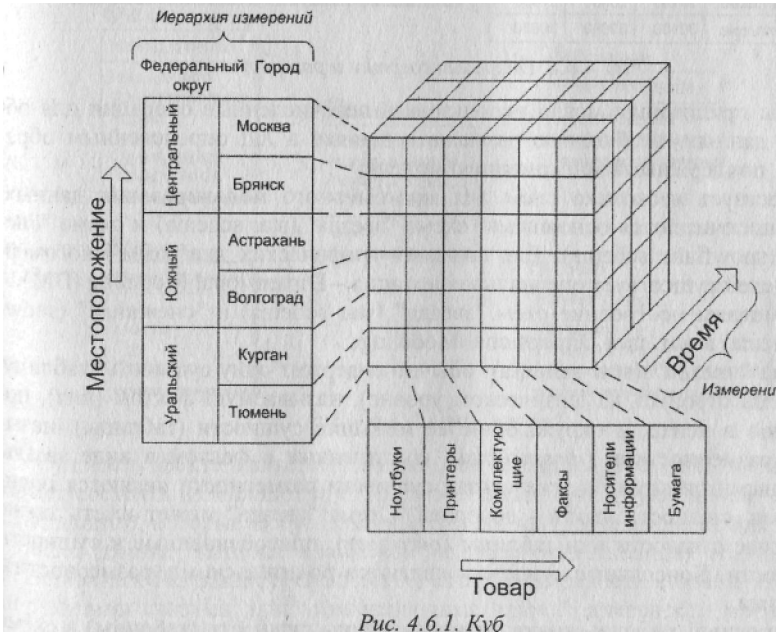

Рис. 1. Пример куба

Многомерное моделирование является основным методом логического Проектирования ХД для OLAP-приложений. Для таких приложений типично выполнение операций свертывания и развертывания данных.

Развертка (drill down) и свертка(drill up) являются операциями перемещения вниз и вверх по уровням иерархии измерения. При выполнении развертки пользователь перемещается от верхних уровней к нижним уровням, которые содержат обычно более подробные данные. При выполнении свертки пользователь перемещается от нижних уровней иерархии к верхним, тем самым обобщая информацию на каждом уровне. При выполнении этих операций путь навигации определяется иерархиями измерений.

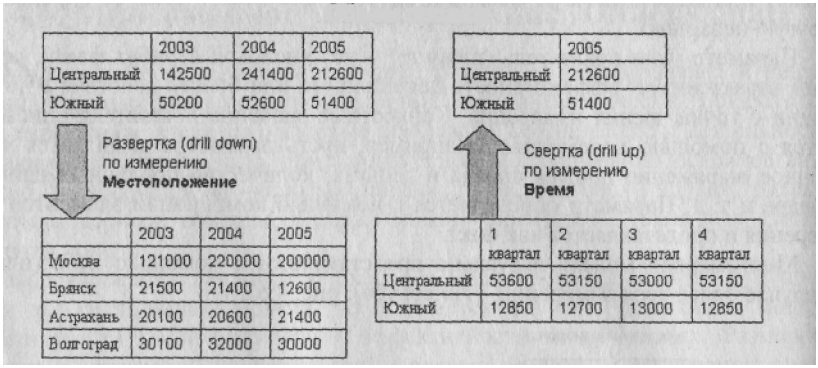

Рис.2. Операции свертки и развертки

Чтобы приложение могло использовать перечисленные операции для обработки данных, необходимо размещать данные в ХД определенным образом, т. е. поддерживать многомерную модель.

Существует несколько схем для многомерного моделирования данных. Две из них считаются основными: схема "звезда" (star schema) и схема "снежинка" (snowflake schema). Для создания графических диаграмм многомерных моделей существует специальная нотация – Dimensional Modeling (DM).

Схема "звезда" обычно содержит одну сущность (таблицу, если схема строится на физическом уровне), называемую фактом (fact), помещенную в центр, и окружающие ее меньшие сущности (таблицы), называемые размерностями (dimensional), соединенные с фактом в виде звезды радиальными связями. В этих связях сущности размерности являются родительскими, сущности факта - дочерними. Схема "звезда" может иметь также консольные сущности или таблицы

(outrigger), присоединенные к сущности размерности. Консольные сущности являются родительскими, размерности - дочерними.

Прежде чем создать БД со схемой типа звезды, необходимо проанализировать бизнес-правила предметной области с целью выяснения центрального вопроса, ответ на который наиболее важен. Все прочие вопросы должны быть объединены вокруг этого основного вопроса, и моделирование должно начинаться с этого основного вопроса. Данные, необходимые для ответа на этот вопрос, должны быть помещены в центральную сущность факта. Например, если необходимо создавать отчеты об общей сумме дохода от продаж за определенный период как по типу товара, так и по продавцам, следует разрабатывать модель так, чтобы каждая запись в сущности факта представляла сумму продаж, осуществленных тем или иным продавцом, с указанием доходов по каждому покупателю и типов проданных товаров (рис.3). В примере сущность факта содержит суммарные данные о продажах (Продажа), а сущности размерности содержат данные о заказчике и заказах (Клиент), продуктах (Товар), продавцах (Продавец) и периодах времени (Время).

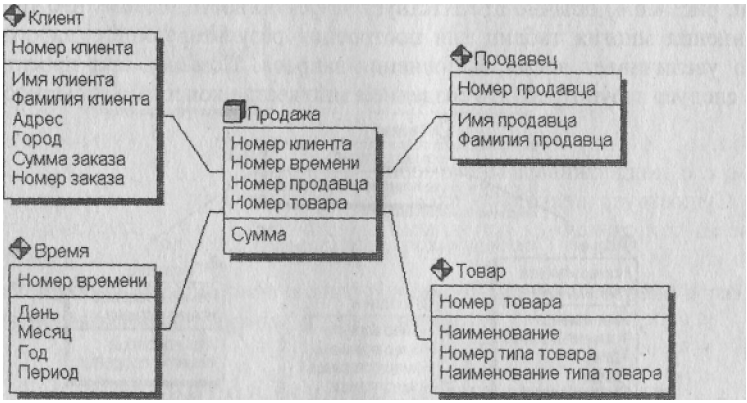

Рис.3. Схема "звезда"

Сущность факта является центральной таблицей в схеме "звезда". Она может состоять из миллионов строк и содержать суммирующие или фактические данные, которые могут помочь ответить на требуемые вопросы. Она соединяет данные, которые хранились бы во многих таблицах традиционных реляционных баз данных. Сущности факта и размерности связаны идентифицирующими связями, при этом первичные ключи размерности мигрируют в сущность факта в качестве внешних ключей. В размерной модели направления связей явно не показываются - они определяются типом сущности в схеме. Первичный ключ сущности факта целиком состоит из первичных ключей всех сущностей размерности. В примере (факта Продажа) первичный ключ составлен из четырех внешних ключей: Номер клиента, Номер, продавца, Номер времени и Номер товара.

Сущности размерности содержат описательную информацию. В примере на рис. 3 Продажа - сущность факта; Клиент, Время, Продавец и Товар - размерности, которые позволяют быстро извлекать информацию о том, кто и когда сделал покупку, какой продавец и на какую сумму продал и какие именно товары были проданы.

Консольные сущности (outrigger), могут быть связаны только с сущностями размерности, причем консольная сущность в этой связи родительская, а размерности - дочерняя. Связь может быть идентифицирующей или неидентифицирующей. Консольная сущность не может быть связана с сущностью факта. Она используется для нормализации данных в таблицах размерности. Нормализация данных полезна при моделировании реляционной структуры, но она уменьшает эффективность выполнения запросов к хранилищу данных. В размерной модели главной целью является обеспечение высокой эффективности просмотра данных и выполнения сложных запросов.

Схема "снежинка" (так называется размерная модель, в которой консольные сущности используются для нормализации каждой сущности размерности, рис.4) обычно препятствует эффективности, потому что требует объединения многих таблиц для построения результирующего набора данных, что увеличивает время выполнения запроса. Поэтому при проектировании не следует злоупотреблять созданием множества консольных сущностей.

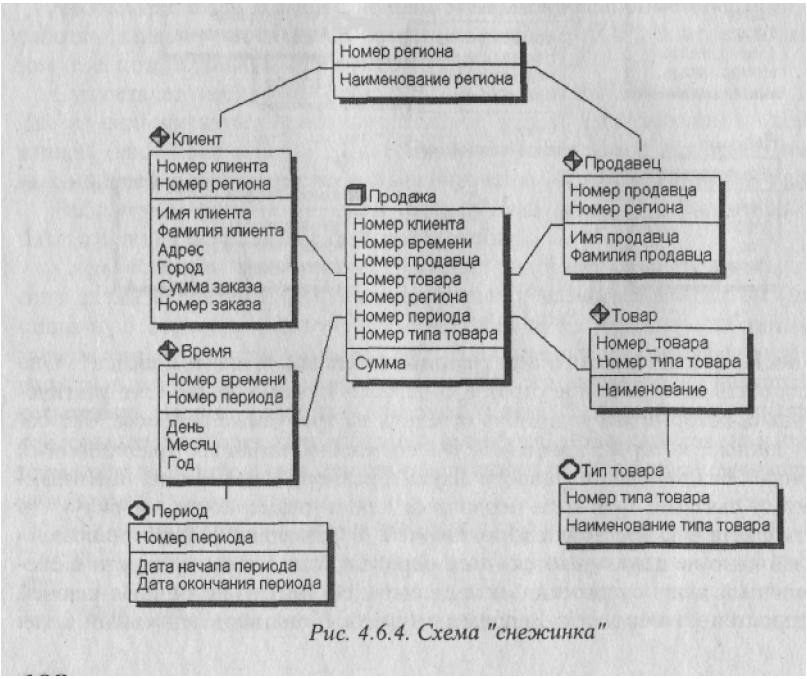

Рис. 4 Схема «снежинка»

Если денормализованная размерность получается слишком большой (таблица размерности содержит слишком много строк), при этом к части колонок запросы делаются чаще, чем к остальным, целесообразно для повышения эффективности разбить одиночную размерность на две отдельные. Две полученные сущности можно связать неидентифицирующей связью. В примере на рис. 3 Товар содержит как информацию о конкретном товаре, так и информацию о типах товаров. Если запросы, связанные с типами товаров, делаются чаще, чем по отдельным товарам, можно создать новую сущность Тип товара и перенести в нее информацию о типах (рис. 5). В этом случае за счет того, что колонки, к которым наиболее часто обращаются запросы, переносятся в новую таблицу, уменьшается время выполнения запроса.

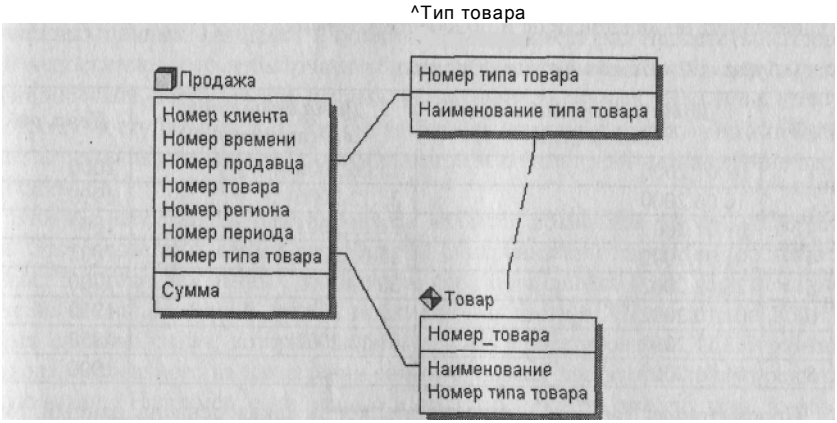

Рис..5. Нормализация размерности

Разработайте структуры вашего хранилища данных на основе схемы «звезда».

3. Разработка комплекса метаданных хранилища данных

Разработка логической структуры метаданных для спроектированного хранилища - метаданных модели, метаданных фактов, метаданных измерений, метаданных источников;

Изучите материал раздела, внимательно просмотрите приметы и предложите логическую структуру комплекса метаданных для разработанного вами на предыдущих занятиях хранилища данных

Базовые компоненты метаданных ХД не сильно отличаются от базовых компонентов систем оперативной обработки данных. Это описание таблиц, их атрибутов и ключей. Существенное отличие для ХД - поддержка версионности метаданных. Базовые компоненты говорят нам, какие данные сохраняются в ХД. Следующая, характерная для ХД, группа компонентов метаданных - это описание

преобразований. Как правило, описание преобразований данных для ХД включает в себя:

идентификацию полей источников данных;

соответствие между атрибутами сущностей источников данных и атрибутами объектов ХД;

преобразования атрибутов;

физические характеристики преобразований;

преобразования таблиц кодировки и ссылочных таблиц;

изменения наименований (соответствие имен источников и объектов ХД);

изменение ключевых атрибутов;

значение полей по умолчанию;

логика (алгоритмы) формирования данных ХД из нескольких источников (приоритетность источников);

алгоритмы трансформации данных и т. д.

Компоненты преобразования говорят нам о том, как данные в ХД были получены.

Немаловажным компонентом метаданных ХД является история поступления в него данных. Компонент метаданных история экстрагирования (поступления) данных, говорит нам о том, когда данные поступили в ХД, а также позволяет судить о полноте представления данных в ХД. Для проведения анализа данных такая информация является очень важной, поскольку на ее основе формируются утверждения пользователей о корректности анализа и надежности его результатов.

Информация о синонимии, или о терминологическом соответствии понятий, - это еще один компонент метаданных ХД. Он включает в себя альтернативные наименования (алиасы) для данных ХД. Такая информация, как правило, делает ХД более "дружелюбным" для пользователей.

Одним из важных компонентов метаданных является информация о состояниях и статистики использования данных ХД. Эта

информация составляет основу для оптимизации производительности ХД. К такой информации относятся данные о числе строк в таблицах, скорости роста таблиц, статистических профилях использования таблиц (среднее и максимальное число запросов на день), статистика архивирования и удаления данных, индексирование таблиц, частота использования индексов в запросах и т. п.

Еще одним компонентом метаданных ХД являются алгоритмы агрегации и суммирования данных, критерии выборки из источников, правила преобразования данных источников перед загрузкой в ХД, описание взаимосвязей между объектами ХД, их кардинальность и т. п. Такая информация играет важную роль при проведении анализа данных и часто требуется аналитиками для решения вопросов надежности результатов анализа.

Информация о том, кто отвечает за содержание и актуальность различных источников данных, составляет еще один компонент метаданных. Эта информация важна для группы сопровождения ХД и позволяет организационно решать вопросы качества, точности и надежности данных в ХД.

Часто в метаданные включаются компоненты, описывающие шаблоны доступа к данным (когда и как данные мигрировали на другой уровень хранения). Они используются также для оптимизации физического потока данных в ХД и для оптимизации производительности. Иногда, алгоритмы обработки данных в ХД используют информацию из объектов внешних систем, так называемые таблицы расширения (таблицы кодировок и электронных справочников). В этом случае в метаданных ХД необходимо фиксировать описание таких таблиц и историю их изменения, поскольку в случае изменения кодов необходимо провести соответствующие изменения в обработке данных ХД, чтобы не потерять исторические связи в данных.

Часто проектировщики ХД включают в состав метаданных дополнительную информацию, важную с их точки зрения.

Из сказанного выше ясно, что проектирование метаданных ХД является достаточно сложной и креативной задачей для

проектировщика, решение которой требует часто литературного мастерства, знания предметной области ХД и очень много времени.

Рассмотрим, как можно описать метаданные на примере киоска данных, предназначенных для анализа продаж некоторой гипотетической компании. Компания занимается производством и реализацией своей продукции. Киоск данных используется аналитиками для детального изучения взаимосвязи расходов и доходов компании от реализации продукции и подготовки отчетности о продажах для руководства.

Допустим, что наша гипотетическая компания открыла сеть точек продаж (склады розничной и оптовой торговли), сеть складов, т. е. сделала расширение своей деятельности. Руководство компании хочет оценить эффективность сделанного расширения и иметь более подробную информацию о зависимости между продажами и производством по затратам и доходам.

Компания выпускает около 200 видов (моделей) некоторой продукции. Каждый продукт имеет базовый набор комплектующих компонентов. Дополнительные комплектующие компоненты используются для создания специфической модели продукта. Политика компании строится таким образом, что число выпускаемых моделей остается постоянным. Это означает, что количество новых моделей приблизительно равно количеству моделей, снятых с производства.

Для каждой модели каждого продукта в зависимости от спроса применяется гибкая система скидок. Как правило, размер скидки для покупателей больших партий продукции определяет заведующий складом розничной продажи

Когда принято решение приостановить производство продукции данной модели, информация о ней сохраняется в БД компании в течение 6 мес. после того, когда вся оставшаяся продукция будет реализована или списана.

Данные о продукции удаляются в тот момент, когда удаляются данные о последней модели этой продукции.

Компания поддерживает 2 способа реализации продукции: через склад оптовой торговли и через склад розничной торговли. Склад оптовой торговли продает товар только оптовым покупателям. Покупатель считается оптовым, если он покупает более 20 партий товара в год. Оптовый покупатель может предоставлять счет либо непосредственно на склад оптовой торговли, либо по факсу в центральный офис компании. Любой покупатель может покупать на нескольких складах компании.

Склад розничной торговли продает за наличный расчет. Независимо от предоставления скидок цена товара меняется. Хотя на каждую продажу продукции оформляется счет, компания не ведет учет покупателей для розничной продажи.

Киоск данных нашей компании предназначен для решения задач анализа показателей расхода и дохода. Типовые запросы, на которые система должна давать ответы, следующие:

Какова величина общих издержек и общей прибыли по каждой модели товара, проданной сегодня и просуммированной по точкам продажи, типу точки продажи, по региону и по складам оптовой торговли?

Какова величина общих издержек и общей прибыли для каждой модели товара, проданной сегодня и просуммированной по заводам и по регионам?

Какой процент моделей получили скидки и какие из них были проданы по факту со скидкой (в процентах) в складах розничной продажи для всех продаж на этой неделе? В этом месяце?

Для каждой модели товара, проданной в текущем месяце, определить, какой был процент продаж с розничной торговли, с оптовой торговли по безналичному расчету, с оптовой торговли продавцами?

Какие модели и какого типа не продавалась в течение последнего месяца? В течение последней недели?

Какие 5 моделей, проданных за последний месяц, принесли наибольшую прибыль? По продажам за квартал? По всем продажам?

Источником данных для киоска данных является фрагмент БД системы оперативной обработки данных компании. Одна из возможных структур данных киоска данных, полученная в результате проектирования, приведена на рис 6.

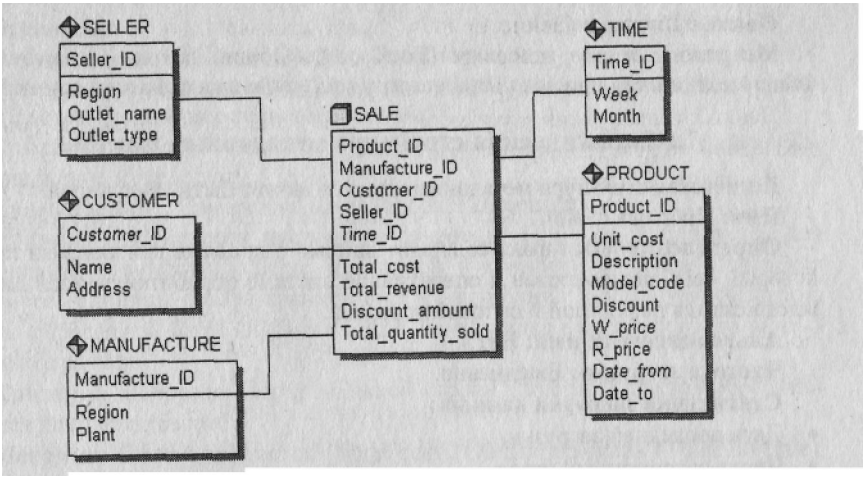

Рис. 6. Многомерная модель киоска данных для анализа продаж компании

Рассмотрим описание метаданных для такого киоска данных. Отметим, что приведенное описание является примером одного из возможных подходов, его нельзя считать полным и законченным.

Логическая структура метаданных модели. В этом разделе приводятся логические схемы метаданных для ХД для взятого нами примера. Пример не претендует на полноту, но дает ясное представление о подходах к описанию метаданных.

Логическая структура модели метаданных может быть следующей.

Имя: Продажи (Sales).

Определение: Модель метаданных содержит метаописание данных о продажах компании для каждого вида продукции, в соответствие с каждым оплаченным счетом на ежедневной основе.

Назначение: Назначением данной модели является предоставление аналитикам и руководству компании возможностей для анализа продаж,

Ответственное лицо за корректность данных: Региональный менеджер по продажам.

Измерения: Покупатель (Customer), Производитель (Manufacture), Продукт (Product), Продавец (Seller) и Время (Time

Факты: Продажа (Sale).

Метрики: Общие издержки (Total cost), Общий доход (Total revenue), Общее количество продаж (Total quantity sold) и Скидка (Discount amount).

Логическая структура метаданных фактов.

Логическая структура метаданных фактов может быть следующей.

Имя: Продажа (Sale).

Определение: Этот факт содержит данные о продаже для каждого заказа, который был зафиксирован в оперативной системе обработки заказов для каждого склада розничной и оптовой торговли.

Альтернативное имя: Нет.

Частота загрузки: Ежедневно.

Статистика загрузки данных:

- Дата последней загрузки.
- Число загруженных строк.

Статистика использования данных:

- Среднее число запросов в день.
- Среднее число выбранных записей на запрос.
- Средне время выполнения запроса.
- Максимальное число запросов в день.
- Максимальное число выбранных записей в запросе.
- Максимальное время выполнения запроса.

Правила архивирования данных: Данные будут архивироваться по истечении 36 мес. на ежемесячной основе.

Статистика архивирования: Последняя дата архивации.

Правила удаления данных: Данные будут удаляться по истечении 48 мес. на ежемесячной основе.

Статистика удаления: Последняя дата удаления.

Качество данных: Допускаются ошибки персонала при комплектовании заказов. Однако записи, представленные в БД, являются точными.

Точность данных: Метрики этого факта являются на 100 % точными, поскольку представляют уже осуществленные продажи.

Гранулированность измерения Время: Метрики данного факта представляют продажу данного товара по данному заказу.

Ключевое поле: Ключом для факта продажи является комбинация ключей измерений: Покупатель (Customer), Производитель (Manufacture), Продукт (Product), Продавец (Seller) и Время (Time).

Метод генерирования ключа: Временная часть ключа есть просто дата продажи товара. Ключ товара, ключ производителя, ключ продавца и ключ покупателя выбираются из справочников оперативной БД компании.

Источники

Наименование: Таблица заказов (Order Table).

Правила преобразования: Строки из таблицы заказов копируются в таблицу фактов продаж на ежедневной основе.

Критерий выборки данных: Выбираются строки, для которых заказ был завершен на текущую дату.

Наименование: Измерение Товар (Product Dimension).

Правила вычисления значения: Измерение Продукт используется для вычисления стоимости модели, проданной в конкретном заказе. Заводская стоимость единицы товара сравнивается с закупочной или отпускной ценой, чтобы определить, была ли дана скидка. Если скидка имела место, то вычисляется ее размер.

Критерий выборки: Перед вставкой строки в таблицу фактов обрабатываются данные о товаре.

Метрики: Общая стоимость (Total cost), Общая прибыль (Total revenue), Общее количество продаж (Total quantity sold) и Скидка (Discount amount).

Изменения: Покупатель (Customer), Производитель (Manufacture), Продукт (Product), Продавец (Seller) и Время (Time).

Сотрудник, ответственный за данные: Директор завода производителя.

Логическая структура метаданных измерений

Логическую структуру метаданных для измерений приведем на примере измерений Покупатель и Время. Она может быть следующей:

Для измерения Покупатель

Имя: Покупатель (СизСотег).

Определение: Покупатель - это любое физическое или юридическое лицо, которое приобретает продукцию компании. Покупатель может приобретать товары в нескольких точках продаж компании.

Альтернативное имя: Нет.

Иерархия измерения: Данные по этому измерению могут суммироваться на двух уровнях. Первый уровень суммирования (нижний) есть адрес отгрузки товара покупателю. Данные по каждому адресу юридического лица могут быть позднее просуммированы по каждому покупателю.

Правила изменения: Адреса отгрузки товара по каждому юридическому лицу вставляются как новые строки в измерение. Изменение существующих адресов покупателей выполняется обновлением непосредственно в таблице измерения.

Частота загрузки: Ежедневно.

Статистика загрузки:

- Последняя дата загрузки.
- Количество загруженных строк.

Статистика использования:

- Среднее число запросов за день.
- Среднее число выбранных строк на запрос.
- Среднее время выполнения запроса.
- Максимальное число запросов за день.
- Максимальное число выбранных строк на запрос.
- Максимальное время выполнения запроса.

Правила архивации: Данные этого измерения не архивируются.

Статистика архивации: Дата последней архивации.

Правила удаления: Покупатели, которые не приобретали продукцию компании в течение последних 5 лет, удаляются из таблицы измерения на ежемесячной основе.

Статистика удаления: Дата последнего обновления.

Качество данных: Когда новый покупатель добавляется в измерение, выполняется поиск, чтобы определить, не было ли продаж товара данному покупателю по другому адресу. Независимо от того, были ли такие продажи, покупатель с новым адресом отгрузки товара вставляется как новая строка.

Точность данных: Допускается 5 %-неточность в определении связей между покупателем и его адресами отгрузки.

Ключ измерения: Сгенерированное системой число, которое идентифицирует покупателя.

Метод генерации ключа: Когда запись о покупателе копируется из подающей системы, выполняется проверка на присутствие покупателя в ХД. Если такого покупателя нет в ХД, новый идентификатор генерируется и запись вставляется в измерение.

Источники

Имя (Name): Таблица Покупатель (Customer).

Правила преобразования: Строки из таблицы Покупатель подающей системы копируются ежедневно.

Критерий выборки: Выбираются только новые или модифицированные на текущую дату строки.

Имя: Таблица Адреса покупателей (Customer Location).

Правила преобразования: Строки из таблицы Адреса покупателей копируются ежедневно в таблицу измерения. Для существующих адресов покупателей адрес отгрузки обновляется. Для новых адресов покупателей ключ генерируется и записи вставляются.

Критерий выборки: Выбираются только те записи, которые на текущую дату были обновлены или добавлены.

Атрибуты

Имя: Идентификатор покупателя (Customer Key).

- **Определение**: Это есть произвольно выбранное число, гарантирующее уникальность каждого покупателя и его адреса.
- **Правила изменения**: После вставки в измерение значение этого атрибута никогда не изменяется.
- **Тип данных**: Числовой.
- **Домен**: 1-999999999.
- **Правила вычисления значения**: Сгенерированный системой ключ.
- **Источник:** Генерируется системой**.**

Имя: Наименование (Name).

- **Определение**: Наименование, под которым покупатель известен компании.
- **Правило изменения**: При изменении наименования покупателя оно обновляется для всего этого измерения.

- **Тип данных**: Символьный.
- **Домен:** Допустимая строка символов.
- **Правила вычисления значения**: Для того чтобы различать покупателей из разных организаций с одинаковым названием, к названию организации будет добавляться число.
- **Источник:** Поле Наименование (Name) из таблицы покупателей (Customer) подающей системы.

Имя: Адрес отгрузки (Ship-to Addres).

- **Определение**: Для юридических лиц - это адрес, по которому отгружается товар. Допускается, что одно юридическое лицо может иметь несколько адресов отгрузки. Для физических лиц и розничных покупателей это поле не поддерживается. Таким образом, для таких покупателей в таблице измерения поддерживается только одна запись.
- **Правила изменения**: При изменении адреса отгрузки выполняется обновление этого значения в измерении.
- **Тип данных**: Символьный.
- **Домен**: Запись адреса в допустимом формате.
- **Правила вычисления значения**: Адрес отгрузки копируется из таблицы источника.
- **Источник:** Поле Адрес отгрузки (Ship-to Addres) из таблицы Адреса покупателей (Customer Location) подающей системы.

Факты: Продажа (Sale).

Метрики: Общая стоимость (Total cost), Общая прибыль (Total revenue), Общее количество продаж (Total quantity sold) и Скидка (Discount amount).

Ответственный за поставку данных: Вице-президент по продажам и маркетингу

Для измерения Время

Имя: Время (Time).

Определение: Измерение Время содержит моменты времени, когда компания фиксирует данные о продажах.

Альтернативное имя: Нет.

Иерархия измерения: Наименьший уровень суммирования данных есть день. Данные для этого дня могут быть просуммированы либо за неделю, либо за месяц.

Правила изменения: Записи вставляются в измерение один раз за текущий год. Никакие обновления в этом измерении не допускаются.

Частота загрузки: По мере необходимости.

Статистика загрузки:

- Дата последней загрузки.
- Число загруженных строк.
- Статистика использования:
- Среднее число запросов за день.
- Среднее число выбранных строк на запрос.
- Среднее время выполнения запроса.
- Максимальное число запросов за день.
- Максимальное число выбранных строк на запрос.
- Максимальное время выполнения запроса.

Правила архивации: Данные этого измерения не архивируются.

Правила удаления: По истечении 5 лет данные будут удаляться на ежегодной основе.

Статистика удаления: Дата последнего удаления.

Качество данных: Никаких ошибок в данных этого измерения не предполагается.

Точность данных: Данные этого измерения всегда точны.

Ключ измерения: Ключ измерения Время есть дата в формате ГГГГММДД.

Метод генерации ключа: Дата, представленная в строке, используется как значение ключа.

Источник

Имя: Календарь, поддерживаемый администратором.

Правила преобразования: Все строки календаря вставляются один раз в год.

Критерий выборки: Все строки выбираются.

Атрибуты

Имя: Идентификатор (Time_ID).

- **Определени**е: Это есть дата в формате ГГГГММДД.
- **Альтернативное имя**: Нет.
- **Правила изменения**: После вставки значение этого поля никогда
  не изменяется.
- **Тип данных**: Числовой.
- **Домен**: допустимое знание для даты.
- **Правила вычисления значения**: Дата есть копия значения источника.
- **Источник**: Числовое значение даты из календаря.

Имя: Месяц (Month).

- **Определение**: Номер месяца в году.
- **Альтернативное имя**: Нет.
- **Правила изменения**: После вставки значение этого поля никогда
  не изменяется.
- **Тип данных**: Числовой.
- **Домен**: 1-12.
- **Правила вычисления значения**: Значение копируется из источника.
- **Источник**: Номер месяца в году из календаря.

Имя: Неделя (Week).

- **Определение**: Номер месяца в году.
- **Альтернативное имя**: Нет
- **Правила изменения**: После вставки значение этого поля никогда не изменяется.
- **Тип данных**: Числовой.
- **Домен**: 1-52.
- **Правила вычисления значения**: Значение копируется из источника.
- **Источник**: Номер недели в году из календаря.

Факты

Продажа (Sale).

Метрики: Общие издержки (To1a1 cost), Общий доход (To1a1 revenue), Общее количество проданного товара (To1a1 quantity sold) и Скидки (Discount amount).

Ответственный сотрудник: Администратор ХД.

7.5.4. Логическая структура метаданных для метрик

Логическую стуктуру метаданных для метрик дадим на примере метрик Общие издержки, Общий доход и Общее количество продаж. Она может быть следующей.

Имя: Общие издержки (Total Cost).

Определение: Это есть стоимость всех компонентов, используемых для создания данного вида (модели) продукции, которая была продана.

Альтернативное имя: Нет.

Тип данных: Числовой.

Домен: 0.01-9999999.99.

Правила вычисления значения: Общие издержки равны произведению стоимости единицы товара (модели) на количество проданных моделей.

Статистика использования:

- Среднее число запросов в день.
- Максимальное число запросов в день.

Качество данных: Эта метрика формируется только исходя из стоимости комплектующих деталей на момент продажи данного вида товара. Никакие другие виды издержек на производство товара не учитываются.

Точность данных: Предполагается, что разброс значений в стоимости комплектующих деталей данного вида товара составляет +/- 5 %.

Факты: Продажа (Sale).

Измерения: Покупатель (Customer), Производитель (Manufacture), Продукт (Product), Продавец (Seller) и Время (Time).

Имя: Общий доход (Total Revenue).

Определение: Общий доход равен произведению проданных единиц товара на отпускную цену этого товара на момент продажи.

Тип данных: Числовой.

Домен: 0.01-999999999.

Правила вычисления значения: Общий доход есть произведение отпускной цены модели товара на количество проданных моделей товара.

Статистика использования:

- Среднее число запросов в день.
- Максимальное число запросов в день.

Качество данных: Эта метрика представляет количество проданных моделей товара.

Точность данных: С точки зрения построения трендов продаж и шаблонов поведения покупателей высокая точность данных не требуется.

Факты: Продажа (Sale).

Измерения: Покупатель (Customer), Производитель (Manufacture), Продукт (Product), Продавец (Seller) и Время (Time).

Имя: Общее количество продаж (Total Quantity Sold).

Определение: Это есть число проданных единиц моделей товара.

Тип данных: Числовой. Домен: 1- 9999999.

Правила вычисления значения: Это значение берется непосредственно из графы количество для каждой позиции счета.

Статистика использования:

- Среднее число запросов в день.
- Максимальное число запросов в день.

Качество данных: Это поле представляет только количество проданного товара.

Точность данных: С точки зрения построения трендов продаж и шаблонов поведения покупателей высокая точность данных не требуется.

Факты: Продажа (Sale).

Измерения: Покупатель (Customer), Производитель (Manufacture), Продукт (Product), Продавец (Seller) и Время (Time).

Логическая структура метаданных источников

Логическая структура метаданных источников может быть следующей (на примере описания таблиц Счет из подающей системы).

Имя таблицы: Счет (Order).

Метод извлечения данных: В исходной таблице выбираются записи с законченными на текущую дату операциями для добавления в ХД.

График извлечения данных: Ежедневно по завершении рабочего дня.

Статистика извлечения данных:

- Последняя дата экстрагирования.
- Число строк.

Изучив приведенные примеры разработайте логическую структуру метаданных разработанного вами хранилища данных.

Методические указания для организации самостоятельной работы

Цели самостоятельной работы

Целью самостоятельной работы является получение знаний о новых направлениях в развитии информационных технологий хранилищ данных и применении их в сфере государственного и муниципального управления.

В рамках самостоятельной работы студентам предлагается более глубоко рассмотреть ряд вопросов. Это направлено на расширение кругозора и уяснение роли информатизации в управленческой деятельности. Студент либо сам предлагает для изучения интересующий его вопрос либо ему дается конкретная тема, по которой он представляет рефераты и (по желанию студента) можно выступить с докладом на практических занятиях. Основой данного вида работ является сбор материала, анализ собранного материала, консультации с преподавателем, сообщение по итогам изучения материала.

В течение данного курса самостоятельная работа предусмотрена в виде ряда форм, в числе которых:

1. Проработка лекционного материала;
2. Подготовка к лабораторным работам;
3. Изучение тем (вопросов) теоретической части курса, отводимых на самостоятельную проработку;

Список тем, вынесенных на самостоятельную проработку, приводится ниже.

1. Использование Microsoft Excel в качестве OLAP-клиента.
2. Модели Data Mining. Предсказательные и описательные модели.
3. OLAP и Web-технологии
4. Специальные средства очистки и инструменты ETL
5. Особенности моделей «звезда» и «снежинка».
6. Реляционная модель хранилища данных.

7. Современный взгляд на взаимное соотношение концепции ХД и концепций анализа данных
8. Современный рынок средств создания Хранилищ данных
9. «Виртуальное» хранилище данных
10. Отличия в характере данных и в требованиях к средствам реализации оперативных и аналитических систем.

При подготовке к лабораторным вопросам особое внимание рекомендуется уделить следующим вопросам.

1 Проектирование структуры и функционального наполнения OLTP систем

Создание структуры OLTP системы, необходимой для поддержки принятия решений в процессе управления одним из подразделений, решающим задачи ведения недвижимого имущества (земельный комитет, бюро технической инвентаризации, учреждение регистрации прав на недвижимое имущество, риэлтерская контора, комитет по налогам и сборам). Используя данные сети Интернет изучите особенности функционирования данных организаций и присущие им функциональные особенности.

.

2 Проектирование структуры хранилища данных

Для повышения эффективности разработки структуры реляционного Хранилища Данных, ориентированного на поддержку принятия решений в области управления недвижимым имуществом территории, проанализируйте все возможные источники данных и, в том числе, данные OLTP-систем, разработанных в рамках предыдущей темы. Используйте данные сети Интернет.

3 Разработка комплекса метаданных хранилища данных

Для повышения эффективности разработки логической структуры метаданных для спроектированного хранилища - метаданных модели, метаданных фактов, метаданных измерений, метаданных источников

изучите, на базе данных из Интернет, имеющиеся подходы к стандартизации задания и определения метаданных хранилищ данных.

Особое внимание уделите рассмотрению примеров и предложите логическую структуру комплекса метаданных для разработанного вами на предыдущих занятиях хранилища данных

Список литературы

1. Туманов В.Е., Маклаков С.В. Проектирование реляционных хранилищ данных. – М.: Диалог-МИФИ, 2007 – 333 с.
2. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP – СПб.: БХВ-Петербург, 2008, - 376 с
3. Ехлаков Ю. П. Информационные технологии учета и регистрации недвижимости : Препринт / Ю. П. Ехлаков, О. И. Жуковский; НИИ автоматики и электромеханики при ТУСУР. - Томск: Издательство ТУСУР, 1998. -