ДИЗАЙН ХРАНИЛИЩ ДАННЫХ. СРАВНЕНИЕ АРХИТЕКТУР ИНМОНА И КИМБАЛЛА

ПУШКАШ Яна

Universitatea Tehnică a Moldovei

Аннотация: *В данной статье определяются цели построения хранилищ данных, а так же выделяются два типа архитектур. Приведены особенности, общие черты и различия архитектур Билла Инмона и Ральфа Кимбалла. Приведены графические представления данных архитектур, а также перечислены сферы, в которых есть преимущество у каждой из архитектур.*

Ключевые слова: *business intelligence, data werehouse, Inmon, Kimball.*

1. Введение

Business intelligence (сокращённо BI) — это методы и инструменты для перевода необработанной информации в осмысленную, удобную форму. Эти данные используются для бизнес-анализа. Технологии BI обрабатывают большие объёмы неструктурированных данных, чтобы найти стратегические возможности для бизнеса.

Цель BI — интерпретировать большое количество данных, заостряя внимание лишь на ключевых факторах эффективности, моделируя исход различных вариантов действий, отслеживая результаты принятия решений.

Ключевую роль в управлении компанией в целом и ее отдельными функциями играет информация. Данные, которые доступны менеджерам и аналитикам непосредственно из корпоративных информационных систем, не унифицированы, разрозненны и в общем случае неподготовлены для анализа. Системы business intelligence - это как раз тот класс информационных систем, который позволяет превратить сырые данные в полезные для бизнеса информацию и знания, используемые в управлении, на основе которых можно принимать решения.

Хранилище данных является неотъемлемым элементом большинства корпоративных систем business intelligence. Определение хранилища данных первым выявил Б. Инмон. Он видел хранилище данных, как предметно-ориентированную, интегрированную, содержащую исторические данные, не разрушаемую совокупность данных, предназначенную для поддержки принятия управленческих решений.

Более общее определение будет звучать, как «Храни́лище да́нных (англ. *Data Warehouse*) — предметно-ориентированная информационная база данных, специально разработанная и предназначенная для подготовки отчётов и бизнес-анализа с целью поддержки принятия решений в организации. Строится на базе систем управления базами данных и систем поддержки принятия решений. Данные, поступающие в хранилище данных, как правило, доступны только для чтения»

Как правило, даже небольшие компании используют несколько информационных систем для автоматизации различных сфер деятельности. Кроме того, большинство компаний использует отдельные системы в филиалах и региональных офисах. Данные, получаемые от различных структурных элементов компании не унифицированы, часто противоречивы, и показатели, используемые для анализа и управления, не могут быть из них получены напрямую.

Использование хранилища данных, как специализированного источника для аналитической обработки информации необходимо, т.к. именно на стадии сбора и интеграции данных происходит объединение данных, их унификация и другие преобразования.

В хранилище данных, в которое в зависимости от решаемых задач, пройдя предварительную обработку, стекаются данные из самых разнообразных источников, включая корпоративные информационные системы, локальные файлы (таблицы Excel, Access), данные, предоставляемые или каким-то образом получаемые от контрагентов, данные по рынку и др.

Практически вся полезная информация в подготовленном виде находится в хранилище данных, а методы обработки и типы систем бизнес-анализа зависят от конкретных задач.

Помимо задачи интеграции и унификации данных хранилище решает вопросы:

- Повышения производительности обработки запросов и позволяет на порядки сократить время подготовки отчетов и ускорить процесс получения информации
- Хранения снимков данных, что позволяет в любой момент времени оценить ситуацию в прошлом и сравнить изменения

- Обнаружения изменения в практически статических данных (проблема медленно меняющихся размерности), что обеспечит правильное распределение показателей по категориям

2. Концепции Построения Хранилищ Данных

Инмона и Кимбалла можно назвать двумя полюсами data warehousing. Параллельное существование этих двух ортогональных по своим взглядам на предмет персонажей отличает все, что связано с хранилищами данных от остальных технологий, предназначенных для работы с данными.

Основой технологического подхода Инмона служит единая корпоративная информационная матрица (Corporate Information Factory, CIF), а Кимбалл «ставит» на шину хранилищ данных (Data Warehouse Bus).

С прагматической точки зрения, эти подходы не противоречат друг другу, более того, они имеют много общего. Предприятиям требуется хранить, анализировать и интерпретировать данные для принятия решений. Предприятиям необходимо развивать обратные связи между средствами накопления данных и системами принятия решений. Оба подхода соответствуют этим критериям. Более того, у Инмона и Кимбалла сходные взгляды на витрины данных; они сходятся в том, что наличие разрозненных витрин данных не решает проблему информированности предприятия. Но в вопросах интеграции данных они занимают противоположные позиции. Кимбалл считает возможным объединить с помощью шины отдельные витрины данных в информационную инфраструктуру, имеющую топологию *звезды*, а Инмон считает необходимым загружать все данные в *единое хранилище*.

Без использования корпоративного хранилища картина накопления и представления данных выглядит так, как показано на рис. 2.1.

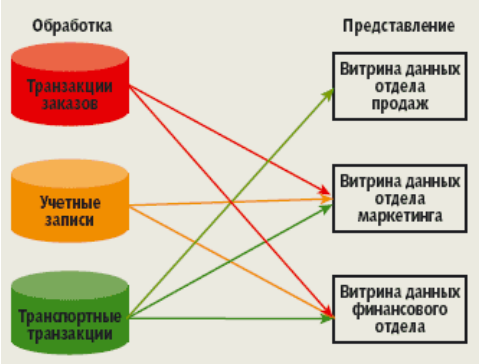

Рис. 2.1 - Схема накопления и представления данных без использования корпоративного хранилища.

Подход Кимбалла предполагает трансформацию исходных данных непосредственно на этапе обработки с соблюдением требований по производительности и качеству. Часть этих действий осуществляется централизованно, а некоторые действия остаются распределенными. Преобразование начинается с выборки требуемых данных из операционных источников. Полученная в результате параметрическая модель по своему содержанию мало отличается от нормализованной, но она лучше подготовлена для использования. Она может содержать как атомарные данные (то есть данные в их исходном виде), так и обобщенные данные, упакованные в реляционные таблицы или в многомерные кубы.

3. Корпоративная Информационная Фабрика (Corporate Information Factory)

На рис. 3.1 представлен подход, используемый в Хранилищах данных с архитектурой CIF. Когда-то этот подход был известен под названием корпоративного Хранилища данных (enterprise data warehouse, сокр. EDW). Работа такого Хранилища начинается со скоординированного извлечения данных из источников. После этого загружается реляционная база данных[1] с третьей нормальной формой[2], содержащая атомарные данные. Получившееся нормализованное Хранилище используется для того, чтобы наполнить информацией дополнительные репозитории презентационных данных, т.е. данных, подготовленных для анализа. Эти репозитории, в частности, включают специализированные Хранилища для изучения и "добычи" данных (Data Mining), а также витрины данных.

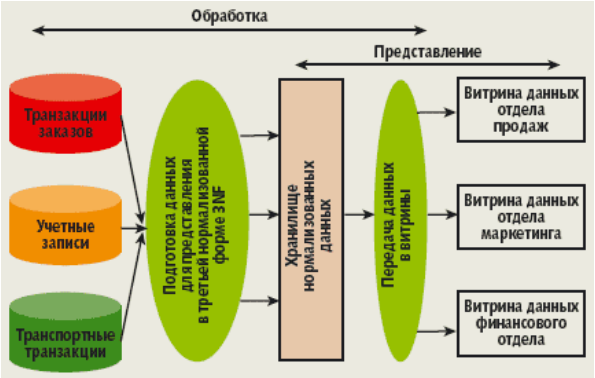

Рис. 3.1 - Нормализованное Хранилище данных с пространственными витринами итоговых данных (CIF).

При таком сценарии конечные витрины данных создаются для обслуживания бизнес-отделов или для реализации бизнес-функций и используют пространственную модель[3] для структурирования суммарных данных. Атомарные данные остаются доступными через нормализованное Хранилище данных. Очевидно, что структура атомарных и суммарных данных при таком подходе существенно различается.

В качестве отличительных характеристик подхода Билла Инмона к архитектуре Хранилищ данных можно назвать следующие пять характеристик.

1. Использование реляционной модели организации атомарных данных и пространственной - для организации суммарных данных.
2. Использование итеративного или "спирального" подхода при создании больших Хранилищ данных, т.е. "строительство" Хранилища не сразу, а по частям. Это позволяет при необходимости вносить изменения в небольшие блоки данных или программных кодов и избавляет от необходимости перепрограммировать значительные объемы данных в Хранилище. То же самое можно сказать и о потенциальных ошибках: они также будут локализованы в пределах сравнительно небольшого массива без риска испортить все Хранилище.
3. Использование третьей нормальной формы для организации атомарных данных, что обеспечивает высокую степень детальности интегрированных данных и, соответственно, предоставляет корпорациям широкие возможности для манипулирования ими и изменения формата и способа представления данных по мере необходимости.
4. Хранилище данных - это проект корпоративного масштаба, охватывающий все отделы и обслуживающий нужды всех пользователей корпорации.
5. Хранилище данных - это не механическая коллекция витрин данных, а физически целостный объект.

4. Хранилище Данных С Шинной Топологией (Data Warehouse Bus)

Рис. 4.1 представляет альтернативный подход к архитектуре Хранилищ данных, известный как Хранилище с архитектурой шины или подход Ральфа Кимболла.

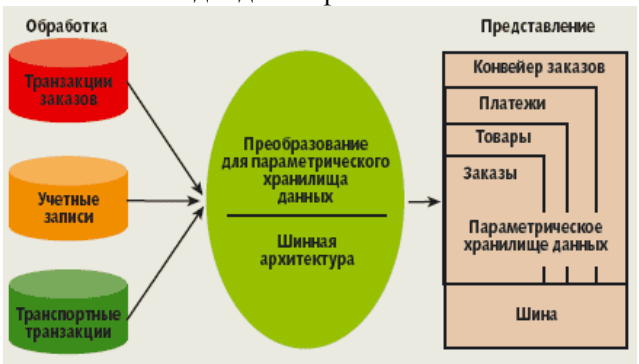

Рис. 4.1 - Пространственное Хранилище данных.

В этой модели первичные данные преобразуются в информацию, пригодную для использования, на этапе подготовки данных. При этом обязательно принимаются во внимание требования к скорости обработки информации и качеству данных. Как и в модели Билла Инмона, подготовка данных начинается со скоординированного извлечения данных из источников. Ряд операций совершается

централизованно, например, поддержание и хранение общих справочных данных, другие действия могут быть распределенными.

Область представления пространственно структурирована, при этом она может быть централизованной или распределенной. Пространственная модель Хранилища данных содержит ту же атомарную информацию, что и нормализованная модель (см. подход Билла Инмона), но информация структурирована по-другому, чтобы облегчить ее использование и выполнение запросов. Эта модель включает как атомарные данные, так и обобщающую информацию (агрегаты в связанных таблицах или многомерных кубах) в соответствии с требованиями производительности или пространственного распределения данных. Запросы в процессе выполнения обращаются к все более низкому уровню детализации без дополнительного перепрограммирования со стороны пользователей или разработчиков приложения.

В отличие от подхода Билла Инмона, пространственные модели строятся для обслуживания бизнес-процессов (которые, в свою очередь, связаны с бизнес-показателями или бизнес-событиями), а не бизнес-отделов. Например, данные о заказах, которые должны быть доступны для общекорпоративного использования, вносятся в пространственное Хранилище данных только один раз, в отличие от CIF-подхода, в котором их пришлось бы трижды копировать в витрины данных отделов маркетинга, продаж и финансов. После того, как в Хранилище появляется информация об основных бизнес-процессах, консолидированные пространственные модели могут выдавать их перекрестные характеристики. Матрица корпоративного Хранилища данных с архитектурой шины выявляет и усиливает связи между показателями бизнес-процессов (фактами) и описательными атрибутами (измерениями).

Суммируя все вышесказанное, можно отметить типичные черты подхода Ральфа Кимболла.

1. Использование пространственной модели организации данных с архитектурой "звезда" (star scheme).
2. Использование двухуровневой архитектуры, которая включает стадию подготовки данных, недоступную для конечных пользователей, и Хранилище данных с архитектурой шины как таковое. В состав последнего входят несколько витрин атомарных данных, несколько витрин агрегированных данных и персональная витрина данных, но оно не содержит одного физически целостного или централизованного Хранилища данных.
3. Хранилище данных с архитектурой шины обладает следующими характеристиками:
   – оно пространственное;
   – оно включает как данные о транзакциях, так и суммарные данные;
   – оно включает витрины данных, посвященные только одной предметной области или имеющие только одну таблицу фактов (fact table);
   – оно может содержать множество витрин данных в пределах одной базы данных.
4. Хранилище данных не является единым физическим репозиторием (в отличие от подхода Билла Инмона). Это "виртуальное" Хранилище. Это коллекция витрин данных, каждая из которых имеет архитектуру типа "звезда".

5. **Факторы Выбора Одного Из Подходов**

Как мы уже видели, подход к проектированию хранилища данных зависит от бизнес-целей организации, характера бизнеса, времени и финансовых затрат, а также уровня зависимостей между различными функциями. Метод Инмона является подходящим для стабильных предприятий, которые могут позволить себе время, необходимое для проектирования и связанных с этим, расходов. Также, при изменении бизнес-условий, они не меняют дизайн; вместо этого они приспосабливают их к существующим моделям. Однако, если локальной оптимизации достаточно, и акцент делается на быстрое достижение цели, то целесообразно использовать архитектуру Кимбалла. Учитывая это, выделим различия обоих подходов по конкретным отделам:

Страхование: Очень важно получить общую картину по отношению к отдельным клиентам, группам, истории претензий, тенденций смертности, демографии, рентабельности каждого плана и агентов и т.д. Все аспекты взаимосвязаны, и поэтому подходит для архитектуры Инмона.

Маркетинг: Это специализированное подразделение, которое не требуют предприятия склада. Требуются только витрины данных. Следовательно, предпочтительна архитектура Кимбалла.

CRM в банках: Особое внимание уделяется таким параметрам, как проданные продукты, повышение и снижение цен на уровне клиента. Это не обязательно для получения общей картины бизнеса. Например, нет необходимости связать детали клиента с казначейским департаментом, занимающимся валютными операциями и правилами. Поскольку цели ограничены, вы можете

использовать метод Кимбалла. Однако, если все процессы и подразделения в банке должны быть связаны, предпочтителен выбор дизайна Инмона.

Производство: присутствует множество функций, независимо от используемого бюджета. Таким образом, там, где присутствует системная зависимость, необходимо использовать модель предприятия. Для таких случаев метод Инмона является идеальным.

При проектировании хранилища данных, во-первых, вы должны смотреть на ваши бизнес-цели - краткосрочные и долгосрочные перспективы. Обратите внимание на связанные и отдельные функции. Проанализируйте источники данных по качеству и количеству. И, наконец, оцените уровень ваших ресурсов, временные рамки и кошелек. Это поможет вам решить, какой метод использовать Инмона или Кимбалла, а может комбинацию обоих.

6. Заключение

В качестве заключения выявим преимущества и недостатки обеих архитектур. Очевидно, вопрос о лучшем из двух подходов не имеет однозначного ответа. В целом оба этих подхода сходятся в главном - в необходимости современных средств управления информационными потоками для принятия своевременных и обоснованных решений при ведении бизнеса и, соответственно, в необходимости создания соответствующих структур для хранения данных, их координации и интеграции. Выбор того или иного технического решения определяется нуждами бизнеса и его конкретными особенностями.

Преимущества и недостатки каждого из подходов напрямую вытекают из их архитектурных решений. Считается, что пространственная организация с архитектурой "звезда" облегчает доступ к данным и требует меньше времени на выполнение запросов, а также упрощает работу с атомарными данными. С другой стороны, сторонники подхода Билла Инмона критикуют эту схему за отсутствие необходимой гибкости и уязвимость структуры, полагая, что в пространственно-организованные атомарные данные труднее вносить необходимые изменения.

Реляционная схема организации атомарных данных замедляет доступ к данным и требует больше времени для выполнения запросов в силу разной организации атомарных и суммарных данных. Но, с другой стороны, эта схема предоставляет широкие возможности для манипулирования атомарными данными и изменения их формата и способа представления по мере необходимости.

Подводя итог, можно сказать, что, несмотря на кажущиеся глубокие различия между двумя подходами к архитектуре Хранилищ данных, это различия в основном технического плана, а в целом Хранилища обоих типов оказываются достаточно похожими по своим функциям и задачам, которые можно решать с их помощью.

Литература

1. Оценка эффективности внедрения хранилищ данных - некоторые аспекты [Электронный ресурс]. – Режим доступа: http://www.bipartner.ru/resources/roi.html
2. Joerg Reinschmidt, Allison Francoise. Business Intelligence Certification Guide. IBM Red books;
3. Inmon W. Building the Data Warehouse. – New York: John Willey & Sons, 1992;
4. Спирли, Эрик. Корпоративные хранилища данных. Планирование, разработка, реализация. Том. 1: Пер. с англ. – М.: Издательский дом "Вильямс", 2001;
5. Joe Celko. Trees in SQL: Intelligent Enterprise, October 20, 2000;
6. Ralph Kimball. Slowly Changing Dimensions: DBMS April 1996;
7. Ralph Kimball: The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Wiley 1996;
8. Леонид Черняк, «Взгляд Ральфа Кимбалла на хранилища данных» [Электронный ресурс]. – Режим доступа: http://www.osp.ru/os/2007/05/4265198/
9. По материалам зарубежных сайтов, «Основные подходы к архитектуре хранилищ данных» [Электронный ресурс]. – Режим доступа: (http://www.iso.ru/rus/document6082.phtml#2)
10. Сергей Кузнецов, «Основы хранилищ данных и BI по Ральфу Кимбаллу» [Электронный ресурс]. – Режим доступа: http://citforum.ru/gazeta/47/