Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра математической
кибернетики и компьютерных наук

ПРОЕКТИРОВАНИЕ И РЕАЛИЗАЦИЯ СИСТЕМЫ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ СРЕДСТВАМИ SQL SERVER

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 411 группы
направления 02.03.02 — Фундаментальная информатика и информационные технологии
факультета КНиИТ
Романова Алексея Ивановича

Научный руководитель
Старший преподаватель _____ М. И. Сафрончик

Заведующий кафедрой
к.ф.-м.н. _____ С. В. Миронов

Саратов 2016

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Различные организации по всему миру ежедневно формируют огромные потоки информации. Однако не все понимают эффективность аналитической обработки этой информации. Результаты анализа массива данных, накопленного предприятием за определенный срок, могут быть использованы в дальнейшем для принятия стратегически важных решений. С каждым годом все больше компаний приходят к выводу, что правильное использование аналитических систем и систем обработки больших массивов данных дает существенное конкурентное преимущество на рынке.

Целью настоящей работы является создание системы аналитической обработки данных, включающей все основные компоненты систем бизнес-аналитики.

В данной работе ставятся следующие задачи:

- изучить теорию аналитических систем;
- создать хранилище данных на основе базы данных "Магазин стройматериалов";
- реализовать средствами MS SQL Server 2014 OLAP-систему на основе созданного ранее хранилища данных;
- разработать несколько типовых моделей интеллектуального анализа данных;
- сравнить преимущества и недостатки различных подходов к созданию структуры интеллектуального анализа данных — на основе OLAP-куба и на основе хранилища;
- создать OLAP-отчет на основе куба.

Программные средства, используемые в работе: Microsoft SQL Server 2014 Enterprice Edition, Microsoft SQL Server Data Tools, Microsoft Visual Studio 2012 — оболочка интегрированная.

1 Разработка хранилища данных на основе базы данных магазина стройматериалов

Проектируемое хранилище данных создается на основе уже существующей базы данных оперативного доступа в среде Microsoft SQL Server 2014. Основной задачей на данном этапе является определение бизнес процессов, которые требуют углубленного анализа, определение ключевых показателей, по которым будут анализироваться эти бизнес процессы, а так же избавление от лишних связей между сущностями базы данных, которые в первоначальном своем варианте были высоконормализированными. Как уже было описано ранее, высокая степень нормализации базы данных приводит к существенному ухудшению производительности на этапе многомерной обработки данных. [1]

Разработанное хранилище данных имеет структуру снежинки. В качестве анализируемых бизнес процессов были выбраны 2 сферы деятельности предприятия: покупки и поставки. Для представления этих объектов в хранилище данных были созданы таблицы: `факт покупки` и `факт поставки`, соответственно. Анализ фактов покупки производился по измерениям: покупатель, способ оплаты, офисы продаж, товар, даты покупок, продавцы, доставка, валюта, производитель. Анализ фактов поставки производился по измерениям: поставщик, товар, даты поставок, валюта, производитель, способ оплаты. Для представления описанных измерений в хранилище были созданы соответствующие таблицы. Таблицы фактов соединены с таблицами измерений связью один ко многим.

1.1 Интеграция данных из различных источников в хранилище при помощи средств SQL Server Integration Services

Для работы с данным инструментом предварительно была установлена среда разработки Microsoft Visual Studio 2012 и Microsoft SQL Server 2014. Так же дополнительно был установлен компонент SQL Server Data Tools Business Intelligence.

Проект служб SQL Server Integration Services (далее-SSIS) создается в среде Visual Studio. Поскольку в хранилище данных регулярно будет поступать новая информация, для того чтобы в таблицах не возникали дубликаты данных, все таблицы перед выполнением очередного пакета полностью очищаются. Чтобы очистить таблицу, в потоке управления необходимо создать

здачу Выполнение SQL. Далее необходимо перейти к редактированию задачи. В свойстве Connection необходимо выбрать заранее настроенное подключение к базе данных, из которого будет удаляться таблица. В нашем случае это база Хранилище, которая располагается на сервере RER-BULL. После этого в свойстве SQLStatement необходимо прописать скрипт:

```
delete from [название таблицы]
```

Для обеспечения пересылки данных необходимо в потоке управления создать задачу потока данных. В окне редактирования этой задачи необходимо указать источники данных и назначение. После этого необходимо настроить сначала источник данных. Разрабатываемое хранилище использует данные из различных типов источников. Туда входит база данных OLTP, файл Exel, файл txt, файл XML. Для доступа к ним выбираются соответствующие типы источников данных. В окне редактирования источника необходимо выбрать требуемое соединение и таблицу источник. Далее нужно аналогичным образом настроить назначение данных и связать связью источник с назначением.

Для того, чтобы выделить более детализированную информацию из некоторых ячеек, к примеру, если необходимо из полного адреса выделить в отдельности город, район и улицу, то необходимо между источником и назначением вставить объект производный столбец.

Таким образом, необходимо организовать согласованную пересылку данных: сначала необходимо заполнить родительские таблицы, только после этого можно приступать к заполнению дочерних. Так же, для того чтобы некоторые ячейки таблиц ссылались на соответствующие им даты и адреса, сначала необходимо эти данные собрать в хранилище, после чего необходимо их переслать обратно в базу, откуда уже нужно будет доставать образованные даты и адреса, в паре с соответствующими им строками других таблиц, т.к. изначально база не имеет отдельно выделенных таблиц Дата и География. Они формируются на основе уже имеющейся в базе информации на этапе выполнения пакета интеграции данных из базы в хранилище. Сформировав окончательную структуру пакета пересылки данных можно приступать к выполнению этого пакета.

2 Создание многомерной базы данных, на основе хранилища

Для того чтобы создать многомерную структуру данных, именуемую OLAP-кубом, необходимо наличие в источнике данных таблиц фактов связанных с таблицами измерений, а так же в таблице фактов должны иметься атрибуты мер. Значения этих мер будут высчитываться в различных разрезах, при построении аналитических отчетов.

Создав проект Analysis Servises в среде Visual Studio 2012, необходимо сначала связать его с источником данных. В данном случае это база Хранилище, которая содержится на сервере RED-BULL. Далее необходимо создать представление источника данных. Оно формируется на основе хранилища и позволяет выбрать только интересующие нас компоненты хранилища. После создания представления источника данных можно приступать к созданию кубов.

В настоящей работе был создан куб:
Аналитика покупок и поставок. Структуру можно посмотреть в приложении. При создании каждого из кубов необходимо указать таблицу фактов, которая будет являться основой для соответствующего куба. Так же мастер создания кубов предложит выбрать меры и измерения. После создания куба автоматически создаются измерения. Остается только правильно их настроить. В тех измерениях, которые не предполагают содержания в себе никаких иерархий, необходимо только добавление интересующих атрибутов в окне редактирования измерения. Измерения Даты покупок и Даты поставок содержат две иерархии, представленные на рисунке 1.

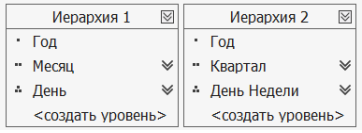

Рисунок 1 – Иерархии измерений Даты покупок и Даты поставок

После создания иерархий необходимо настроить связи атрибутов. Иллюстрация представлена на рисунке 2.

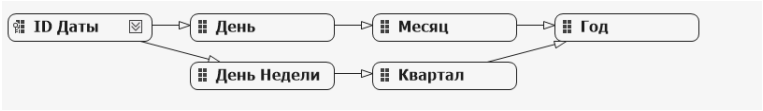

Рисунок 2 – Связи атрибутов иерархий измерений Даты покупок и Даты поставок

Чтобы уникально идентифицировать каждый из атрибутов, кроме ключевого. Необходимо настроить свойства `KeyColumn` и `NameColumn`. К примеру, для атрибута `День`, в качестве ключевых атрибутов необходимо выбрать коллекцию из атрибутов `День` и `ID_даты`, после этого необходимо привязать эту коллекцию к столбцу `День`.

Помимо календарной иерархии, аналогичным образом была создана географическая иерархия (см. рисунок 3), а так же, каждый конкретный товар был отнесен к конкретному классу, в зависимости от категории товара (см. рисунок 4).

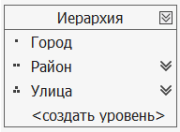

Рисунок 3 – Иерархия по географическому признаку

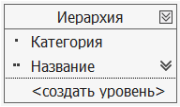

Рисунок 4 – Иерархия товаров

Аналогичным образом необходимо настроить иерархии в дочерних измерениях, которые содержат в качестве родителей таблицы, предполагающие создание иерархий. После создания иерархии в измерении, необходимо выполнить обработку этого измерения.

После настройки всех измерений необходимо выполнить обработку и развертывание всех кубов. Если она была выполнена успешно, то можно приступать к созданию структур интеллектуального анализа данных, на основе созданного куба.

3 Интеллектуальный анализ данных средствами SQL Server Analysis Services

Интеллектуальный анализ данных в SSAS подразумевает использование двух компонент:

- структуры интеллектуального анализа данных;
- модели интеллектуального анализа данных.

Структура выступает для модели в роли базы для построения. При создании структуры DM данные подготавливаются к анализу и представляются в специальной форме в виде метаданных. После обработки структуры, данные можно использовать для анализа. Структуры интеллектуального анализа данных можно создавать как на основе хранилища, так и на основе OLAP-куба. Модель же, в отличии от структуры, представляет собой алгоритм или каскад различных алгоритмов решения задач DM. В используемом программном решении имелись в распоряжении следующие методы интеллектуального анализа данных:

- алгоритм временных рядов (Microsoft);
- алгоритм дерева принятия решений (Microsoft);
- алгоритм кластеризации (Microsoft);
- алгоритм кластеризации последовательностей (Microsoft);
- алгоритм линейной регрессии (Microsoft);
- алгоритм логистической регрессии (Microsoft);
- алгоритм нейронной сети (Microsoft);
- правила взаимосвязей (Microsoft);
- упрощенный алгоритм Байеса (Microsoft).

Средства SSAS предоставляют разработчику удобный мастер создания структур интеллектуального анализа данных с подробным описанием каждого шага. Создание структуры начинается с выбора источника для построения: это либо база данных OLTP или хранилище, либо OLAP-куб. При выборе основы следует руководствоваться следующими факторами: имеется-ли куб, позволяющий создать требуемую для анализа структуру, насколько велики объемы анализируемых данных. Если объемы небольшие, то можно для анализа использовать реляционную базу или хранилище. Конструктор, основанный на данном методе достаточно гибок и прост в использовании. Если же объем данных велик, предпочтительнее использовать куб, т.к. при обработке куба

выполняется значительная часть вычислений, и данные в нем представлены в готовом формате, соответственно транзакции для такого набора будут выполняться значительно быстрее и использовать при этом меньше вычислительных ресурсов. Данный конструктор структуры интеллектуального анализа данных так же позволяет использование более сложных и разнообразных конструкций таблиц, с возможностью анализа иерархий.

Следующий шаг требует определения выбора метода интеллектуального анализа данных, который будет использоваться для решения поставленной задачи. Все эти методы описаны выше, и выбор того или иного метода зависит от типа самой задачи, ее особенностей, а так же от предпочтений аналитика. Так, например, одну и ту же задачу можно решить с использованием различных методов.

Третий шаг для обоих методов будет отличаться. При создании структуры на основе реляционной базы или хранилища потребуется выбрать соответствующее представление источника данных. Оно может представлять собой либо хранилище данных, либо базу OLTP, либо выборка таблиц из этих баз. Если на первом шаге нами был выбран способ создания структуры на основе куба, то на третьем шаге необходимо будет выбрать соответствующее измерение куба, по которому будет производиться анализ.

На 4 шаге для 1 способа необходимо выбрать таблицу вариантов и вложенную таблицу. Вложенная таблица должна ссылаться на таблицу вариантов. Для 2 способа на данном шаге требуется выбрать ключевой атрибут для используемого измерения.

На следующем шаге для 1 способа требуется указать используемые при обучении столбцы, а так же выбрать ключевой столбец для вложенной таблицы. Для 2 способа необходимо указать столбцы уровня вариантов структуры интеллектуального анализа данных. Туда входят атрибуты выбранного измерения, а так же вычисляемые меры таблицы фактов.

После выбора используемых при анализе столбцов мастер предлагает выбрать типы данных и типы содержимого для выбранных столбцов. В типе содержимого указывается какую роль будет играть выбранный атрибут в созданной структуре данных. Для ключевых столбцов это может быть просто ключ, ключ последовательности или временной ключ. Для остальных атрибутов содержимым могут быть дискретные значения, непрерывные, либо

значения, которые необходимо дискретизировать. В альтернативном подходе, на данном шаге мастер предлагает выбрать прогнозируемые атрибуты, либо добавление вложенной таблицы. Вложенная таблица используется для расширения представляемой информации таблицей фактов, соответственно, на нее таблица фактов ссылается. Мастер при создании вложенной таблицы предлагает выбрать интересующее аналитика измерение, ключевой атрибут выбранного измерения, а так же используемые в анализе атрибуты этого измерения, для которых нужно указать тип: либо это входной атрибут, либо прогнозируемый. Следующий шаг данного метода аналогичен предыдущему в альтернативном методе. Так же после выбора типа данных и содержимого для используемых атрибутов, мастер предлагает указать критерии фильтрации значений, т.е. необходимо указать срез исходного куба.

Заключительные шаги аналогичны для обоих методов создания структуры интеллектуального анализа данных: определяются процент проверочных данных (не для всех методов), с заданным по умолчанию значению 30, а так же максимальное количество вариантов в наборе проверочных данных. После задания этих значений входной набор данных разбивается на 2 подмножества — обучающий и проверочный. Обучающий набор будет задавать параметры алгоритмов, а проверочный, соответственно проверять точность прогнозов. Если заданы оба значения, то применяться будут оба ограничения. На завершающем шаге мастер предложит указать имя структуры и модели интеллектуального анализа данных, разрешить детализацию данных, создать измерение модели интеллектуального анализа данных и куб с использованием измерения этой модели. Если на втором шаге работы мастера был выбран пункт — создание структуры интеллектуального анализа данных, то на данном шаге потребуется указать только имя этой структуры. После создания структуры интеллектуального анализа данных ее необходимо обработать. После этого уже можно приступать к изучению созданной модели интеллектуального анализа данных.

В настоящей работе было создано 3 структуры интеллектуального анализа данных:

- определение взаимосвязей потребительской корзины;
- кластеризация покупателей физ. лиц;
- прогноз покупок.

4 Обеспечение общего доступа к OLAP-отчету

Для того, чтобы создать отчет, необходимо сначала создать проект Reporting Services. После этого необходимо настроить соединение с источником данных, выбрав в качестве источника базу данных SSAS, хранящуюся на сервере RED-BULL. Отчеты в Reporting Services формируются при помощи запросов к подключенной базе данных. Поскольку база SSAS является многомерной, то для формирования отчета необходимо ввести MDX-запрос, либо воспользоваться конструктором запросов.

Покажем пример формирования отчета сравнения показателей продаж по каждому из офисов продаж в городе. В построителе запросов выберем атрибут Адрес измерения Офисы продаж и меры Количество и Итоговая сумма. В окне результатов запросов сразу же отобразятся подсчитанные показатели продаж по каждому из офисов (см. рисунок 5).

Адрес	Количество	Итоговая Сумма
г. Саратов, Волжский, ул. Усть-курдюмская, 23	78	19203
г. Саратов, Кировский, ул. Пензенская, 15	111	32390
г. Саратов, Фрунзенский, ул. Астраханская, 67	127	75236

Рисунок 5 – Показатели продаж по офисам

MDX-запрос будет выглядеть следующим образом:

```
SELECT NON EMPTY { [Measures].[Количество], [Measures].[Итоговая Сумма] }
ON COLUMNS, NON EMPTY { ([Офисы Продаж].[Адрес].[Адрес].ALLMEMBERS ) }
DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME
ON ROWS FROM [Аналитика покупок] CELL PROPERTIES
VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING,
FONT_NAME, FONT_SIZE, FONT_FLAGS
```

Далее выбираем тип отчета табличный. На следующем шаге нужно распределить атрибуты по группам по своему значению. Меры включаем в подробности, а атрибуты, в зависимости от уровня иерархии, относим к первому или ко второму полю. Выбрав макет таблицы, задав стиль и имя отчету, можно приступать к просмотру. Для более наглядного представления сформированных данных, рекомендуется добавить диаграммы.

ЗАКЛЮЧЕНИЕ

В данной работе были рассмотрены основные этапы построения BI-системы при помощи средств Microsoft SQL Server Data Tools:

- создание хранилища данных на основе базы данных "Магазин стройматериалов";
- реализация пакета для интеграции данных из различных источников в хранилище средствами SSIS;
- реализация OLAP-куба на основе спроектированного хранилища данных средствами SSAS;
- создание структуры и модели интеллектуального анализа данных средствами SSAS;
- сравнение двух подходов к созданию структуры интеллектуального анализа данных — на основе OLAP-куба и на представления данных, созданного на основе базы данных OLTP;
- создание аналитических отчетов средствами SSRS.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Сарка, Д.* Microsoft SQL Server 2012 Реализация хранилищ данных / Д. Сарка, М. Лах, Г. Йеркич. — Санкт-Петербург: ООО «Издательство «Русская редакция», 2014.