электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

 # 11, ноябрь 2015

УДК 004.654

Методика организации потоков данных в корпоративном хранилище данных SAP NetWeaver Business Warehouse

***Высочанский В.А.**, студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Системы обработки информации и управления»

Научный руководитель: Тоноян С.А., к.т.н, доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Системы обработки информации и управления»
chernen@bmstu.ru

Введение

Как известно, обобщённый термин «хранилище данных» подразумевает информационную систему, предназначенную для оптимизированного хранения большого объёма данных и построения аналитической отчётности для поддержки принятия решений в организации. Корпоративное хранилище данных (КХД), в свою очередь, представляет собой систему обработки и многомерного анализа оперативных, исторических и прогнозных данных предприятия на основе системы ключевых показателей эффективности (КПЭ).

Данные, обрабатываемые любым хранилищем, делятся на два класса: транзакционные и основные. К первому классу относятся записи исходной системы обработки транзакций в реальном времени (OLTP-системы), представленные в хранилище в агрегированной форме. Основные данные используются в качестве навигационных атрибутов для формирования отчётности.

В соответствии со спецификой КХД, потоки данных должны иметь более сложную структуру для выполнения функций, не заложенных в классическом хранилище данных. Таким образом, возникает необходимость в описании основополагающей методики, учитывающей специфику КХД и возможности SAP BW.

1. Архитектура классического хранилища данных на базе SAP BW

Неотъемлемой чертой любого хранилища данных является наличие средств

аналитической обработки данных в реальном времени OLAP (*on-line analytical processing*), иногда называемых средствами многомерного анализа. Данные инструменты основаны на концепции многомерной модели базы данных, позволяющей исключить недостатки использования реляционной базы данных с высокой степенью нормализации, которые задействованы в ориентированных на обработку транзакций системах OLTP (*on-line transaction processing*). Платформа SAP BW предоставляет широкий набор инструментов OLAP, в основе которых лежит идея построения OLAP-кубов (инфо-кубов в терминологии SAP).

Хранилище данных в SAP BW строится из набора инфо-кубов, каждый из которых состоит из одной таблицы фактов и нескольких таблиц измерений. Подобная структура денормализована и нередко избыточна с целью повышения скорости выполнения запросов к инфо-кубам, которое достигается отсутствием необходимости в соединении (JOIN в терминологии SQL) множества таблиц при выполнении специализированных запросов. Непосредственно на OLAP-кубах в SAP BW строятся аналитические отчёты, являющиеся конечной целью разработки хранилища данных.

Основной таблицей инфо-куба является таблица фактов, в которой отражаются некоторые события, значимые для дальнейшего анализа. Например, фактом является проведение документа закупки товаров по некоторой цене, в определённом количестве, по номеру партии, в валюте и по другим показателям в исходной OLTP-системе. Очевидно, что для описания любого факта требуется набор параметров, уникально идентифицирующих данный факт, и набор числовых значений, характеризующих его. В SAP BW параметры фактов содержатся в таблицах измерений, а числовые характеристики называются показателями и находятся непосредственно в таблице фактов. Таким образом, значение показателя соответствует уникальной комбинации из ключевых полей измерений.

Таблицы измерений в OLAP-кубе содержат данные из справочников-классификаторов, сгруппированные по предметным областям. К примеру, измерение «Географические данные» может включать справочники «Страны», «Географические сегменты» и «Части света». Стоит отметить, что справочники в одном измерении нередко составляют иерархию: справочник «Части света» является корнем иерархического дерева, а «Географические сегменты» и «Страны» представляют различные уровни иерархии. Составной ключ из идентификаторов справочников уникально характеризует одно измерение. Очевидно, что таблицы измерений состоят в отношении «один ко многим» с таблицей фактов.

Структура классического OLAP-куба, часто называемая схемой-звездой, представлена на рис. 1.

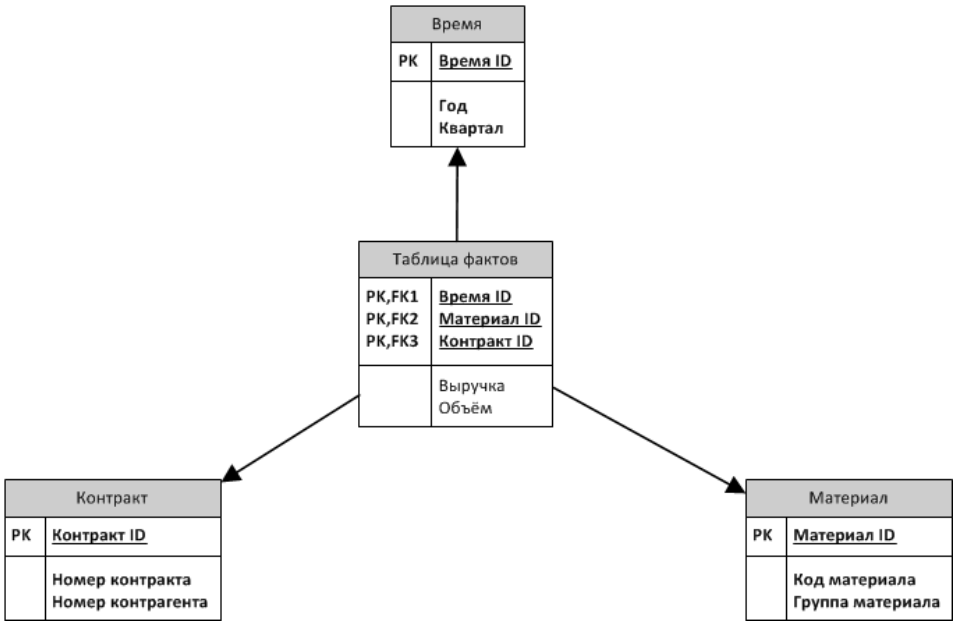

Рис. 1. Структура классического OLAP-куба

В SAP BW используется усовершенствованная схема-звезда, которая отличается от классической схемы тем, что значения справочников не хранятся в таблицах измерений. В схеме SAP для каждого справочника генерируется т.н. «суррогатный ключ» (SID), который заменяет ключ справочника в таблице измерения. Таким образом, несколько инфо-кубов могут независимо использовать одни и те же справочники в таблицах измерений.

Расширенная схема-звезда, используемая в SAP BW, представлена на рис. 2.

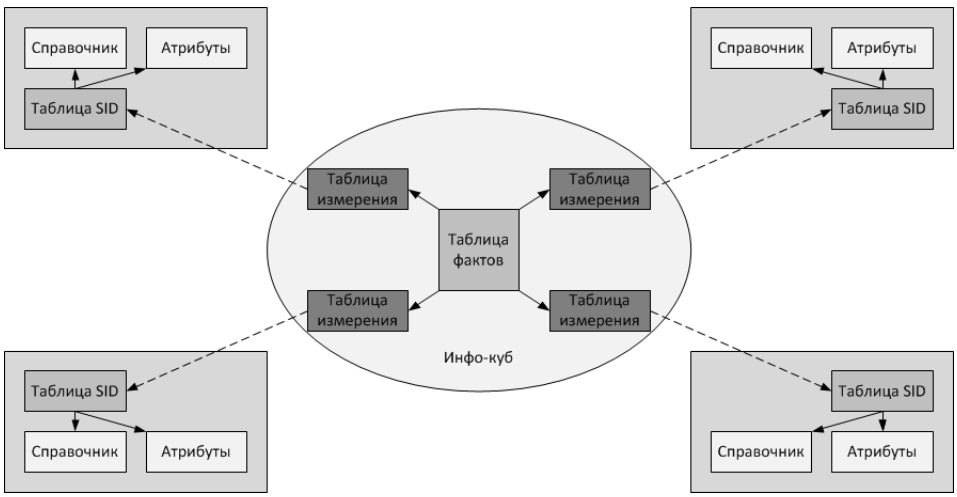

Рис. 2. Расширенная схема-звезда SAP BW

Как было отмечено ранее, структура OLAP-куба оптимизирована для запросов чтения благодаря денормализации. Однако системы, из которых данные поступают в хранилище (т.е. в инфо-кубы) очень часто построены по совершенно иным принципам: к примеру, системы ERP базируются на реляционных базах данных в третьей нормальной форме. Это означает, что между хранилищем данных и исходной системой должен существовать промежуточный уровень, осуществляющий преобразование информации без потерь и агрегирования.

Основным источником данных для любого хранилища, как было отмечено ранее, является OLTP-система, которая в семействе программных продуктов SAP представлена платформой R/3. Для создания потока данных между системами SAP R/3 и SAP BW необходимо задействовать процессы ETL (*extract*, *transform*, *load* — извлечь, преобразовать, загрузить), содержащие следующие стадии:

— извлечение данных из таблиц OLTP-системы при помощи т.н. «структур извлечения» — особых виртуальных таблиц, являющихся проекциями реально существующих таблиц базы данных;

— преобразование извлечённых данных посредством стандартных операций (перевод в другой формат) или программ, написанных на встроенном объектно-ориентированном языке ABAP;

— загрузка в хранилище данных.

В хранилищах SAP BW основным компонентом уровня сбора записей исходной системы является т.н. «источник данных», представленный таблицей PSA (*persistent staging area* — область постоянного хранения). Для загрузки данных в PSA предусмотрены следующие технологии:

— DB Connect — позволяет получить доступ непосредственно к реляционным базам данных при помощи технологии DB MultiConnect;

— UD Connect — предназначена для соединения не только с реляционными, но и с многомерными источниками данных;

— сервисный интерфейс API BI — служит для интеграции с любыми системами SAP (как ERP, так и BW);

— файл — позволяет импортировать файлы в форматах CSV и ASCII;

— веб-сервисы — служит для считывания данных в формате XML.

В итоге, классическое хранилище данных на базе SAP BW представляет собой систему инфо-кубов, загружаемых через уровень PSA из исходных систем. OLAP-отчёты строятся на базе инфо-кубов, а необходимые преобразования данных выполняются

программно при передаче записей из источников в кубы.

2. Требования к проектированию корпоративного хранилища данных

Корпоративное хранилище данных, в отличие от классического, должно содержать данные разной степени агрегирования. Это означает, что одна и та же запись, извлечённая из исходной системы, хранится в КХД на нескольких уровнях:

— изначальное представление — наиболее детализированная информация;

— промежуточное представление — данные агрегированы по нескольким признакам, например: выручка в рублях из сбытовых контрактов суммируется по странам, т.е. исключается признак «Сбытовой контракт», но остаётся признак «Страна»;

— окончательное представление — данные находятся в наиболее обобщённом виде, в котором набор измерений ограничен 5-10.

Отсюда следует, что для построения КХД недостаточно использовать систему инфо-кубов, на каждом из которых базируется OLAP-отчёт, т.к. подобная схема позволяет хранить данные только на одном уровне детализации.

Следующей особенностью КХД является наличие ключевых показателей эффективности. В отличие от показателей классического хранилища данных, повторяющих показатели исходной системы (суммы, объёмы, выручки, расходы, прибыль и т.п.), КПЭ наиболее точно отражают состояние бизнеса. Дело в том, что КПЭ представляют собой сложную систему сбалансированных показателей, каждый из которых рассчитывается на основе данных исходной системы. Обычно, ключевой показатель эффективности характеризует результат деятельности отдельно взятого отдела предприятия за определённый отчётный период времени (например, квартал). Очевидно, что концепция КПЭ подразумевает сложную систему расчёта и извлечения данных (иногда для расчёта одного КПЭ требуются данные из разных несвязанных систем), не предоставляемую классическим хранилищем.

Как правило, КХД внедряют достаточно крупные предприятия-учредители, желающие проводить анализ эффективности работы своих дочерних предприятий. Отсюда следует, что единое КХД разделяется на т.н. «витрины данных» — небольшие хранилища отдельных дочерних обществ, формирующие КПЭ в рамках своей деятельности. Часто возникает необходимость во взаимной интеграции витрин в связи с получением обобщённых КПЭ. Например, если одно дочернее общество компании занимается импортом продукции, а другое — экспортом, то кроме индивидуальных показателей эффективности может потребоваться показатель соотношения импорта/экспорта

продукции.

Обязательное наличие оптимизированной структуры хранения архивных (исторических) данных также отличает КХД от обычного хранилища. Как правило, в каждой витрине данных создаётся архивный OLAP-куб, в который передаются наиболее агрегированные записи (для уменьшения объёма хранимой информации) по закрытым отчётным периодам. Особенностью хранения исторических данных является возможность изменения значений основных данных с течением времени. К примеру, наименование контрагента за несколько лет может изменяться, поэтому хранение исторических записей в разрезе самых актуальных значений справочников является некорректным.

Таким образом, для реализации корпоративного хранилища данных требуется организовать особый подход, учитывающий особенности КХД и предполагающий его использование в рамках крупного предприятия.

3. Организация раздельных потоков основных и транзакционных данных

Как было отмечено ранее, при реализации корпоративного хранилища данных возникает необходимость хранения и организации работы с различными классами транзакционных данных:

- «сырые данные» (*raw data*), которые хранятся в КХД в формате систем-источников;
- интегрированные данные (*integrated data*), приведённые к общему формату корпоративного хранилища данных;
- бизнес-информация (*business information*) — унифицированные, обогащённые данные, которые используются приложениями отчетности и планирования в инстанции КХД.

На техническом уровне предпочтительным средством реализации задачи хранения различных классов данных в корпоративном хранилище данных является концепция многоуровневой масштабируемой архитектуры (LSA). КХД, построенное по принципам LSA, состоит из 5 уровней (слоёв):

- уровень извлечения (*data acquisition layer*);
- уровень корпоративной памяти (*corporate memory layer*);
- уровень гармонизации (*harmonization layer*);
- уровень распределения (*propagation layer*);
- уровень планирования и отчётности (*reporting layer*).

Уровень извлечения является базовым уровнем, который несёт исключительно технические функции. Данный уровень содержит наиболее актуальные данные, загруженные из систем-источников. Это позволяет отследить и проанализировать

возможные ошибки данных в промежуточной области.

Уровень корпоративной памяти формирует историческую память корпоративного хранилища данных. Это означает, что каждая единожды загруженная запись постоянно хранится на этом уровне. Если какие-либо данные будут удалены или архивированы в исходных системах, корпоративная память гарантирует, что они всегда могут быть использованы повторно в будущем.

Уровень гармонизации предусматривает необходимую очистку и соединение данных из различных источников. Данный уровень предназначен для подготовки интегрированных данных, унифицированных под общий формат и семантику корпоративного хранилища данных.

Уровень распределения получает интегрированные данные и делает их доступными для бизнес-приложений. Этот уровень обеспечивает «единый источник правды» для всех аналитических приложений предприятия.

Уровень планирования и отчётности служит для целей построения отчётности и планирования данных. При моделировании данного уровня важно гарантировать, что структуры данных, на которых строятся отчёты, оптимизированы для обеспечения высокой производительности.

Основные данные являются одной из наиболее важных областей моделирования данных. Важно отметить, что таблицы транзакционных данных могут быть в любой момент удалены или реорганизованы. В этом случае подстановки из таблиц основных данных должны быть проведены повторно. Таким образом, таблицы основных данных, которые могут меняться с течением времени, должны иметь версионность, в противном случае, результат подстановок будет отличаться от ранее полученного, и исторический ракурс будет утрачен навсегда. Версионность — зависимость от времени значения основных данных, достигаемая присвоением времени действия каждой записи справочника.

Например, в справочнике «Контрагент» запись с ключом «1» может иметь следующие значения основных данных, представленные в таблице.

Ключ записи	Наименование	Начало действия	Конец действия
1	ООО «Компьютерная помощь»	01.01.1999	31.12.2004
1	ООО «Ай Ти Сервис»	01.01.2005	31.12.9999

В КХД основные данные используются для нескольких задач. С одной стороны, они используются в качестве навигационных атрибутов при формировании отчётности. С

другой стороны, они служат основой для подстановок при обогащении потока транзакционных данных.

В рамках архитектуры LSA поток основных данных организован более простым образом, нежели поток транзакционных данных.

В случае загрузки основных данных, являющихся текстами, иерархиями и атрибутами, которые не могут меняться от времени, используется стандартный подход, в результате которого данные загружаются непосредственно в таблицы основных данных справочников уровня планирования и отчётности.

В случае загрузки основных данных, являющихся атрибутами, для которых важна версионность, данные не следует загружать непосредственно в таблицы основных данных справочников уровня планирования и отчётности. Данная рекомендация должна применяться в случае, если версионность основных данных является необходимым условием.

Описанный ниже подход позволяет сохранять историю изменения атрибутов основных данных в случае, если используемый источник данных не предоставляет данную информацию. Данный подход наиболее эффективен при реализации ежедневной загрузки основных данных. Он основан на использовании стандартной функциональности и не требует дополнительных усилий, связанных с необходимостью написания подпрограмм преобразования данных.

Общая схема потока основных данных в рамках концепции LSA представлена на рис. 3.

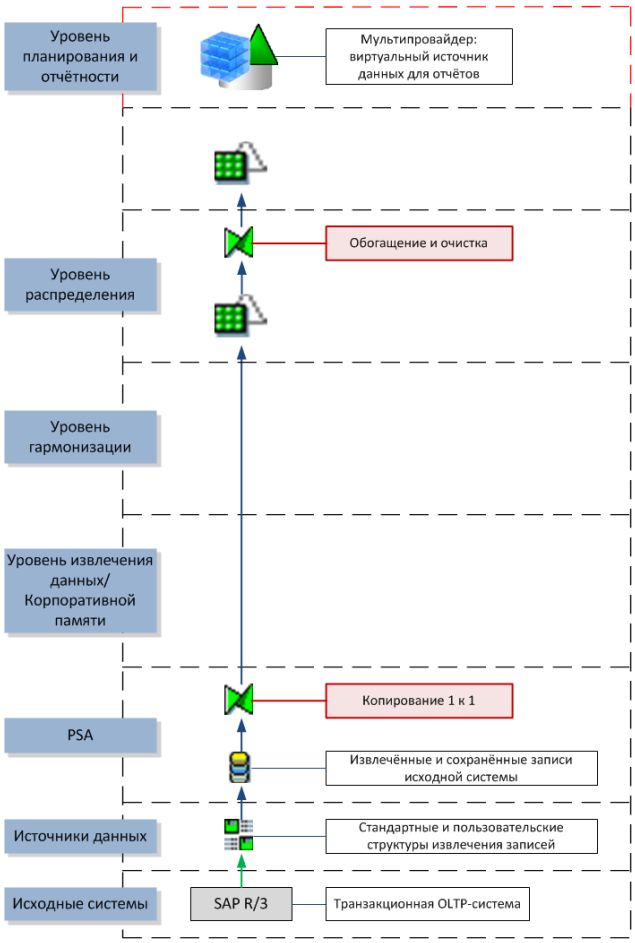

Рис. 3. Поток основных данных концепции LSA для КХД

Вместо загрузки в таблицы основных данных, все записи сначала копируются 1 к 1 и передаются в технический справочник на уровне распределения, в котором всегда содержатся поля с датами «действителен с» и «действителен до» и все поля данных, пришедшие из исходной системы. Для всех атрибутов технического признака в настройках справочника должно быть включено свойство зависимости от времени.

Поля «действителен с» и «действителен по» заполняются в трансформации данных текущей датой (с использованием формулы в трансформации) и даты 31.12.9999 (с использованием константы) соответственно. В результате загрузки актуальной версии данных в технический справочник на уровне распределения будет происходить автоматическая генерация временной зависимости данных. При обогащении потока транзакционных данных атрибутами основных данных следует использовать технический справочник уровня распределения, а не справочник уровня планирования и отчётности.

Справочник на уровне планирования и отчётности может содержать только те атрибуты, которые в настоящий момент используются для целей отчётности, и только

актуальную версию основных данных. Последнее замечание про актуальную версию справедливо при условии, если в отчётности не требуется отображать информацию на конкретную дату. Если такое требование в отчётности присутствует, то атрибуты справочника уровня планирования и отчётности должны также быть настроены как зависимые от времени. Для загрузки актуальной версии данных предлагается использовать фильтрацию в процессе переноса данных из технического справочника в справочник на уровне планирования и отчётности.

Описанный подход позволяет оптимизировать состав атрибутов глобального справочника, сохраняя при этом гибкость в расширении состава атрибутов за счёт добавления уже загруженных атрибутов с уровня распределения.

Выводы

Многие крупные предприятия, уже использующие на протяжении нескольких лет ERP-системы SAP R/3, активно внедряют корпоративные хранилища данных на базе SAP BW для эффективной поддержки принятия стратегических решений при работе с большими массивами накапливаемых данных. Тем не менее, не всегда уделяется должное внимание проектированию поток данных КХД — часто отдаётся предпочтение классической структуре хранилища данных, которую проще реализовать и быстрее внедрить.

Рассмотрена универсальная методика, позволяющая создавать потоки основных и транзакционных данных КХД любого масштаба и сложности, при этом сохраняя прозрачность структуры и возможности интеграции с другими системами.

Список литературы

1. Баллард Ч., Хэрреман Д., Шау Д., Белл Р., Ким Е., Валенчик А. Технологии моделирования данных для хранилищ данных // Режим доступа: http://www.redbooks.ibm.com/redbooks/pdfs/sg242238.pdf (дата обращения 02.09.2014).
2. Балдин А.В., Тоноян С.А., Елисеев Д.В. Анализ избыточности хранения темпоральных данных средствами реляционных СУБД // Наука и инновации. МГТУ им. Н.Э. Баумана. Электрон. журн. 2014. № 4 (24). Режим доступа: http://engjournal.ru/articles/1273/1273.pdf (дата обращения 10.04.2015)
3. Тоноян С.А., Сараев Д.В. Темпоральные модели базы данных и их свойства // Наука и инновации. МГТУ им. Н.Э. Баумана. Электрон. журн. 2014. № 12 (36). Режим доступа: http://engjournal.ru/search/word/page1.html (дата обращения 10.04.2015)

4. Patel B., Palekar A., Shilarkal S. A Practical Guide to SAP NetWeaver Business Warehouse (BW) 7.0. Braintree Massachusetts: Galileo Press, 2008. 689 p.
5. Паклин Н.Б., Орешков В.И. Бизнес аналитика: от данных к знаниям. СПб.: Питер, 2013. 704 с.
6. Барсегян А.А. Анализ данных и процессов: учебное пособие. СПб.: БХВ-Петербург, 2009. 512 с.