ХРАНИЛИЩА ДАННЫХ И ИХ ИСПОЛЬЗОВАНИЕ

Хрестоматия

Министерство образования и науки Российской Федерации

Владивостокский государственный университет
экономики и сервиса (ВГУЭС)

ХРАНИЛИЩА ДАННЫХ И ИХ ИСПОЛЬЗОВАНИЕ

Хрестоматия

Владивосток
Издательство ВГУЭС
2017

УДК 004.652, 004.89
ББК 22.17
Х90

Рецензенты: *П.Г. Рагулин*, канд. техн. наук, проф., Школа естественных наук ДВФУ;
Е.В. Кийкова, канд. техн. наук, доцент, каф. ИТС ВГУЭС

Хранилища данных и их использование: хрестоХ90 матия / сост. А.Б. Кригер ; Владивостокский государственный университет экономики и сервиса. – Владивосток : Изд-во ВГУЭС, 2017. – 120 с.

ISBN 978-5-9736-0456-1

Хрестоматия «Хранилища данных и их использование» составлена из различных публикаций, как научного так и научно-практического характера, посвященных архитектурам хранилищ данных, их разработке и использованию. В нее включены статьи из печатных изданий (журналов), материалы конференций, статьи из электронных он-лайновых журналов, материалы официальных сайтов разработчиков программного обеспечения, материалы web-форумов ИТ-специалистов. Сохранена авторская структура публикаций. Основная задача хрестоматии – познакомить обучающихся с актуальными проблемами разработки и использования систем хранения данных.

Для студентов, обучающихся по направлениям подготовки 09.03.02 «Информационные системы и технологии», 09.03.02 «Прикладная информатика», 38.03.05 «Бизнес-информатика».

УДК 004.652, 004.89
ББК 22.17

ISBN 978-5-9736-0456-1

ВВЕДЕНИЕ

Данное учебное издание составлено из различных публикаций, посвященных архитектурам хранилищ данных, их разработке и использованию.

Цель создания хрестоматии – объединение под одной обложкой научных, практических и учебных материалов, которые бы наиболее полно отвечали на вопросы:

- Что такое хранилища данных (далее ХД)?
- Чем принципиально отличаются ХД от баз данных?
- Каковы цели создания ХД, чем отличаются системы, использующие ХД?
- Теоретические основы ХД – что включает эти понятия?
- Каковы архитектурные решения хранилищ данных?

Так как отцы-основатели систем хранения данных нового типа декларировали два различных подхода к созданию ХД, до сегодняшнего дня существую различные представления об архитектурных решениях ХД. Соответственно, и авторы статей, и участники форумов, и авторы учебных материалов, часто высказывают различные точки зрения. Таким образом, основная задача хрестоматии донести до студентов позиции разработчиков на принципы организации ХД и их использование.

Состав материалов, включенных в хрестоматию, обеспечивает основные **задачи освоения дисциплины**:

– изучение принципов построения и разработки хранилищ данных;

– получение навыков проектирования и настройки витрин данных (локальных хранилищ данных);

– разработка процесса наполнения хранилища данных, реализации запросов к хранилищам данных.

Структура хрестоматии отражает направления исследовательских, проектных и учебных материалов, опубликованных в российских изданиях различных типов. Комплекс материалов логически разделен на четыре тематических модуля:

1. Архитектуры данных. Понятие архитектуры данных. Развитие систем хранения и обработки данных.
2. Многомерные данные. OLAP-технология как ключевой компонент хранилищ данных.
3. Концепция хранилищ данных (ХД).
4. Архитектуры хранилищ данных.

В приложении «Диаграммы, иллюстрации» отдельно представлены как графические материалы из статей, включенных в «Хрестоматию», так и графические материалы из других источников. Предполагается,

что студенты смогут использовать графические материалы как элементы рабочей тетради.

Источники и типы публикаций. Сфера ИТ обладает рядом особенностей:

– динамичностью – результаты разработок, исследовательские материалы часто устаревают до их публикации;

– использованием электронных ресурсов для публикаций результатов (что вполне очевидно!);

– использование телекоммуникационных ресурсов для обучения, организации конференций и семинаров.

Как следствие, в хрестоматию вошли не только печатные публикации, но и материалы электронных он-лайновых журналов, материалы официальных сайтов разработчиков программного обеспечения, материалы web-форумов ИТ-специалистов. Сохранена авторская структура публикаций.

На содержательном уровне публикации, вошедшие в «Хрестоматию», так же разнообразны. В каждом из разделов данного учебного пособия есть работы учебного характера, научные статьи, проектные разработки и проектные предложения. Материалы учебного характера, позволяют изучить основные принципы построения ХД и аналитических систем на основе ХД. Научные статьи позволяют получить представление о формализации задач построения ХД. Проектные разработки и проектные предложения являются примерами реализации ХД для различных задач.

В числе материалов, вошедших в «Хрестоматию», есть и авторская статья, в которой предложена структура (модель данных) ХД для анализа показателей учреждения культуры.

Так как в статьях не весь материал представляет интерес для обучающихся, в ряде случаев приведены сокращенные варианты публикаций – это указано в подзаголовке.

1. АРХИТЕКТУРЫ ДАННЫХ. ПОНЯТИЕ АРХИТЕКТУРЫ ДАННЫХ. РАЗВИТИЕ СИСТЕМ ХРАНЕНИЯ И ОБРАБОТКИ ДАННЫХ

А. Левичев

POWER BL OT MICROSOFT: СЕРВИС БИЗНЕС-АНАЛИТИКИ ДЛЯ КОМПАНИЙ

21.07.2016
Режим доступа: http://computerologia.ru/power-bl-ot-microsoft-servis-biznes-analitiki-dlya-kompanij/
(статья приводится в сокращении)

BI (business intelligence) – сравнительно новомодный термин. Его легко перепутать с бизнес-аналитикой, но между ними есть разница. Если бизнес-аналитика просто применяет статистические инструменты, то business intelligence участвует также и в сборе данных, отыскивая необходимую ей информацию. Всю процедуру можно разбить на пять этапов: поиск полезных данных, их аналитическая обработка, "выверка" показателей на отклонения, бизнес-аналитика и, наконец, отчетность – для удобного представления полученного результата. BI – более обширный и сложный инструмент, зато и итоговая информация может оказаться полезнее.

Технология может использоваться для информирования руководителей компании о прогрессе в достижении ключевых бизнес-целей, составления бенчмарков, ведения аналитики. Она позволяет проводить статистический анализ, прогнозное моделирование, обрабатывать сложные события. Некоторые системы BI ставят на первый план введение понятной и удобной инфраструктуры для корпоративной отчетности – чтобы привести данные в понятный вид и таким образом упростить ведение стратегического управления бизнесом. Сюда входит визуализация информации, технологии OLAP и так далее. Наконец, business intelligence способствует обмену опытом, идеями и мнениями между участниками бизнес-процессов, так что решения принимаются более взвешенно, с учетом всех переменных и с предельно достоверной информацией. Даже если ключевое слово стоит за одним человеком, он может посмотреть на наглядные данные и быть в курсе всех дел.

Одним из ключевых игроков, активно пропагандирующих идею business intelligence в своих проектах, является Microsoft. Для такого гиганта правильная работа с поступающей информацией – важнейший фактор ведения своей активности. Так Microsoft оценивает степень ус-

пешности своих текущих проектов, ищет пути их развития и сравнивает потенциал новых рынков, на которые она может выйти со своими технологиями. Как говорит сама фирма, даже на ИТ-презентациях среди глав корпорации лучше «превратить безликие цифры в информативные, интересные материалы, сделать их понятными и живыми». Тем более что с каждым годом в мозг среднего человека поступает все больше информации, и, если она не является правильно обработанной и структурированной, многие предпочитают ее просто не замечать.

Для решения собственных вопросов, а также для помощи средним и мелким компаниям в 2015-м Microsoft презентовала новый сервис Power BI. Это инструмент для визуализации крупных массивов данных, позволяющий создавать красочные, яркие и понятные отчеты, собирая информацию из большого числа источников. Сервис, как и полагается системе business intelligence, обладает возможностями по сбору и упорядочиванию данных, которые затем представляются в приятном виде и позволяют руководителям бизнеса принять правильное, взвешенное решение.

Продукт разбит на два модуля: один – для получения информации, второй – для ее удобного представления. Причем над вторым модулем, панелью мониторинга, была проведена огромная работа. Собственно, только ради него и стоит пользоваться сервисом. Если данные уже есть у вас в Excel или в Google Analytics, значит, вы примерно понимаете, как и где их собирать. А вот понятная, наглядная визуализация – другой вопрос. Microsoft Power BI позволяет сразу, в одном месте, глянуть на информацию, требующуюся для принятия решений, здесь можно отслеживать самые важные сведения о бизнес-активности вашей компании, тут же, при необходимости, ведется наблюдение за состоянием продукта, маркетинговой компании, отдела или всего предприятия. А если у вас есть особые запросы, можно создать индивидуальную панель мониторинга и поставить туда именно те метрики, которые вас интересуют.

Продукт удобный, простой и удивительно мощный. Здесь можно работать с огромным количеством форматов и сервисов. В Power BI легко загружаются данные из Excel-файлов, SQL-сервера, SharePoint, Oracle, GitHub, Twilio, Marketo, SendGrid, Zendesk, Google Analytics, MySQL, Oracle и так далее. С чем бы фирма ни работала, велика вероятность, что в Power BI найдется инструмент для обработки собранных ею данных. Сервис является безумно полезным для менеджеров, владельцев фирм и индивидуальных предпринимателей. И, что приятно, пользоваться им можно совершенно бесплатно, если не выходить за установленные лимиты.

УДК 629.7

Э.Э. Акимкина, А.Э. Аббасов

АНАЛИЗ ИНСТРУМЕНТАЛЬНЫХ СРЕДСТВ ИНФОРМАЦИОННЫХ СИСТЕМ ДЛЯ ОБРАБОТКИ МНОГОМЕРНЫХ ДАННЫХ

Информационно-технологический вестник. 2016. Т 2. С. 61–75

Проведен анализ информационной инфраструктуры предприятия; описаны критерии сравнения систем бизнес-анализа и дан краткий обзор функциональных возможностей BI-систем от вендоров (SAP, Oracle, IBM и Microsoft); даны рекомендации по выбору инструментальных средств для объединения данных из различных источников информации и проектированию системы оценки ключевых показателей эффективности с использованием инструментов системы бизнес-анализа; предложены организационно-технические мероприятия для определения максимальной производительности многомерных хранилищ данных на основе многомерных кубов.

Многомерные хранилища данных, OLAP-куб, среда моделирования

Введение

На сегодняшний день системы класса BI (от англ. Business Intelligence, системы бизнес аналитики) один из самых популярных инструментов для оценки эффективности обслуживания своих клиентов и поддержки принятия решений. Для этого им необходимо анализировать деятельность своих сотрудников, особенно тех, кто взаимодействует с клиентами напрямую, например, менеджеры продаж, сервисные инженеры. Эффективность работы сотрудника оценивается по определенным показателям бизнеса (сервиса), в соответствии с развитием и формированием комплекса интеграционных стратегий при управлении корпорацией [1]. Чтобы провести системный анализ, необходима информация о том, какой показатель и на сколько процентов выполнен данным сотрудником, а также применение новейших информационных технологий [2-5], позволяющих оперативно оценить множество параметров (данные, модели, функции, дислокация, персонал, временные параметры, цели в представлении, связи между элементами [6]).

Данные о работе сотрудников хранятся в различных информационных источниках организации (оперативные системы, документы офисных приложений и т.д.). Это связано с тем, что области деятельности автоматизируется раздельно как по историческим причинам, так и по

исключительно функциональным возможностям корпоративных систем. Для получения полной картины деятельности сотрудников необходимо объединить данные из различных источников, в которых содержится полезная для целей анализа информация. Но данные, предоставляемые различными источниками, не унифицированы, и показатели эффективности деятельности сотрудников, необходимые для оценки, не могут быть легко получены. В таком случае следует говорить о наличии сырых данных и отсутствии в них полезной для бизнеса информации, и тем более знаний (знания – ключевая компетенция бизнеса).

Методика выбора инструментальных средств информационных систем для анализа многомерных данных

Сначала необходимо проанализировать информационную структуру предприятия [7-12], определить иерархическую структуру данных и выяснить, где необходимо повысить уровень качества и информативности данных, поступающих из источников, до приемлемого, а также организовать их интегрированное хранение в структурах, обеспечивающих их целостность, непротиворечивость, высокую скорость и гибкость выполнения аналитических запросов.

Инструменты Business Intelligence преобразуют сырые данные в полезные информацию и знания, на основе которых можно решать любые задачи по анализу, управлению, прогнозированию и др.

Компоненты BI-системы представлены на рис. 1.

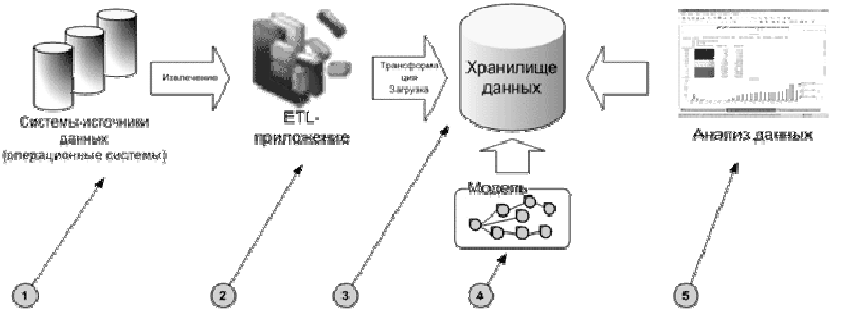


Рис. 1. Компоненты BI-системы

С использованием инструментов BI-системы проектируется KPI-система (от англ. Key Performance Indicators – ключевые показатели эффективности). Система представлена совокупностью аналитических отчетов с выбранными KPI-показателями. Отчеты предназначены руководителям для мониторинга и контроля бизнес-процесса продаж (предоставляемых услуг), а также для выстраивания на основе данных отчетов грамотной политики мотивации сотрудников.

Начальный этап реализации любой аналитической задачи или проекта – консолидация данных, предназначенная для сбора и организации хранения данных в виде, оптимальном с точки зрения их обработки на конкретной аналитической платформе или решения конкретной аналитической задачи, а также для оценки качества данных. При необходимости, данные могут быть преобразованы (очистка данных и обогащение) в соответствии со структурной схемой консолидации данных, приведенной на рис. 2.

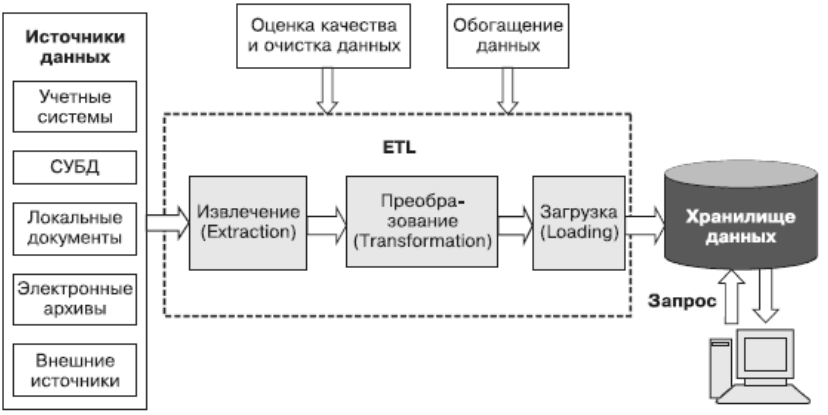


Рис. 2. Процесс консолидации данных

Как показано на рис. 2, данные извлекаются из разнотипных источников (учетные системы, системы управления базами данных (СУБД), локальные документы, электронные архивы, внешние источники). Затем данные преобразуются к виду, пригодному для хранения в определенной структуре. В соответствии с заданным регламентом преобразованные данные загружаются в соответствующую базу или хранилище данных (ХД). На рисунке 3 представлена концептуальная схема ХД.

Данные, хранимые в ХД, представлены в соответствии с заданной моделью. ХД также включает метаданные и специализированные локальные тематические хранилища, подключаемые к централизованному ХД – витрины данных – для обслуживания отдельных подразделений организации или бизнес-процесса.

Для выполнения сложных нерегламентированных запросов целесообразно использовать многомерные хранилища данных (МХД) [5]. МХД – это упорядоченные многомерные массивы, или OLAP-кубы (от англ. On-Line Analytical Processing – оперативная аналитическая обработка). Для реализации OLAP-куба используются универсальные реляционные СУБД или специализированное ПО.

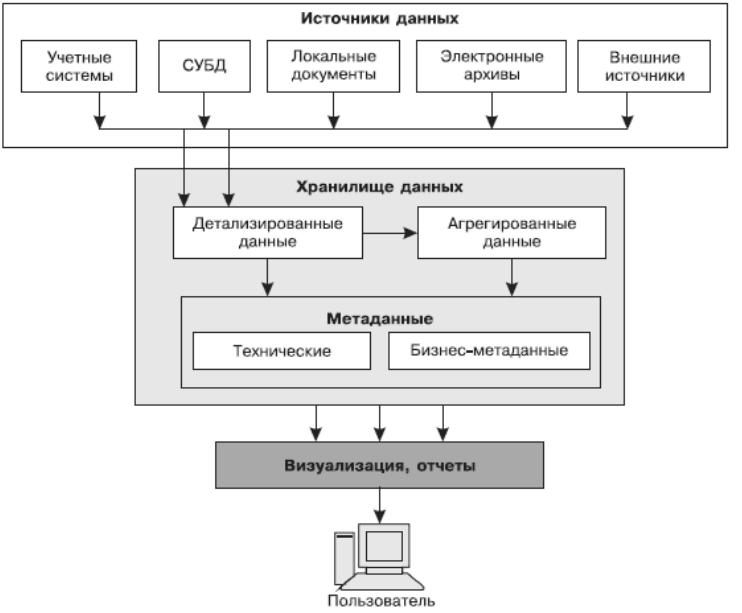
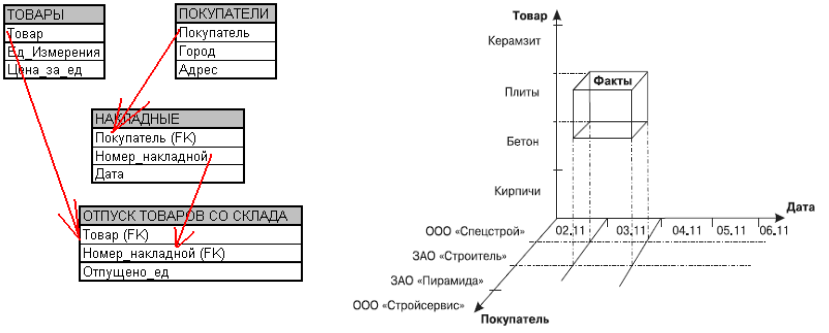


Рис. 3. Концептуальная схема ХД

Сравним совокупность нормализованных таблиц реляционной модели и представление данных в виде многомерных кубов (рис. 4 *а* и *б*). Трехмерное измерение позволяет создавать более наглядные отчеты.



а) данные нормализованной таблицы БД

б) измерения и факты в многомерном кубе

Рис. 4. Представление данных в различных измерениях

После проведенного анализа сформулируем методику выбора инструментальных средств информационных систем для анализа многомерных данных.

1 шаг. Анализ информационной инфраструктуры предприятия (иерархическая структура данных, уровень качества и информативности данных, поступающих из источников, до приемлемого качества).

2 шаг. Выбор системы бизнес-интеллекта. Объединение данных из различных источников информации.

3 шаг. Проектирование KPI-системы с использованием инструментов BI-системы.

4 шаг. Выбор аналитических методов и алгоритмов, подготовка исходных данных для анализа.

5 шаг. Консолидация данных и загрузка их в хранилище данных.

6 шаг. Определение целесообразности многомерных хранилищ данных на основе многомерных кубов.

7 шаг. Построение аналитических отчетов для оценки ключевых показателей эффективности.

Данные аналитического исследования критериев сравнения BI-систем в соответствии с их функциональными возможностями, полученные компанией Gartner за 2014 год, приведены на рис. 5 для BI-систем от вендоров (SAP, Oracle, IBM и Microsoft).

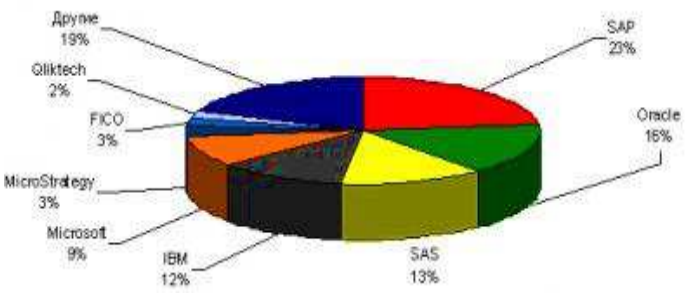


Рис. 5. Лидирующие вендоры на мировом рынке BI по объему выручки, 2014

После сравнения и выбора BI-платформы осуществляется построение корпоративной системы отчетности для реализации KPI-системы.

С помощью BI-системы пользователи и программисты могут разрабатывать отчеты и визуализировать информацию среди множества предусмотренных возможностей и настроек системы.

Существует множество различных критериев сравнения BI-систем. Перечислим наиболее важные из них с точки зрения пользователя.

Возможность создания отчётов пользователями, не имеющими навык программирования самостоятельно, используя стандартные средства.

Возможность построения преднастроенных и не регламентированных отчетов.

Возможность представления данных в виде разнообразных таблиц и диаграмм (комбинирование отчетов и графиков). Возможность использовать различные типы графиков в отчетах.

Возможность настройки фильтров и запроса параметров в отчетах.

Возможность работы с шаблонами отчетов: сохранения построенного шаблона отчета и возможность его передачи другому пользователю, многостраничные шаблоны.

Возможность централизованной рассылки (публикации) отчетов в соответствии с правами доступа к данным отчета.

Возможность создания панелей показателей.

Возможности настройки прав пользователей: создание/импорт классов и пользователей, доступ к отдельным элементам системы отчетности, различные уровни администрирования системы.

Основными архитектурными компонентами системы являются: Oracle BI Server, Oracle BI Web и Oracle Delivers Server.

Oracle BI Server централизованно хранит метаданные об источниках данных и бизнес-объектах (business definitions) в своем репозитории, доступном всем инструментам платформы Oracle BI EE.

Oracle BI Web предоставляет интерфейсы для всех компонентов системы, используемых для визуализации данных. Он взаимодействует с Oracle BI Server и выполняет ряд важнейших функций: отвечает за авторизацию пользователей и персонализацию интерфейса для них, генерацию логических запросов к аналитическому серверу, хранение и администрирование метаданных (Web-каталог) для отчетов и интерактивных панелей, осуществляет дополнительную пост-обработку данных.

Для достижения высокой производительности и масштабируемости системы Oracle BI Server и Oracle BI Web можно объединять в кластеры. Поддерживается возможность балансировки нагрузки, распределения запросов и пользовательских сеансов на разные серверы. В состав платформы Oracle BI Suite EE входит следующий набор инструментов (клиентских приложений): BI Answers – инструмент для выполнения произвольных (ad hoc) запросов и анализа; BI Interactive Dashboard – интерактивные панели показателей, отображающие персонализированную информацию; BI Publisher – масштабируемое средство формирования регламентированных отчетов в разных форматах на основе данных из множества источников и их рассылки по различным каналам; BI Office Plug-In – инструмент работы с аналитическим сервером через такие приложения как MS Word, Excel и Powerpoint и др. Из недостатков системы Oracle BI можно выделить: сложность освоения системы для обычных пользователей; высокая стоимость продукта.

Стоимость лицензий наиболее полной версии Oracle Business Intelligence Suite Enterprise Edition Plus на 50 пользователей составляет около $ 290 000.

Компании SAP и IBM придерживаются разных концепций в реализации функциональности OLAP. В линейке продуктов SAP BusinessObjects отсутствует выделенный источник данных для анализа и используется виртуальная многомерная модель («юниверс» в терминах SAP BusinessObjects). Такая реализация называется DOLAP (Desktop OLAP), выделенного сервера нет, обработка кубов выполняется непосредственно на клиенте или сервере приложений.

DOLAP фактически является «клиентским» ROLAP (Relational OLAP). Т.е. не только не существует физической модели данных, но также отсутствует сервер для работы с единым виртуальным кубом. Принципиальное отличие DOLAP от MOLAP состоит в том, что в DOLAP-кубы (они называются микрокубы), строятся на стороне клиента при его запросе.

Линейка продуктов SAP BusinessObjects для создания OLAP-системы: Desktop Intelligence – средство конечного пользователя для построения запросов и анализа информации, так называемый «толстый клиент».

Web Intelligence – Средство конечного пользователя для построения отчетов и анализа информации, так называемый «тонкий клиент». Оно аналогично Desktop Intelligence, но ориентировано на работу через Web. Представляет собой сервер приложений, работа с которым идет через web-браузер. Он использует трехзвенную архитектуру. С помощью SAP BusinessObjects WebIntelligence можно создавать разнообразные отчеты, с которыми можно работать как в онлайн, так и в оффлайн режимах. Для этого бизнес-пользователям не нужно обращаться к ИТ-специалистам, так как доступ ко всей необходимой информации осуществляется с помощью простого перетаскивания мышью нужных элементов (названных понятными бизнес-терминами). В запросы можно вставлять разнообразные графики, применять опции форматирования и фильтрации, подкрашивать разными цветами тренды и особые ситуации, а также рассылать отчеты по электронной почте. Данный инструмент удивительным образом совмещает функциональность и простоту использования.

Designer. Предназначен для создания графического представления БД, которое будет использоваться в построителе отчетов, и используется специалистом, который знает SQL и понимает структуру БД. Специалист не создает отчеты, а проектирует модель, используя которую конечные пользователи строят свои отчеты.

Central Management Console. Как только специалист создал модель, администратор системы (возможно, это тот же специалист) может использовать Central Management Console для настройки ограничения прав доступа к данным, путем внесения в репозиторий (специальная база данных) информации о пользователе. Таким образом, администратор

управляет доступом к модели данных и собственно БД. Также в данном модуле существует планировщик заданий, который позволяет выполнять отчеты по расписанию или совершению событий и отправлять документы большому количеству пользователей.

В качестве источника данных может использоваться практически любая реляционная СУБД, ODBC-источник, персональные файлы (Excel, DBF, txt и т.п.). Установка дополнительных модулей позволяет использовать API наиболее распространенных MOLAP-серверов. В данной схеме может использоваться также единый репозиторий.

В отчете Business Objects можно объединять выборки, полученные из нескольких источников. В зависимости от того, как построен юниверс, выборка может представлять собой микрокуб. Таким образом, документ Business Objects представляет собой набор микрокубов, к каждому из которых можно применять доступную функциональность. Если пользователь запрашивает данные, которых нет в микрокубе, автоматически формируется запрос для подкачки данных из первоисточников.

Основные особенности архитектуры OLAP-системы на платформе SAP BusinessObjects: OLAP-анализ проводится над данными, взятыми непосредственно в первоисточнике (можно, естественно, работать с выделенным ХД).

На машине клиента SAP BusinessObjects осуществляется предварительное преобразование данных в многомерную модель. Документ SAP BusinessObjects содержит не только модель данных и метаданные, но и сами кубы.

Производительность OLAP-системы играет решающую роль. Если время отклика на запросы исчисляется минутами, такая система не принесет преимуществ, т.к. OLAP подразумевает исследование данных, постоянную смену измерений, проведение вычисления, смену уровней детализаций и применение большого количества прочих функций.

На производительность DOLAP-системы в первую очередь влияют следующие особенности архитектуры. У SAP BusinessObjects отсутствует выделенный многомерный источник данных. Это означает, что запросы, формируемые в терминах многомерной модели, транслируются непосредственно на клиенте в запрос на SQL, который выполняется сервером РСУБД (ODBC).

OLAP-движок расположен на клиенте. Кубы данных расположены непосредственно в пользовательских документах. В процессе анализа информации, в том случае если в кубе не хватает данных, SAP BusinessObjects полностью перестраивает обновленный куб.

Для успешной работы с кубами потребуется: выделенное хранилище/витрина данных; высокая скорость работы сети; высокая производительность каждой машины, на которой идет работа с SAP BusinessObjects.

Несоответствие любому из 3-х вышеперечисленных условий не позволит работать с адекватной скоростью. Более того, прямое обращение к данным, расположенным в БД транзакционных систем может очень сильно замедлить работы последней.

При выборе OLAP-системы следует понимать, какие задачи планируется решать и насколько структура бизнес-информации удачно ложится на многомерную модель данных конкретного OLAP-средства. В программных продуктах различных производителей можно найти много отличий. Одни из важнейших вопросов: работа с несбалансированными иерархиями и особенность построения измерения «Время».

Следствием того, что многомерный куб в SAP BusinessObjects – структура виртуальная, при работе с несбалансированными иерархиями Business Objects и Web Intelligence извлекает избыточное количество информации, что в конечном итоге отрицательно сказывается на скорости работы и нагрузке на вычислительные ресурсы вычислительной системы.

Особенностью модели SAP BusinessObjects является невозможность объединять в одном юниверсе данные из нескольких источников. Однако при необходимости микрокубы документа SAP BusinessObjects могут объединяться в одном микрокубе непосредственно пользователем.

От способа представления информации, ее наглядности и имеющихся функций зависит удобство использования системы – важный фактор, влияющий на успешность внедрения OLAP-решения.

Документ WebIntelligence представляет собой отчет, состоящий из выборок-микрокубов (возможно собранных из нескольких источников). На отчете можно также свободно размещать различные элементы оформления: текстовые надписи, картинки, ссылки на документы других приложений. SAP BusinessObjects является MS Office-совместимым по пользовательскому интерфейсу продуктом, т.е. возможности по форматированию отчета достаточно богаты и просты в использовании. Каждый документ Business Objects может содержать несколько отчетов, размещаемых на отдельных закладках (листах, по аналогии с MS Excel).

Стоимость лицензий наиболее полной версии SAP BusinessObjects на 50 пользователей составляет около $ 170 000.

Cognos использует MOLAP-технологию (Multidimensional OLAP), классическую архитектуру с использованием выделенного физически многомерного источника данных – куба PowerPlay и сервера для работы с ним.

Рассмотрим линейку продуктов Cognos для создания OLAP-системы. PowerPlay User – средство конечного пользователя для OLAP и подготовки отчетов. Специальное пользовательское решение для Excel – PowerPlay Excel. В состав PowerPlay также включен модуль PowerPlay Transformer. PowerPlay Transformer – инструмент для моделирования и генерации аналитического многомерного куба. PowerPlay Enterprise Server – сервер приложений и OLAP-сервер. Visualizer – сред-

ство создания и работы с информационными панелями, наиболее эффективного и наглядного инструмента для представления деловой информации и ее анализа. Access Manager – инструмент сквозного управления привилегиями и правилами доступа пользователей к информации.

Источником данных может являться многомерный куб PowerPlay или MOLAP-сервер сторонних производителей. Используя PowerPlay Transformer, разработчик проектирует куб и определяет расписание или условия, в соответствии с которыми должен обновляться куб. Документ PowerPlay по умолчанию не сохраняет данные. Однако в случае необходимости есть возможность сохранить отчет вместе с кубом.

Основные особенности архитектуры OLAP-системы на платформе Cognos: пользователь работает с предварительно подготовленной структурой данных – физически существующим многомерным кубом. В отчете Cognos можно сохранять только структуру документа, данные располагаются в выделенном кубе.

Многомерный куб PowerPlay физически представляет собой файл и позволяет создавать структуры с практически неограниченным числом измерений, уровней и строк фактов. Имеется возможность выбора параметра оптимизации скорости работы (загрузка куба или анализ данных) и автоматического создания партиций. Обновление куба можно проводить в режиме инкрементальной загрузки данных без остановки работы пользователей – в этом случае естественно необходимо быть уверенным в том, что информация за уже загруженный период не изменилась.

В зависимости от потребностей бизнеса, разработчик может определять график и описывать условия, в соответствии с которыми выполняется обновление куба PowerPlay.

Использование выделенного многомерного источника данных позволяет обеспечить высокую скорость работы даже с очень большими кубами (сотни мегабайт, большие размеры на практике не встречаются – используются выделенные ХД).

Для более удобной работы разработчик может создать несколько кубов (например, для решения задач различных отделов) и связать их для возможности сквозного перемещения в процессе анализа информации.

Интерфейс Cognos PowerPlay настроен под выполнение OLAP-анализа. Аналогично Business Objects, в отчет PowerPlay можно добавлять произвольные текстовые поля, графику, а также колонтитулы.

Все операции по вращению куба и перемещению по измерениям непосредственно доступны в рамках всего многомерного (возможно очень большого) куба. Пользователь перетаскивает необходимые объекты из левой части экрана (вращение) или кликая по ячейкам переходит на другой уровень иерархии (детализация, свертка).

В Cognos используется несколько видов диаграмм и графиков, для каждого из которых предусмотрена настройка различных параметров.

Среди встроенных функций в Cognos PowerPlay имеются: итого, процент от общего, процент от определяемой ячейки, нарастающий процент, возведение в степень, суммирование, вычитание, произведение, деление столбцов/строк, изменение процента и пр. Используя формулы, можно реализовать более сложные вычисления.

Стоимость лицензий наиболее полной версии IBM Cognos на 50 пользователей составляет около $ 155 000.

Дискуссия

При работе с крупными ХД строятся OLAP-кубы для достижения максимальной производительности. Встает вопрос: как оценить максимальную производительность?

Проанализируем основные характеристики OLAP-кубов – режим хранения данных и уровень агрегирования. Режимы хранения представлены в табл. 1.

Таблица 1

Основные режимы хранения OLAP-кубов

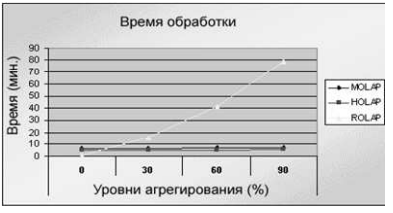
Режим хранения данных	Описание
Реляционный OLAP (ROLAP)	Данные-факты и агрегаты хранятся на сервере реляционной БД
Многомерный OLAP (MOLAP)	Данные-факты и агрегаты хранятся на OLAP-сервере в оптимизированном многомерном формате
Гибридный OLAP (HOLAP)	Данные-факты хранятся на сервере реляционной БД, а агрегаты – на OLAP-сервере в оптимизированном многомерном формате

Агрегаты – это предварительно рассчитанные суммы данных таблицы фактов для определенных комбинаций уровней из каждого измерения. Функции агрегатов: обработка запросов и создание дополнительных агрегатов. При выборе количества агрегатов (в процентах) для включения в куб необходимо учитывать объем хранимой информации и время выполнения запроса. Объем хранимой информации в БД при предварительном расчете всех возможных агрегатов требует значительного увеличения дискового пространства. Необходимое для обработки запроса время также увеличится при расчете агрегатов в момент обработки запроса.

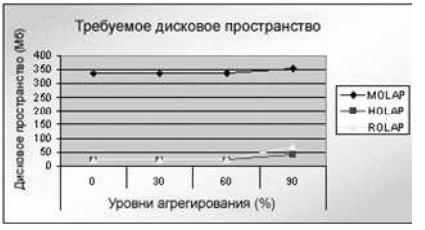
Для анализа производительности OLAP-кубов в среде **SQL Server Analysis Services** проведено тестирование для определения времени обработки запросов, требуемого дискового пространства, мощности

центрального процессора на сервере реляционной базы данных и аналитическом сервере. Эти параметры определялись для многомерной базы данных OLAP, содержащей кубы одинаковой структуры, но с различными режимами хранения (приведенными в табл. 1), а также для режима организации хранения данных по схеме «звезда» в соответствии с ER-диаграммой (от англ. entity-relationship diagram, диаграмма «объекты-отношения»). Тематика запросов – доходы и расходы клиентов в различные периоды времени и по различным продуктам (предоставляемым услугам). Пример обрабатываемого запроса: «Чему равен экономический доход за первые кварталы 2014 и 2015 годов? Сравнить соответствующие значения по каждому потребительскому сегменту». При обработке результатов графики построены для уровней агрегирования 0%, 30%, 60% и 90%, хотя в большинстве случаев используются только значения от 30% до 60% (0% и 90% были включены для сравнения). Следует отметить, что уровень агрегирования характеризует увеличение производительности обработки запросов по сравнению с отсутствием предварительно рассчитанных агрегатов данных.

Определено время обработки для каждого режима хранения. Результаты, представленные на рисунке 6 *а*, получены путем обработки идентичных по структуре кубов с различными режимами хранения и уровнями агрегирования. На рисунке 6 *б* приведены графики, показывающие изменение требуемого пространства на диске в зависимости от уровня агрегирования для каждого режима хранения.



а) время обработки для каждого режима хранения



б) требуемый объем дискового пространства

Рис. 7. Результаты обработки OLAP-кубов

Таблица 2 показывает объем требуемого дискового пространства для MOLAP-куба по сравнению со схемой «звезда», построенной в соответствии с ER-диаграммой.

Таблица 2

Результаты анализа требований к дисковому пространству

Уровень агрегирования (%)	Дисковое пространство для MOLAP-куба	Размер схемы «звезда» (таблицы фактов и измерений с индексами)	Степень сжатия данных при построении кубов MOLAP
60	335.75	5188	93.53
90	353.11	5188	93.19

На основе анализа данных рис. 1 и табл. 2 определены характеристики хранения многомерных данных при различных режимах, сформулированные в таблице 3.

Таблица 3

Характеристики хранения многомерных данных

№ п/п	Характеристики хранения	Примечание
1	При уровне агрегирования 0% ROLAP потребовалось наименьшее количество времени для обработки куба	Данные таблицы фактов и измерений в куб не добавляются и агрегаты не рассчитываются
2	По мере увеличения уровня агрегирования, ROLAP – по сравнению с MOLAP или HOLAP – затрачивает все больше времени на обработку куба	
3	Различие между MOLAP и HOLAP в промежутке 30 - 60% незначительно	
4	Время обработки MOLAP и HOLAP увеличивается в промежутке 60 - 90%, но несильно	
5	Время обработки ROLAP увеличивается экспоненциально в промежутке 60 - 90%	
6	Режим хранения MOLAP требует больше места, чем HOLAP или ROLAP Режим хранения MOLAP требует больше места, чем HOLAP или ROLAP	Кубы MOLAP содержат копии исходных фактов и измерений
7	Различие в количестве потребляемого дискового пространства при режимах MOLAP и HOLAP незначительно в интервале 0…60% и увеличивается по мере приближения к уровню 90%	
8	Режим хранения HOLAP использует наименьшее количество дискового пространства	Копии исходных фактов и измерений отсутствуют в базе данных OLAP, а агрегаты хранятся в оптимизированном многомерном формате
9	Режим хранения ROLAP требует дополнительного места на диске, когда уровень агрегирования превышает 30% и когда он приближается к 90%	График учитывает объем пространства, требуемого для хранения агрегатов данных в реляционной базе данных

Выводы

Сформулирована методика выбора инструментальных средств информационных систем для анализа многомерных корпоративных данных. В качестве BI системы решено выбрать систему SAP Business Objects.

Проведена оценка производительности инструментов многомерных хранилищ данных – OLAP-кубов. Для достижения максимальной производительности занимаемая OLAP-кубом память составляет примерно 7% от объема, требуемого для схемы режима хранения «звезда», построенной в соответствии с ER-диаграммой. Даже при 90%-ом уровне агрегирования удается достичь почти такой же степени сжатия. Проведена классификация режимов хранения OLAP-кубов. Дополнительное пространство, необходимое для куба в различных режимах хранения, зависит от количества уровней в измерении, количества мер и типа данных.

Список литературы

1. Христофорова, И.В. Корпоративное управление в России: формирование и развитие комплекса интеграционных стратегий: монография / И.В. Христофорова, Е.А. Эльканова, Д.Р. Макеева, О.А. Сырейщикова, В.И. Переяславский, В.Я. Вилисов, Н.З. Атаров; под общ. науч. ред. И.В. Христофоровой. – М., 2015. – 248 с.

2. Артюшенко, В.М. Современные исследования в области теоретических основ информатики, системного анализа, управления и обработки информации / В.М. Артюшенко, Т.С. Аббасова, И.М. Белюченко, Н.А. Васильев, В.Н. Зиновьев, Ю.В. Стреналюк, Г.Г. Вокин, К.Л. Самаров, М.Е. Ставровский, С.П. Посеренин, И.М. Разумовский, В.Ю. Фоминский: монография / под науч. ред. д-ра техн. наук, проф. В.М. Артюшенко. – Королев: ГБОУ ВПО ФТА, 2014. – 174 с.

3. Артюшенко, В.М. Системный анализ в области управления и обработки информации: монография / В.М. Артюшенко, Т.С. Аббасова, Ю.В. Стреналюк, Н.А. Васильев, И.М. Белюченко, К.Л. Самаров, В.Н. Зиновьев, С.П. Посеренин, Г.Г. Вокин, А.П. Мороз, В.С. Шайдуров, С.С. Шаврин; под науч. ред. д-ра техн. наук, проф. В.М. Артюшенко. – Королев МО: МГОТУ, 2015. – 168 с.

4. Артюшенко, В.М. Информационные технологии и управляющие системы: монография / В.М. Артюшенко, Т.С. Аббасова, Ю.В. Стреналюк, В.И. Привалов, В.И. Воловач, Е.П. Шевченко, В.М. Зимин, Е.С. Харламова, А.Э. Аббасов, Б.А. Кучеров; под науч. ред. д-ра техн. наук, проф. В.М. Артюшенко. – М.: Изд-во «Научный консультант», 2015. – 185 с.

5. Сидорова, Н.П. Методы и средства моделирования ИТ-инфраструктуры предприятия / Н.П. Сидорова // Вопросы региональной экономики. – 2010. – Т. 3. № 3. – С. 81–90.

6. Сидорова, Н.П. Информационные технологии оперативного анализа данных / Н.П. Сидорова, Н.В. Логачева, В.Ю. Добродеев // Информационно-технологический вестник. – 2014. – Т. 01. № 1. – С. 64–74.

7. Аббасов, А.Э. Оценка качества программного обеспечения для современных систем обработки информации / А.Э. Аббасов, Т.Э. Аббасов // Информационно-технологический вестник. – 2015. – № 3(05). – С. 15–27.

8. Аббасов, Э.М. Достижение максимальной производительности при работе с крупными хранилищами данных / Э.М. Аббасов, Э.Э. Акимкина // Информационные технологии. Радиоэлектроника. Телекоммуникации (ITRT-2016): сб. ст. VI международной заочной научно-технической конференции. – Тольятти: Изд-во: ПВГУС, 2016.

9. Аббасова, Т.С. Принципы и средства оптимизации высокоскоростных информационных каналов / Т.С. Аббасова // Информационно-технологический вестник. – 2014. – № 2(02). –С. 10–16.

10. Аббасова, Т.С. Применение диффузионной модели для балансировки нагрузки в неоднородных вычислительных системах / Т.С. Аббасова, Д.М. Двоеглазов, А.А. Борисов: сб. ст. II-ой Международной заочной научно-технической конференции «Информационные технологии. Радиоэлектроника. Телекоммуникации». – Тольятти: ПВГУС, 2012. Ч. 1. – С. 14–20.

11. Привалов, В.И. Повышение эффективности центров обработки данных для информационных систем / В.И. Привалов, Ю.В. Боковой, В.М. Зимин, Е.С. Харламова, В.А. Четкин, Е.П. Шевченко // Двойные технологии. – 2014. – № 4. –С. 75–78.

12. Аббасов, Э.М. Экспертная оценка качества программного обеспечения для трехмерного моделирования / Э.М. Аббасов, А.Э. Аббасов: сб. тр. по материалам Международной научно-практической конференции «Инновационные исследования: проблемы внедрения результатов и направления развития» (23.02.2016). – Уфа, 2016. – С. 18–21.

2. МНОГОМЕРНЫЕ ДАННЫЕ. OLAP-ТЕХНОЛОГИЯ, КАК КЛЮЧЕВОЙ КОМПОНЕНТ ХРАНИЛИЩ ДАННЫХ

Дм. Лобач

ОСНОВЫ OLAP

18.08.2003 /Доступно из
http://www.softkey.info/reviews/review.php?ID=465

Трудно найти в компьютерном мире человека, который хотя бы на интуитивном уровне не понимал, что такое базы данных и зачем они нужны. В отличие от традиционных реляционных СУБД, концепция OLAP не так широко известна, хотя загадочный термин "кубы OLAP" слышали, наверное, почти все. Что же такое OnLine Analytical Processing, где он обитает, и с чем его едят, мы и попытаемся разобраться.

OLAP – это не отдельно взятый программный продукт, не язык программирования и даже не конкретная технология. Если постараться охватить OLAP во всех его проявлениях, то это совокупность концепций, принципов и требований, лежащих в основе программных продуктов, облегчающих аналитикам доступ к данным. Несмотря на то, что с таким определением вряд ли кто-нибудь не согласится, сомнительно, чтобы оно хоть на йоту приблизило неспециалистов к пониманию нашего предмета. Поэтому в своем стремлении к познанию OLAP мы пойдем другим путем. Для начала мы выясним, зачем аналитикам надо как-то специально облегчать доступ к данным.

Дело в том, что аналитики – это особые потребители корпоративной информации. Задача аналитика – находить закономерности в больших массивах данных. Поэтому аналитик не будет обращать внимания на отдельно взятый факт, что в четверг четвертого числа контрагенту Чернову была продана партия черных чернил – ему нужна информация о сотнях и тысячах подобных событий. Одиночные факты в базе данных могут заинтересовать, к примеру, бухгалтера или начальника отдела продаж, в компетенции которого находится сделка. Аналитику одной записи мало – ему, к примеру, могут понадобиться все сделки данного филиала или представительства за месяц, год. Заодно аналитик отбрасывает ненужные ему подробности вроде ИНН покупателя, его точного адреса и номера телефона, индекса контракта и тому подобного. В то же время данные, которые требуются аналитику для работы, обязательно содержат числовые значения – это обусловлено самой сущностью его деятельности.

Итак, аналитику нужно много данных, эти данные являются выборочными, а также носят характер «набор атрибутов – число». Последнее означает, что аналитик работает с таблицами следующего типа:

Страна	Товар	Год	**Объем продаж**
Аргентина	Бытовая электроника	1988	105
Аргентина	Бытовая электроника	1989	117
Аргентина	Бытовая электроника	1990	122
Аргентина	Резиновые изделия	1989	212
Аргентина	Резиновые изделия	1990	217
Бразилия	Бытовая электроника	1988	313
Бразилия	Бытовая электроника	1989	342
Бразилия	Бытовая электроника	1990	337
Бразилия	Резиновые изделия	1988	515
Бразилия	Резиновые изделия	1989	542
Бразилия	Резиновые изделия	1990	566
Венесуэла	Бытовая электроника	1988	94
Венесуэла	Бытовая электроника	1989	96
Венесуэла	Бытовая электроника	1990	102
Венесуэла	Резиновые изделия	1988	153
Венесуэла	Резиновые изделия	1989	147
Венесуэла	Резиновые изделия	1990	162

Рис. 1

Здесь «Страна», «Товар», «Год» являются атрибутами, а «Объем продаж» – тем самым числовым значением. Задачей аналитика, повторимся, является выявление стойких взаимосвязей между атрибутами и числовыми параметрами. Посмотрев на таблицу, можно заметить, что ее легко можно перевести в три измерения: по одной из осей отложим страны, по другой – товары, по третьей – годы. А значениями в этом трехмерном массиве у нас будут соответствующие объемы продаж.

Рис. 2. Трехмерное представление таблицы. Серым сегментом показано, что для Аргентины в 1988 году данных нет

Вот именно такой трехмерный массив в терминах OLAP и называется кубом. На самом деле, с точки зрения строгой математики кубом такой массив будет далеко не всегда: у настоящего куба количество элементов во всех измерениях должно быть одинаковым, а у кубов OLAP такого ограничения нет. Тем не менее, несмотря на эти детали, термин «кубы OLAP» ввиду своей краткости и образности стал общепринятым. Куб OLAP совсем не обязательно должен быть трехмерным. Он может быть и двух-, и многомерным – в зависимости от решаемой задачи. Особо матерым аналитикам может понадобиться порядка 20 измерений – и серьезные OLAP-продукты именно на такое количество и рассчитаны. Более простые настольные приложения поддерживают где-то 6 измерений.

Измерения OLAP-кубов состоят из так называемых меток или членов (members). Например, измерение «Страна» состоит из меток «Аргентина», «Бразилия», «Венесуэла» и так далее.

Должны быть заполнены далеко не все элементы куба: если нет информации о продажах резиновых изделий в Аргентине в 1988 году, значение в соответствующей ячейке просто не будет определено. Совершенно необязательно также, чтобы приложение OLAP хранило данные непременно в многомерной структуре – главное, чтобы для пользователя эти данные выглядели именно так. Кстати именно специальным способам компактного хранения многомерных данных, «вакуум» (незаполненные элементы) в кубах не приводят к бесполезной трате памяти.

Однако куб сам по себе для анализа не пригоден. Если еще можно адекватно представить или изобразить трехмерный куб, то с шести- или девятнадцатимерным дело обстоит значительно хуже. Поэтому перед употреблением из многомерного куба извлекают обычные двумерные таблицы. Эта операция называется «разрезанием» куба. Термин этот, опять же, образный. Аналитик как бы берет и «разрезает» измерения куба по интересующим его меткам. Этим способом аналитик получает двумерный срез куба и с ним работает. Примерно так же лесорубы считают годовые кольца на спиле.

Соответственно, «неразрезанными», как правило, остаются только два измерения – по числу измерений таблицы. Бывает, «неразрезанным» остается только измерение – если куб содержит несколько видов числовых значений, они могут откладываться по одному из измерений таблицы.

Если еще внимательнее всмотреться в таблицу, которую мы изобразили первой, можно заметить, что находящиеся в ней данные, скорее всего, не являются первичными, а получены в результате суммирования по более мелким элементам. Например, год делится на кварталы, кварталы на месяцы, месяцы на недели, недели на дни. Страна состоит из регионов, а регионы – из населенных пунктов. Наконец в самих городах можно выделить районы и конкретные торговые точки. Товары можно объединять в товарные группы и так далее. В терминах OLAP такие

многоуровневые объединения совершенно логично называется иерархиями. Средства OLAP дают возможность в любой момент перейти на нужный уровень иерархии. Причем, как правило, для одних и тех же элементов поддерживается несколько видов иерархий: например день-неделя-месяц или день-декада-квартал. Исходные данные берутся из нижних уровней иерархий, а затем суммируются для получения значений более высоких уровней. Для того, чтобы ускорить процесс перехода, просуммированные значения для разных уровней хранятся в кубе. Таким образом, то, что со стороны пользователя выглядит одним кубом, грубо говоря, состоит из множества более примитивных кубов.

Рис. 3. Пример иерархии

Вот, кстати, мы и подошли, к одному из существенных моментов, которые привели к появлению OLAP – производительности и эффективности. Представим себе, что происходит, когда аналитику необходимо получить информацию, а средства OLAP на предприятии отсутствуют. Аналитик самостоятельно (что маловероятно) или с помощью программиста делает соответствующий SQL-запрос и получает интересующие данные в виде отчета или экспортирует их в электронную таблицу. Проблем при этом возникает великое множество. Во-первых, аналитик вынужден заниматься не своей работой (SQL-программированием) либо ждать, когда за него задачу выполнят программисты – все это отрицательно сказывается на производительности труда, повышаются штурмовщина, инфарктно-инсультный уровень и так далее. Во-вторых, один-единственный отчет или таблица, как правило, не спасает гигантов мысли и отцов русского анализа – и всю процедуру придется повторять снова и снова. В-третьих, как мы уже выяснили, аналитики по мелочам не спрашивают – им нужно все и сразу. Это означает (хотя техника и идет вперед семимильными шагами), что сервер корпоративной реляционной СУБД, к которому обращается аналитик, может задуматься глубоко и надолго, заблокировав остальные транзакции.

Концепция OLAP появилась именно для разрешения подобных проблем. Кубы OLAP представляют собой, по сути, мета-отчеты. Разрезая мета-отчеты (кубы, то есть) по измерениям, аналитик получает, фактически, интересующие его «обычные» двумерные отчеты (это не обязательно отчеты в обычном понимании этого термина – речь идет о структурах данных с такими же функциями). Преимущества кубов очевидны – данные необходимо запросить из реляционной СУБД всего один раз – при построении куба.

Поскольку аналитики, как правило, не работают с информацией, которая дополняется и меняется «на лету», сформированный куб является актуальным в течение достаточно продолжительного времени. Благодаря этому, не только исключаются перебои в работе сервера реляционной СУБД (нет запросов с тысячами и миллионами строк ответов), но и резко повышается скорость доступа к данным для самого аналитика. Кроме того, как уже отмечалось, производительность повышается и за счет подсчета промежуточных сумм иерархий и других агрегированных значений в момент построения куба. То есть, если изначально наши данные содержали информацию о дневной выручке по конкретному товару в отдельно взятом магазине, то при формировании куба OLAP-приложение считает итоговые суммы для разных уровней иерархий (недель и месяцев, городов и стран).

Конечно, за повышение таким способом производительности надо платить. Иногда говорят, что структура данных просто «взрывается» – куб OLAP может занимать в десятки и даже сотни раз больше места, чем исходные данные.

Теперь, когда мы немного разобрались в том, как работает и для чего служит OLAP, стоит, все же, несколько формализовать наши знания и дать критерии OLAP уже без синхронного перевода на обычный человеческий язык. Эти критерии (всего числом 12) были сформулированы в 1993 году Е.Ф. Коддом – создателем концепции реляционных СУБД и, по совместительству, OLAP. Непосредственно их мы рассматривать не будем, поскольку позднее они были переработаны в так называемый тест FASMI, который определяет требования к продуктам OLAP. FASMI – это аббревиатура от названия каждого пункта теста:

- Fast (Быстрый). Приложение OLAP должно обеспечивать минимальное время доступа к аналитическим данным – в среднем порядка 5 секунд;
- Analysis (Анализ). Приложение OLAP должно давать пользователю возможность осуществлять числовой и статистический анализ;
- Shared (Разделяемый доступ). Приложение OLAP должно предоставлять возможность работы с информацией многим пользователям одновременно;
- Multidimensional (Многомерность). См. выше;

Information (Информация). Приложение OLAP должно давать пользователю возможность получать нужную информацию, в каком бы электронном хранилище данных она не находилась.

Работа с OLAP-системами может быть построена на основе из двух описанных ниже схем.

Для «легковесного» применения подойдут OLAP-средства, встроенные в настольные приложения. Такие средства, как правило, имеют множество ограничений: на количество измерений, на допустимые иерархии и так далее. К подобным средствам, например, относится модуль Pivot Table, позволяющий работать с кубами в Microsoft Excel. Pivot Table входит в Microsoft Office с незапамятных времен и до недавнего времени был единственным OLAP-продуктом в его составе. В этом случае данные извлекаются модулем-клиентом непосредственно из реляционной СУБД.

В «тяжелых» случаях применяют двухступенчатую схему «клиент-сервер». Сервер обеспечивает непосредственно извлечение информации из СУБД и все прочее, необходимое для создания кубов. Специализированное же приложение-клиент предназначено для удобного (а главное – эффективного) просмотра кубов и выявления тех самых аналитических закономерностей, с которых мы начинали наш экскурс. В линейке продуктов Microsoft серверная часть представлена в лице Microsoft Analysis Services, которые входят в MS SQL Server. Сравнительно недавно в состав MS Office включен OLAP-клиент под названием Microsoft Data Analyzer.

С. Федечкин

ХРАНИЛИЩЕ ДАННЫХ: ВОПРОСЫ И ОТВЕТЫ

Об авторе: директор производственного центра
Datagy компании «Диасофт»; #31,26.08.2003
Режим доступа: http://www.olap.ru/basic/hd.asp

Хранилище данных как важнейший инструмент управления и развития бизнеса приковывает к себе все большее внимание. Публикации на эту тему обычно затрагивают технический и технологический аспекты. Мы же обратимся к некоторым концептуальным вопросам построения хранилищ и области их применения в банковском секторе.

Специалисты определяют хранилище данных как предметно-ориентированный, интегрированный, зависимый от времени набор данных, предназначенный для поддержки принятия решений различными группами пользователей. Так как хранилище носит предметно-ориентированный характер, его организация нацелена на содержательный анализ информации, а не на автоматизацию бизнес-процессов. Это свойство определяет архитектуру построения хранилища и принципы

проектирования модели данных, отличные от тех, что применяются в оперативных системах.

Интегрированность означает, что, например, данные о клиентах, подразделениях и банковских продуктах, полученные из различных источников, хранятся согласованно и централизованно. При этом полная информация о клиенте может включать данные, поступившие как из основной автоматизированной банковской системы (АБС), так и из фронт-офисного или иного приложения.

Хранилище содержит исторические данные, или зависимый от времени набор данных. Иными словами, если в оперативных источниках представлены самые последние значения (например, текущее наименование клиента или его физический адрес), то хранилище данных будет содержать в себе всю их предысторию с указанием периода, когда те или иные данные были актуальны. Хранилище данных предназначено для поддержки принятия решений, и его пользователи — это высший и средний менеджмент банка, аналитики, представители подразделений финансового анализа и маркетинга.

Предпосылки создания

Сегодня существует несколько движущих сил, или предпосылок, формирующих потребность в создании хранилищ данных.

Ужесточение конкуренции

После того как банковское сообщество пережило «кризис ликвидности» 1998 г., банки незаметно для себя вошли в полосу «кризиса доходности», характеризующуюся отсутствием высокодоходных финансовых инструментов и невысокой средней нормой прибыли. И сейчас многие из них по-настоящему поняли, что привлекать новых клиентов довольно дорого и трудоемко, так как большинство из них уже определились с выбором кредитного учреждения.

Развитие систем управления взаимоотношениями с клиентами (CRM)

Процесс выстраивания взаимоотношений с клиентами нацелен на сохранение старых и привлечение новых клиентов, что трудно осуществить без автоматизации продаж, маркетинга и совершенствования обслуживания. Для построения эффективной стратегии таких взаимоотношений необходимо хранилище данных, с помощью которого легко определить, какой клиент является наиболее прибыльным и выгодным для банка. Это даст любому кредитному учреждению возможность выработать единую и эффективную политику по отношению к каждому клиенту.

Разрозненность данных

Несмотря на то что банки склонны к централизации всех систем автоматизации в рамках единой АБС, достичь этого им удается далеко не всегда, поскольку неизбежно присутствуют разнородные источники информации. И хотя отдельные системы автоматизации позволяют получить отчет по определенной группе смежных банковских продуктов (чаще всего они отражают бухгалтерскую прибыль), этих данных недостаточно для управления бизнесом.

Возможные заблуждения и рекомендации по их разрешению

1. Хранилище данных – это OLAP

OLAP является аналитическим инструментом и одним, но далеко не единственным средством анализа данных в хранилище. Важно отметить, что средства OLAP могут быть использованы и вне хранилища. OLAP-анализ данных, находящихся в своих источниках, может быть произведен без их извлечения и загрузки в хранилище. Однако эффективность многомерного анализа при наличии хранилища данных резко возрастает.

Во избежание разночтений полезно провести демонстрацию конкретного OLAP-средства и на концептуальном уровне представить архитектуру хранилища данных. Обычно это позволяет определить единые понятия, необходимые для дальнейшего развития проекта.

2. Построение хранилища данных – задача только информационных технологий

Хранилище данных можно построить исключительно в тесном контакте ИТ- и бизнес-подразделений. Дело в том, что ИТ-специалисты компетентны в вопросах структуры источников данных и методов доступа к ним, а представители основных подразделений лучше понимают потребности бизнеса. Необходимо, чтобы конкретный заказчик внутри банка обладал достаточными полномочиями для поддержки проекта. Рекомендуется сформировать рабочую (проектную) группу или комитет, ответственный за создание и развитие хранилища данных.

3. Загрузка данных – это просто

Недооценка сложности процедур загрузки данных приводит к провалу большей части проектов, которые банки начинают делать самостоятельно.

Существует возможность минимизировать риски, связанные с загрузкой данных, за счет четкой формализации целей и задач проекта и исследования информационных источников на предмет достаточности и согласованности данных для решения поставленных задач. Благодаря

этому можно с самого начала выявить потенциальные трудности, связанные с исходными данными, и скорректировать потребности бизнеса, а также произвести нужные доработки в информационных системах.

4. Сначала загрузим все в хранилище, а уж затем определим цели

Загрузка данных – достаточно сложный процесс. Проведение его без определения целей анализа может привести либо к неполной востребованности хранилища данных, либо к необходимости в дальнейшем его серьезной переработки.

Перед началом проекта следует провести исследование потребностей бизнеса. Основная цель такого исследования – определение согласованных с руководством потребностей бизнеса в анализе. В итоге очень важно получить скоординированный с руководством заказчика документ, описывающий задачи анализа информации в порядке убывания их приоритета, а также результаты, которые может принести решение данных задач бизнесу. Это позволит осуществить декомпозицию задач анализа и разбить их решение на этапы. Следующим важным шагом должно стать исследование информационных источников, призванное гарантировать выполнение работ в поставленные сроки.

5. Хранилище данных – это готовая программа

Построение хранилища данных – проект, требующий серьезной проработки и усилий со стороны бизнеса и поставщика информационных технологий. Наиболее эффективным подходом здесь будет совместный проект банка и компании, специализирующейся в этой области.

Общемировая практика показывает, что хранилища данных создаются под конкретного заказчика. Серьезным преимуществом является наличие квалифицированного персонала, типовых витрин данных для бизнес-заказчиков, а также отраслевой модели данных.

6. Хранилище данных можно построить за пару недель

Цикл создания хранилища данных и решения первой значимой для бизнеса задачи не превышает трех месяцев. Сроки можно и сократить, но качество при этом заметно ухудшится. Хотя хранилище развивается итерационно, уже на первом этапе надо заложить серьезный фундамент не только для решения первой задачи, но и для развития аналитики в стратегической перспективе.

7. Централизованное хранение метаданных решит все проблемы

При построении хранилища данных необходимо использовать принцип централизации метаданных, но при этом важно понимать, что на нынешнем этапе развития информационных технологий централизовать хранение метаданных довольно сложно. Например, в технических

метаданных должны содержаться информация об источниках и их структуре, описание потоков данных и процессов перегрузки. Если первые два набора обычно поставляются вместе с информационной системой, то вторые, как правило, формируются в рамках проекта по созданию хранилища и размещаются на сервере перегрузки данных.

Основные элементы хранилищ данных

Рассмотрим некоторые компоненты хранилища данных на примере решения Datagy, созданного компанией Diasoft.

В целом такие компоненты подразделяются на два вида: структурообразующие и структурные. Первые представлены на схеме вертикальными прямоугольниками, а вторые – горизонтальными.

Оперативные источники данных

Конечно, желательно, чтобы АБС реализовывала все функции автоматизации бизнес-процессов и содержала все необходимые для анализа данные, но достичь этого практически невозможно. Вот почему при построении хранилища нужно быть готовым к подключению самых разнородных источников данных. Наибольшую сложность представляют слабоструктурированные пользовательские файлы (например, файлы MS Excel), строение которых порой трудно формализовать. Кстати, надо учитывать, что данные, извлеченные из всех этих разнородных источников, требуют согласования.

Процедура загрузки данных

Как показывает практика, ресурсоемкость процесса загрузки прямо пропорциональна сложности структуры каждого источника данных и экспоненциально зависит от их количества. Поставляющие информацию оперативные системы далеко не всегда обладают достаточным уровнем качества данных, поэтому процесс загрузки этих данных в хранилище не ограничивается простым их копированием или репликацией, а включает в себя очистку, согласование и контроль качества.

Отраслевая модель данных

Хранилище данных может быть реализовано как на реляционной, так и на многомерной СУБД. Но, как показывает практика, хранилища серьезного объема реализованы в основном на реляционных СУБД. Центральным компонентом хранилища является отраслевая модель данных, и ее тщательная проработка во многом определяет успешность проекта в целом.

Витрины данных

Витрины, построенные на основе хранилища данных или на базе первичных источников, проектируются для удовлетворения потребно-

стей определенной группы пользователей, ориентированных на решение конкретных аналитических задач. Витрины позволяют сравнительно легко обеспечить приемлемую производительность, так как содержат меньший объем данных, заблаговременно их агрегируют и востребованы ограниченным кругом пользователей. Для построения такой витрины можно использовать как реляционные, так и многомерные СУБД.

Представление данных и способы их анализа

Существует несколько подходов к анализу данных в хранилище. Основными считаются:

– интерактивный анализ данных (Online Analytical Processing, OLAP) – компьютерное приложение, поддерживающее многомерное представление и визуализацию данных с целью их анализа и подготовки отчетов;

– периодически выпускаемая отчетность (Reporting) – отчеты в стандартных формах;

– нерегламентированная отчетность (Ad-Hoc Reporting) – возможность получать быстрый доступ к реляционной базе данных для ответов на запросы, формируемые менеджерами «на лету»;

– интеллектуальный анализ данных (Data Mining) – процесс анализа больших наборов данных, применяемый для обнаружения связей между различными их элементами и поиска скрытых закономерностей.

Методология

На протяжении своего жизненного цикла хранилище данных итерационно модифицируется, и очень важно, чтобы каждый такой этап не только решал конкретные задачи бизнеса, но и оставлял возможность для развития. При правильно выбранной методологии, опираясь на хранилище данных, можно сформировать единый подход к решению аналитических задач банка.

Метаданные

Метаданные можно разделить на два класса: технические и бизнес-метаданные; последние представляют собой описание данных на языке бизнес-пользователей. Иными словами, бизнес-метаданные — это слой абстракции, который позволяет бизнес-пользователям работать с системой, концентрируя свое внимание исключительно на предмете анализа, а не на технических деталях системы. Качество и полнота бизнес-метаданных во многом определяют степень успешности проекта по созданию хранилища данных.

Технические метаданные включают в себя статистику загрузки данных в хранилище и их использования, описание моделей данных, структуры источников и реципиентов, а также метаданные приложений.

Типичные задачи, решаемые с помощью хранилищ данных

Известно, что существуют определенные классы задач, которые лучше решать в рамках хранилища данных. К ним относятся, в частности, анализ клиентской базы, анализ продаж и анализ доходов, а также управление пассивами и активами.

Анализ клиентской базы позволяет сформировать целевые сегменты клиентов и использовать эту информацию при продаже банковских продуктов и услуг. Целевые сегменты формируются на основе демографических и фирмографических сведений, финансовых показателей (например, оборота или прибыли), отраслевых признаков и других параметров клиентов.

Одним из наиболее важных вопросов является выделение сегментов прибыльных клиентов, нацеленное на их последующее удержание. В частности, за счет более детальной сегментации подразделения маркетинга начинают лучше понимать потребности клиентов и могут использовать эти данные при проведении маркетинговых кампаний. Анализ клиентской базы и сегментация дают возможность приблизиться к реализации концепции индивидуального маркетинга и более эффективно применять систему управления взаимоотношениями с клиентами.

Анализ продаж помогает выявлять тенденции, планировать продажи по продуктам, клиентам, подразделениям и, исходя из результатов сбыта, строить механизмы стимулирования клиентских и продуктовых подразделений. Благодаря использованию хранилища данных можно получить интегрированное представление о результатах продаж и взять эту информацию на вооружение при формировании планов.

Анализ доходов актуален для любого банка, причем более всего востребован анализ в разрезе клиентов. Очень важно также иметь представление о распределении доходов по продуктам и услугам, каналам предоставления услуг и подразделениям банка. Анализ доходов в разрезе клиентов и продуктов позволяет формировать «уникальные» предложения для каждого «уникального» клиента с целью максимизации прибыли в долгосрочной перспективе. Он способствует формированию ценовой политики банка, выделению сегментов, продуктов и услуг, которые стратегически важны для него.

Управление активами и пассивами. С помощью хранилища данных можно проводить эффективный анализ активов и пассивов и управлять не только ими, но и мгновенной ликвидностью банка на основе инструментального и портфельного подходов. Эти задачи решаются при минимальных затратах на подготовку специальных данных и с учетом лишь ограниченного объема информации, собираемой из источников в филиалах. Программный комплекс обеспечивает загрузку из информационных источников семи типов и позволяет формировать несколько десятков отчетов.

УДК 004.891

М.М. Горохов, Д.А. Переведенцев

ФОРМИРОВАНИЕ МНОГОМЕРНОЙ МОДЕЛИ ДАННЫХ ДЛЯ ЦЕЛЕЙ OLAP-АНАЛИЗА В ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЕ УПРАВЛЕНИЯ НАУЧНЫМИ ПРОЕКТАМИ

Вестник ВГУ, Серия: Системный анализ и информационные технологии. 2016. № 3.
Поступила в редакцию 11.09.2016 г.

Аннотация. *В статье дается описание многомерной модели данных с целью выявления измерений и атрибутов OLAP-куба для анализа данных в информационно-аналитической системе (ИАС), представлены принципы формирования аналитических срезов OLAP-куба. На основе выделенных центров принятия решений в процессе управления научными проектами и ER-модели данных определены возможности поддержки принятия управленческих решений и направления многомерного анализа имеющихся данных на примере разработанной ИАС.*

Ключевые слова: *многомерная модель данных, MD-модель, OLAP-анализ, экспертная система, управление научными проектами, информационно-аналитическая система.*

Annotation. *The paper describe the multi-dimensional data model in order to identify measurements and attributes of the OLAP-cube to analyze the data in the information-analytical system (IAS), presented the principles of formation of analytical sections of the OLAP-cube. Based on the identified centers of decision-making in the process of scientific project management and ER-model, data identified the ability to support management decision-making and directions of multi-dimensional analysis of the available data on the example developed by IAS.*

Keywords: *multidimensional data model, MD-model, OLAP-analysis, Expert System, management of science project, information-analytical system.*

Введение

Эффективность практического использования информационно-аналитической системы (ИАС) оценивается степенью своевременной и адекватной помощи в принятия решений в типовых, формализованных

ситуациях, а также в возможности прогнозирования ситуации с вероятностным исходом на основе накопленной информации, поскольку специализированные системы позволяют накапливать данные, используя их в дальнейшем для принятия более обоснованных управленческих решений в заданной ситуации путем обработки и удобного представления информации.

Для решения данной задачи наиболее эффективным является подход к обработке данных на основе технологии OLAP, поскольку данная технология представляет большие в сравнении, например, с SQL, возможности для исследования данных, манипуляции ими и формирования различных отчетов, а также их графическое представление. Кроме того, интеграция технологии OLAP в корпоративную ИС является перспективным с точки зрения развития ее функциональных возможностей в будущем, выступая и в качестве эффективного средства использования данных предшествующих периодов, основанных на стандартных (относительно простых) методах формирования отчетов и группировки итогов во многих разрезах аналитических статей, ресурсным показателям, так и инструмента дополнительного получения прогнозов, адекватно данным предшествующих периодов, основанных на экономико-математическом моделировании. Позволяя выявлять закономерности и давать оптимальные управленческие рекомендации согласно построенным прогнозам.

Таким образом, направление анализа, форма и содержание необходимых отчетов задается целями предприятия и конкретными научными интересами ее сотрудников, а также возможностью и степенью автоматизации отдельных процедур управления научными проектами. А интеллектуальный анализ данных с помощью ИАС [1] сводится к выбору набора данных и подходящей модели и структуры, которые используются для обработки, выявления и создания необходимой информации.

Описание методики

Системам поддержки принятия решений присущи определенные информационные и предметные стандарты, отражающие традиции профессиональной области приложений. Специфика предметной области данных представлена так называемой информационной базой (ИБ). ИБ является, по существу, совокупностью конфигурации и данных, доступных в пользовательском режиме согласно типовой (или специфической) конфигурации.

Конфигурируемость ИБ обеспечивает реализацию большого спектра предметных специфических задач автоматизации учёта и управления. Одинаковая конфигурация разных ИБ обеспечивает похожие условия пользовательской эксплуатации. Данные, регистрируемые в ИБ один раз, могут быть востребованы многократно [2].

На примере разработанной ИАС «UNIProject» [1], единые ИБ позволяют конструировать необходимые отчеты и автоматизировать отдельные бизнес-процессы, к примеру, процесс формирования заявки на конкурс (рис. 1).

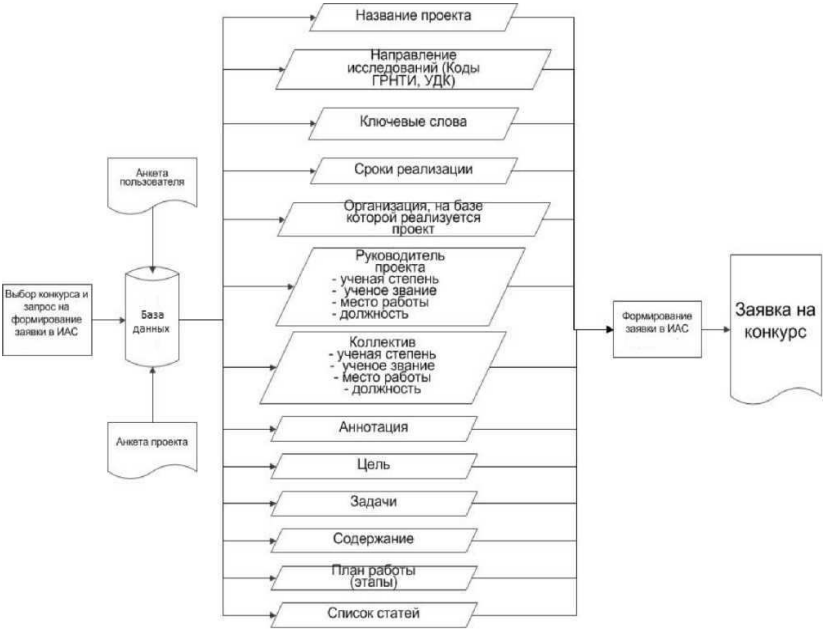


Рис. 1. Процесс автоматического формирования заявки

Таким образом, разно профильные ИБ обеспечивают создание единого информационного пространства на предприятии, поскольку конфигурацию конкретной ИБ уместно рассматривать как информационную модель определенной предметной области, сферы учета, управления. И при этом эффективно реализуемы связи между ИБ.

Систему автоматизации учета и управления в качестве надстройки над СППР уместно рассматривать как современную инструментальную среду для оптимизации разных совершенствуемых со временем бизнес-процессов, способствующих принятию решений [2].

Согласно проведенному анализу [4], в обозначенной предметной области выделяется несколько центров принятия решений и потребления отчетов:

1. Руководитель предприятия или подразделения;
2. Аналитик (менеджер проектов);
3. Сотрудник предприятия, руководитель проекта.

Тогда состав принимаемых решений на основе имеющейся информации и аналитических отчетов будет иметь следующую структуру (рис. 2).

Рис. 2. Структура информационных потоков и принимаемых решений

Из рисунка 2 видно, что основными создателями и потребителями аналитических отчетов являются специалисты по работе с научными проектами. Он формирует запрос в БД с целью извлечения аналитической информации. Поэтому кубы данных проектируются для поддержки аналитических запросов и строятся в соответствии с определенными специалистом измерениями. Ответственными за формирование аналитических отчетов являются специалисты по работе с научными проектами, в нашем случае – менеджер проектов. Он формирует запрос в многомерную БД с целью извлечения аналитической информации.

Важными для аналитика научных проектов являются метрики OLAP-куба, поскольку, являясь результатом вычислений и переменной величиной, они соответствует фокусу исследования данных.

Количество возможных измерений данных для анализа задается многомерной моделью концептуального уровня, которая может быть получена на основе построенной ранее ER-модели [4]. Данный подход отличается тем, что в ER-модели дополнительно вычленяются в отдельные сущности атрибуты, существенные в плане анализа, после чего

связи типа «один-ко-многим» рассматриваются как потенциальные гиперкубы (атрибуты связей – как меры гиперкуба, а связываемые сущности – как измерения гиперкуба). Идея преобразования ER-модели в многомерную модель (MD-модель) заключается в том, что каждая связь типа «один-ко-многим» следует рассматривать как потенциальную факт-сущность MD-модели, при этом связь один-ко- многим» задают иерархии в измерениях кубов. Это обстоятельство служит отправной точкой проектирования MD-модели.

Поскольку исходная ER-модель может содержать несколько связей типа «один-ко-многим», которые связывают одни и те же сущности, отражая различные аспекты взаимодействия связываемых сущностей, это обычно является следствием нормализации модели данных. При переходе к MD-модели целесообразнее иметь укрупненные кубы, позволяющие выполнять анализ по различным аспектам. Тогда добавление нового измерения в обобщенный куб, базирующийся на общих измерениях, позволит селектировать срезы исходных кубов [5].

При этом для обеспечения возможности анализа по хронологии событий в MD-модели атрибуты времени вычленены в самостоятельную сущность «Период», фиксирующую даты конкретных событий. Полученная MD-модель представлена на рис. 3.

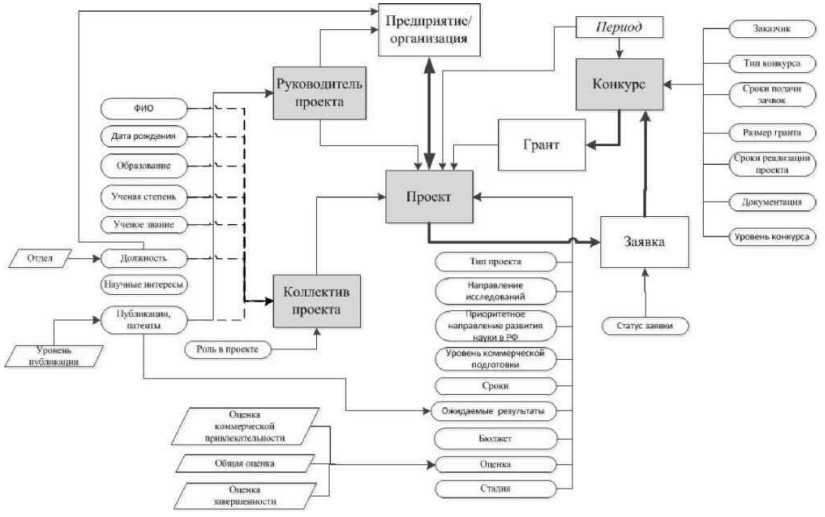


Рис. 3. MD-модель данных

На схеме четко выделяется 8 элементов, являющихся потенциальными центрами для анализа, а поскольку они имеют несколько связей типа «один-ко-многим», то они объединяют в себе несколько кубов,

назовем такую конструкцию гиперкубом. Описание элементов множественных связей дано в табл. 1.

Таблица 1

Элемент	Описание	Составляющие элементы
Руководитель проекта	Список параметров, характеризующих исполнителя проекта в роли «руководитель»	ФИО, дата рождения, образование, ученая степень, ученое звание, должность, научные интересы, публикации (патенты)
Коллектив проекта	Список параметров, характеризующих исполнителей проекта в других ролях	
Параметры проекта	Список параметров, характеризующих содержание проекта	Тип, направление исследований, приоритетное направление в РФ, уровень коммерческой подготовки, сроки, ожидаемые результаты, бюджет, оценка, стадии
Конкурс	Список параметров, характеризующих содержание требований к заявкам и условия их подготовки	Заказчик, тип, сроки подачи заявок, размер гранта, сроки реализации проекта в рамках гранта, уровень конкурса
Заявка	Является промежуточным атрибутом между проектом и конкурсом. Агрегирует и параметры проекта и конкурса	Содержит параметры проекта и конкурса
Период	Атрибуты хронологии событий. Имеет иерархию	Год, месяц
Публикации	Является как атрибутом исполнителя и его проектов, так и измерением	Уровень публикации
Предприятие/ организация	Является составляющим атрибутом всех центров анализа	Название предприятия, организации

Для выполнения OLAP-анализа отдельные кубы вычленяются из гиперкуба, при этом исходные измерения могут дополняться унаследованными измерениями и мерами (в качестве измерений) родительских кубов, а «родные» меры – сводными мерами кубов-наследников.

Разработанная многомерная модель концептуального уровня в совокупности с описанием формы принимаемых решений определяет направления и содержательную часть анализа имеющихся данных, что дает возможность преобразовать MD-модель в соответствии с выделенными измерениями и атрибутами, при этом в связи с одинаковым набором атрибутов центры анализа «Руководитель проекта» и «Коллектив проекта» объединим в один гиперкуб «Коллектив проекта» (рис. 4).

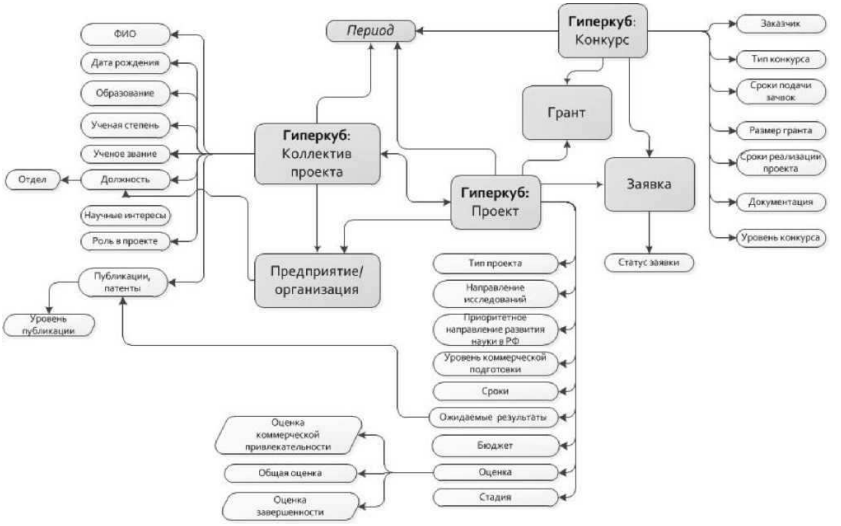


Рис. 4. Преобразованная MD-модель данных

Реализация принципов работы с данными на основе представленной структуры многомерной модели позволяет получить количественную информацию по множеству задаваемых параметров, что значительно облегчает проведение анализа научных проектов. А гибкая настройка модуля на предметную область конкретного предприятия предоставляет возможности:

– анализа пользователей ИАС в разрезе количества и уровня публикаций;

– построения профиля проектов в разрезе типа, стадии, уровня научного и коммерческого потенциала, отчетов по выигранным грантам и т.д.;

– анализа заявок на конкурсы в разрезе их статуса, процесса подготовки и успешности участия;

– анализа команды проектов;

– подготовки необходимых справок и аналитических отчетов.

Источником данных служит база данных ИАС «UNIProject» [1], реализованная на платформе Microsoft SQL Server, а использование в аналитическом модуле свободно распространяемого компонента Microsoft Office позволяет специалисту (менеджеру проектов):

– корректировать вид и наполнения формы визуализации данных;

– при необходимости созданные формы сохранить в системе в качестве шаблонов;

– дополнить или изменить структуру анализируемых данных;

– представлять данные в графическом и табличном виде.

Заключение

Работа с данными на основе представленной модели значительно расширяет возможности ИАС в интеллектуальном анализе больших объемов накопленных данных, позволяя менеджеру проектов в короткие сроки решить практическую любую нестандартную задачу по управлению научными проектами, поскольку специалист имеет оперативный доступ ко всей информации по проекту, собранной в единой базе данных и может проводить ее всесторонний анализ, реализуемый аналитическим модулем ИАС.

Список литературы

1. Благодатский, Г.А. Информационно-аналитическая система поддержки научной деятельности предприятий и вузов «UNIProject» / Г.А. Благодатский, Д.А. Переведенцев: сб. материалов XX Республиканской выставки-сессии студенческих инновационных проектов. – Ижевск: Издательство Иннова, 2015. – С. 31–37.

2. Попов, А.Л. Системы поддержки принятия решений: учебно-метод. пособие / А.Л. Попов – Екатеринбург: Урал. гос. ун-т, 2008. – 80 с.

3. Барсегян, А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – СПб.: БХВ-Петербург, 2004. – 336 с.

4. Переведенцев, Д.А. Разработка UML-модели информационно-аналитической системы перспективных научных проектов / Д.А. Переведенцев // Вестник ИжГТУ имени М. Т. Калашникова. – 2015. – № 4. – С. 58–60.

5. Макарова, Е.С. Проектирование концептуальной модели данных для задач web-olap на основе ситуационно-ориентированной базы данных / Е.С. Макарова, В.В. Миронов // Вестник УГАТУ – 2012. – № 6 (51). – С. 177–188.

Д. Кузнецов, Ю.А. Кудрявцев

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ OLAP-КУБОВ

Программирование. 2009. № 5. С. 43-36
(статья приводится в сокращении)

В 1993 году Э. Коддом была предложена концепция OLAP-систем (Online Analytical Processing), включающая в себя 12 правил представления данных пользователю. Подобные системы, как следует из названия, предназначены для анализа данных в интерактивном режиме. В связи с этим основной задачей OLAP-средств является представление больших объемов данных в виде, удобном для анализа конечными пользователями. Представление данных в виде многомерных кубов на сегодняшний день является de facto стандартом пользовательской работы с большими массивами данных.

В данной статье вводятся основные понятия OLAP-систем, которые затем формализуются с использованием математического аппарата теории решеток. В рамках введенной формализации доказывается оптимальность (с точки зрения объема хранимых элементов) представления OLAP-кубов замкнутыми решетками или эквивалентными им Quotient-решетками.

OLAP. Базовые понятия и терминология

Термин OLAP (Online Analytical Processing) был введен в 1993 г. Эдгаром Коддом [1]. Цель OLAP-систем – облегчение решения задач анализа данных. Кодд сформулировал 12 признаков OLAP-данных, и большинство современных OLAP-средств отвечают этим постулатам. Однако 12 признаков в дальнейшем трансформировались в 4 ключевых определения, сформулированные Найджелом Пендзом (см. [2]), на которые теперь ссылаются при определении OLAP- систем.

FASMI-тест. OLAP-система должна быть:

- Fast – быстрой, обеспечивать почти мгновенный отклик на большинство запросов.
- Multidimensional – многомерной, данные должны представляться в виде многомерных кубов.
- Information – данные должны быть полны с точки зрения аналитика, т.е. содержать всю необходимую информацию.

Большинство существующих OLAP-средств удовлетворяют всем этим признакам. Однако в реализации подобных приложений возникает ряд проблем, прежде всего связанных с увеличением объема данных, которые необходимо хранить.

В 1995 г. группа исследователей во главе с Джимом Греем [3], проанализировав создававшиеся тогда пользовательские приложения баз данных,

предложила расширение языка SQL – оператор CUBE. Этот оператор отвечает в SQL за создание многомерных кубов. Концепция многомерного представления данных является, наряду с моделью транзакций, одной из самых известных идей Кодда. В этой работе исследователи указали ряд эвристических рекомендаций по реализации новой структуры данных.

CUBE представляет собой обобщение операторов GROUP BY по всем возможным комбинациям измерений с разными уровнями агрегации данных. Каждая сгруппированная таблица относится к группе ячеек, описываемых кортежами из измерений, по которым формируется куб. Оператор, расширяющий SQL, называется CUBE BY (синтаксис такой же, как и у GROUP BY).

В стандарт SQL'99 был включен набор операторов для работы с OLAP-данными (запросы grouping set, rollup by, cube by, window by, rank, rownum и пр.).

Многомерные кубы, определение и свойства

Рассмотрим базовую (фактическую) таблицу *r*, на основе которой будет строиться OLAP-куб. Множество атрибутов *r* условно делят на 2 группы:

1. Набор измерений (категорий, локаторов), которые служат критериями для анализа и определяют многомерное пространство OLAP-куба. За счет фиксации значений измерений получаются срезы (гиперплоскости) куба. Каждый срез представляет собой запрос к данным, включающий агрегации.
2. Набор мер – функции, которые каждой точке пространства ставят в соответствие данные.

Из атрибутов *r* создаются измерения, содержащие проекцию *r* по атрибуту, с введенной иерархией (например, для таблицы, в которой хранятся фактические данные по продажам магазина, возможно наличие измерения под названием «Время», содержащего иерархию вида «Год-Месяц-Неделя-День»). Куб представляет собой декартово произведение измерений, где для каждого элемента произведения назначен набор мер. В кубе введены отношения обобщения и специализации (roll-up/drill-down) по иерархиям измерений (подробнее об иерархиях см. раздел 2.3). Ячейка высокого уровня иерархии может «спускаться» (drill-down) к ячейке низкого уровня (для примера 2.1 *(R1, ALL,* весна) может «спуститься» к ячейке (R1, книги, весна)) и наоборот, "подняться" (roll-up) (от (R1, книги, весна) к *(R1,ALL,* весна) по измерению «продукты»).

2.1. Пример

Рассмотрим пример, который будет в дальнейшем использоваться в этой статье.

Размер куба данных определяется по формуле $\Pi_d(c_i+1)$, где d – количество измерений («столбцов»), c_i – размерность измерения, т.е.

количество различных значений кортежей по этому измерению. Эквивалентный SQL-запрос: *SelectCount (distinct dimension) from cube_Table))*, +1 отвечает за «значение» ALL, агрегирующее все возможные значения измерения.

2.2. Измерения

Измерения куба – это набор доменов, по которым создается многомерное пространство. Важной особенностью OLAP-моделей является разделение измерений на локаторы (задающие точки) и меры (задающие значение). Как отмечено в [4], данное разделение может носить как условный, так и жесткий характер. В случае условного разделения измерения можно «разворачивать» как данные и как аналитики, создавая новую аналитику куба по продажам – «количество продаж». Таким образом, возрастает гибкость моделей и уровень абстракции. Однако этот подход, несмотря на свою привлекательность, сложен в реализации (в частности, отметим необходимость создания оптимальных алгоритмов хранения абстрактных типов данных) и, насколько нам известно, нигде промышленно не реализован. Теоретически, вкупе с моделированием решеток кубов логикой предикатов первого порядка, абстрагирование понятия «измерения» дает очень интересные результаты.

Локаторы куба отличаются иерархической структурой, и для получения значений мер на каждом уровне агрегирования вводятся агрегирующие функции.

2.3. Иерархии и агрегирование

Иерархичность данных – одно из важнейших свойств многомерных кубов. Иерархии призваны добавлять новые уровни в аналитическое пространство пользователя. Самым распространенным примером иерархии является «день–неделя–месяц–год». Между элементами разных уровней иерархии существуют отношения обобщения и специализации (roll-up/drill-down).

Таблица 1

Фактические данные для примера

Регион	Продукт	Время года	AVG (Продажи)
$R1$	книги	Весна	9
$R1$	Еда	Осень	3
$R2$	книги	Осень	6

Таблица 2

Куб для таблицы 1. Агрегирующая функция AVG

Регион	Продукт	Время года	AVG (Продажи)
$R1$	книги	Весна	9
$R1$	Еда	Осень	3
$R2$	книги	Осень	6
$R1$	книги	ALL	9
$R1$	ALL	Весна	9
ALL	книги	Весна	9
$R2$	ALL	ALL	6
ALL	Еда	ALL	3
ALL	ALL	Весна	9
ALL	ALL	ALL	6

Все иерархии можно разбить на 2 типа, о которых пойдет речь ниже. Основой разбиения будет служить расстояние d от корня ({*ALL, ALL,.ALL*}) до листьев. В случае, если *d = const*, – иерархии называются уровневыми (*leveled*), иначе – несбалансированными (*ragged*).

Примеры типов иерархий:

Уровневые: день – месяц – год; улица – город- страна.

Несбалансированные: Организационная диаграмма, различная группировка продуктов.

2.4. Агрегирующие функции, меры и формулы

Неотъемлемой частью OLAP-модели является задание функций агрегирования. Поскольку цель OLAP – создание многоуровневой модели анализа, данные на уровнях, отличных от фактического, должны быть соответствующим образом агрегированы. Важно отметить, что по каждому измерению можно задавать собственную (и не одну) функцию агрегации.

Таким образом, в случае куба с n измерениями функция агрегирования имеет вид

$$f(x) = [\{f_{1,1}, \dots f_{1,k_1}\}, \dots\dots, \{f_{n,1}, \dots f_{n,k_n}\}],$$

где x – точка куба, а $f_{i,j}$ – j-я функция агрегирования по i-*му* измерению.

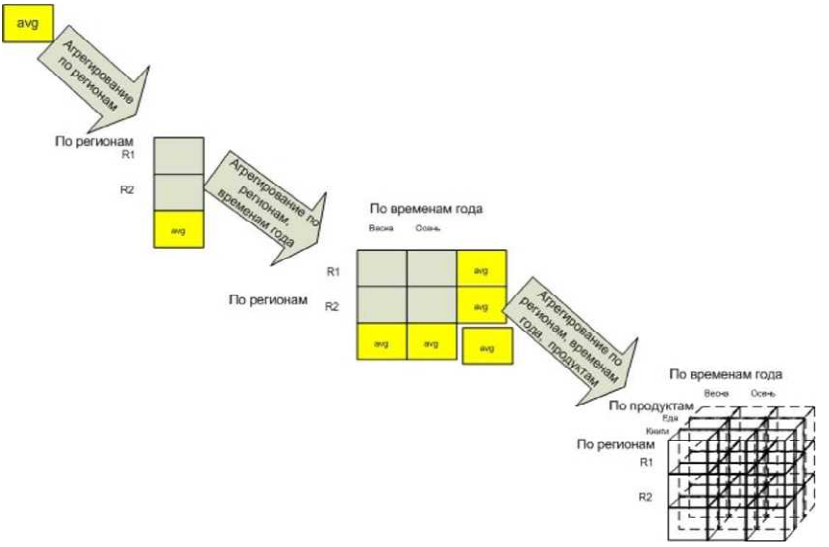


Рис. 1. Схема агрегирования данных для формирования куба

В [5] приведена следующая классификация агрегирующих функций с точки зрения сложности распараллеливания.

Таблица 3

Категории агрегирующих функций имеет вид

Категория	Примеры
Дистрибутивные	*Sum(), Count(), Minimum(), Maximum()*
Алгебраические	*Average(), Standard-Deviation(), Center-of-Mass(), MaxN(), MinN()*
Холистические	*Median(), Most Frequent (), Rank()*

Дистрибутивные функции позволяют разбивать входные данные и вычислять отдельные итоги, которые потом можно объединять.

Алгебраические функции возможно представить в виде комбинации из дистрибутивных функций (например, $Average()$ можно представить как $\frac{sum()}{count()}$)

Холистические функции невозможно вычислять на частичных данных или представлять каким-либо образом.

Выводы и направление дальнейших исследований

В данной статье представлена математическая модель OLAP-данных, проведена связь между представленной моделью и теорией решеток, доказана оптимальность представления OLAP-кубов замкнутыми решетками и quotient- решетками.

Дальнейшими направлениями исследований являются:

- Развитие математической модели. Модель должна удовлетворять требованиям, перечисленным в работах [10] и [11]. Основными задачами в данном направлении являются:

– Поддержка различных типов иерархий. Возможность задавать несбалансированные, неонтологичекие, нестрогие иерархии.

– Вероятностные меры. Возможность вводить данные с некоторым уровнем точности (часто точное число неизвестно) и на основе этих данных получать корректные результаты запросов.

– Объединение данных различных уровней гранулярности. Данные могут быть представлены на разных уровнях гранулированности (например, продажи на уровне региона, а не в конкретной кассе). В таком случае, данные должны корректно отображаться и позволять проводить анализ.

- Реализация парралельного алгоритма создания замкнутых решеток кубов. Представленная модель будет реализована в рамках открытого проекта MROLL (Map/Reduce OLAP Lattices) на базе кластера Apache/Hadoop, с учетом существующих работ по распараллеливанию обработки OLAP-кубов (см. [12] и [5]).

Список литературы

1. *Codd E.F.* Providing OLAP for end-user analysis: An IT mandate // ComputerWorld. 1993.
2. *Pendse N.* Olapreport: What is olap, 2005.
3. *Gray J, Bosworth A., Layman A., Pirahesh H.* Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals // Microsoft Lab, 1995.
4. *Thomsen E.* OLAP Solutions: Building Multidimensional Information Systems. Second Edition. Wiley Computer Publishing John Wiley & Sons, Inc., 2002.
5. *Goil S., Choudhary A.* High performance olap and data mining on parallel computers. Center of Parallel and Distributed Computing Technical Report TR-97-05. 1997.
6. *Гретцер, Г.* Общая теория решеток / Г. Гретцер. – М.: Мир, 1981.
7. *Casali A.* Mining borders of the difference of two datacubes // DaWaK. 2004.

8. *Nedjar S., Casali A., Cicchetti R., Lakhal L.* Emerging cubes for trends analysis in olapdatabas- es / Il Yeal Song, Eder J., Tho Manh Nguyen (eds.) DaWaK // Lecture Notes in Computer Science. Springer, 2007. V. 4654. P. 135-144.

9. *Yan Zhao.* Quotient cube and qc-tree: Efficient summarizations for semantic olap, 2003.

10. *Pedersen T.B., Jensen C.S., Dyreson C.E.* A foundation for capturing and querying complex multidimensional data // Inf. Syst. 2001. V. 26, № 5. P. 383-423.

11. *Rafanelli M.* (ed.) Multidimensional Databases: Problems and Solutions. Idea Group Publishing, 2003.

12. *Ge Yang, Ruoming Jin, Gagan Agrawal.* Implementing data cube construction using a cluster middleware: algorithms, implementation experience, and performance evaluation // Future Gener. Comput. Syst. 2003. V. 19, № 4. P. 533–550.

3. КОНЦЕПЦИЯ ХРАНИЛИЩ ДАННЫХ

А.И. Волков

ПРОБЛЕМЫ ИНТЕГРАЦИИ ХРАНИЛИЩ ДАННЫХ С ОТКРЫТЫМИ И БОЛЬШИМИ ДАННЫМИ И ПОДХОДЫ К ИХ РЕШЕНИЮ

Международная конференция CPT2014, Республика Кипр, Ларнака, 11-18 мая 2014 г. International conference CPT2014, Cyprus, Larnaca, May 11-18, 2014

В работе рассмотрены вопросы, связанные с комплексным использованием различных источников финансовой информации, методы интеграции данных для решения задач финансовой организации. Предложены подходы по включению в хранилища финансовых данных сведений из внешних источников, таких как открытые данные, информация, получаемая из государственных автоматизированных систем и др. Рассмотрено использование процессного подхода при разработке сервисной ИТ-компанией средств автоматизации обработки информации финансовой организации.

Ключевые слова: *финансовая информация, хранилище данных, открытые данные, большие данные, процессный подход в обработке финансовой информации, информационный объект*

Введение

В современном обществе используются различные виды информации, одним из наиболее востребованных из них является финансовая информация.

Наряду с традиционной финансовой информацией, касающейся сведений об объектах и субъектах финансовых операций, все большее значение получают связанные с ней другие виды данных, становящиеся доступными в автоматизированных системах как внутри организации, так и за ее пределами.

Актуальными направлениями расширения «информационного поля» применяемого в финансовых приложениях являются открытые данные (ОД), большие данные, сведения, предоставляемые государственными автоматизированными системами, а также метаданные о свойствах социальной и технологической среды, которую обслуживает финансовая система. Следует отметить, что в основе любого анализа данных лежит модель предметной области, некоторая концепция, кото-

рая используется для обработки данных и интерпретации получаемых результатов.

Работа выполнена и опубликована при финансовой поддержке РФФИ, гранты 14-07-00362, 14-07-06022.

Рассмотрим технологии получения, хранения и обработки различных видов данных, а также актуальные методические и технологические решения, использование которых позволяет повысить качество обработки финансовой информации.

Одним из основных элементов системы автоматизации финансовой организации (ФО) является хранилище данных (ХД), консолидирующее основную информацию организации. Поскольку «построение хранилищ данных – это не просто создание очень большой базы данных, это процесс» [1], мы рассмотрим также основные организационно-технологические вопросы, связанные с использованием в решении финансовых задач новых видов информации.

В основе функциональных требований к ХД лежат стратегические цели компании, которые определяют приоритеты в формировании структуры информации и особенности ее обработки. Данные являются результатом реализации бизнес-процессов компании, содержание которых определяется некоторым «процессным офисом». В процессе выполнения этих бизнес-процессов используются соответствующие программные средства и объединяется для последующего применения получаемая в результате их работы информация [2].

Требования к ХД меняются при изменениях рынка, стратегии компании, а также при появлении новых возможностей по получению и обработке информации о деятельности финансовых и других организаций. Структура ХД и методы работы с получаемой из него информацией должны меняться в «точках стратегического перелома» бизнес-процессов (БП) компании, когда конкурентная обстановка вокруг компании меняется и она должна адаптироваться к новым условиям [3].

Виды информации, применяемой в современной ФО

В современном обществе появляется возможность использования в режиме «реального времени» новых информационных возможностей. Этот процесс в полной мере касается ФО. Под ФО мы понимаем такую организацию, в основе деятельности которой находится предоставление финансовых услуг. К ФО относятся в первую очередь банки, страховые компании, организации обслуживающие платежные системы и др.

Рассмотрим использование основных видов информационного наполнения автоматизированных систем для решения финансовых задач.

На рисунке 1 показаны основные составляющие информационного наполнения систем обработки финансовой информации и наиболее существенные информационные связи между ними.

Рассмотрим эту схему более подробно применительно к деятельности некоторой обобщенной ФО. Основным источником информации о деятельности ФО является транзакционная система. В банках эта система обычно называется «автоматизированная банковская система» (АБС). Она обеспечивает механизмы авторизации пользователей, средства взаимодействия с клиентом с использованием принадлежащих ему компьютерных и коммуникационных устройств, возможность подключения к существующим системам безналичных денежных расчетов. Важной особенностью такой системы является согласованность, обеспечение достоверности и непротиворечивости всех выполняемых действий.

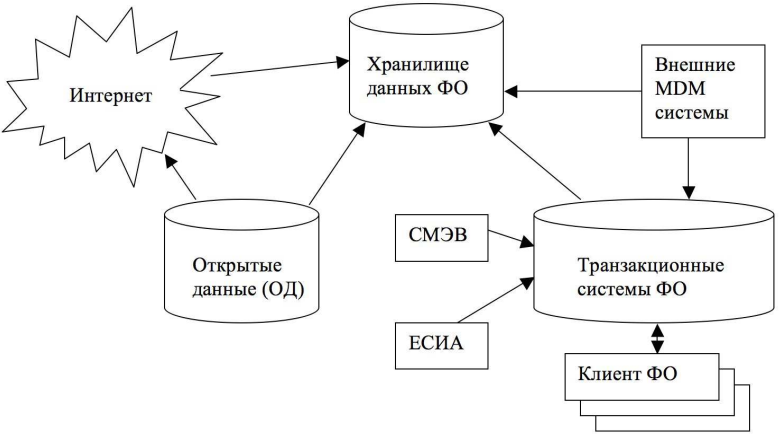


Рис. 1. Виды информации, применяемые в ФО и их взаимодействие

В перспективе произойдет интеграция с транзакционными системами ФО появившихся в государстве официальных систем идентификации и аутентификации для решения задач взаимодействия с другими организационными структурами. Это в первую очередь касается единой системы идентификации и аутентификации (ЕСИА). Заверенная государством информация на бумажных носителях (паспорта, различные документы государственных организаций и др.) будет постепенно замещаться их электронными аналогами и интегрироваться с информационной системой ФО, а также доступными внешними источниками данных.

Система межведомственного электронного взаимодействия (СМЭВ) является перспективным средством коммуникации, обработки запросов к различным государственным структурам и местным органам власти. Информация, появляющаяся в государственных структурах, должна быть увязана с данными собственной транзакционной системы

ФО. Поскольку сейчас эти средства автоматизации развиваются независимо, средства информационного обмена между ними, как правило, являются недостаточно эффективными. Поэтому ФО должна формировать новые и совершенствовать имеющиеся механизмы взаимодействия с формирующейся государственной информационной инфраструктурой.

Транзакционная система реализует взаимодействие с клиентами как с помощью собственных сотрудников, так и с использованием программного обеспечения для доступа клиентов к предоставляемым услугам. В этой деятельности во все большей степени применяется инфраструктура Интернет, средства автоматизации госструктур и партнеров. Например, все больше различных платежей за услуги государственных организаций осуществляются с использованием в той или иной степени средств автоматизации этих организаций и ФО должны обеспечивать взаимодействие клиентов с ними в автоматизированном режиме. Для информирования и подтверждения транзакций используются принадлежащие клиентам средства коммуникации с использованием в качестве партнеров телекоммуникационных компаний.

Информация транзакционной системы ФО выгружается в ХД – среду хранения и обработки информации, обеспечивающей ее эффективную аналитическую обработку (показано в верхней части рис. 1). В ХД также поступает информация из других систем обработки информации: внешних систем управления справочными данными – Master Data Management (MDM), систем предоставления открытых данных (ОД) и Интернет источников. Рассмотрим взаимодействие этих систем.

В процессе работы ФО в ней формируется внутренняя система классификации и кодирования, обеспечивающая работу транзакционной системы и ХД. Однако и в государстве в целом, в финансовой сфере формируется среда унифицированного представления информации, которую можно назвать внешней MDM-системой, обеспечивающей общепринятую классификацию и кодирование данных. Такая система формируется в виде общепринятых форматов и определений данных, унифицированных значений классификаторов, стандартизованных форм документов, онтологий.

Упомянутая внешняя MDM-система это не единая автоматизированная система, а некоторое множество систем, содержащих возникающую в автоматизированном режиме в других организациях информацию, необходимую для работы ФО. С этой инфраструктурой в условиях единой коммуникационной среды каждой ФО приходится считаться во все большей степени, обеспечивая, в том числе, прозрачность собственной среды автоматизации относительно общепринятой системы смыслов. В настоящее время мы находимся на этапе формирования внешней MDM инфраструктуры ФО, что усложняет процесс автоматизации. Однако со временем, с развитием средств интеграции на

методическом, технологическом, нормативном уровне, острота проблемы совместного использования информации будет снижаться. В результате возникнет новое качество в функциональных возможностях обработки финансовой информации.

Под ОД обычно понимаются данные, доступные для свободного использования и представленные в виде, удобном для использования в автоматизированных системах [4]. ОД в ФО являются средством связывания внутренней и внешней информации, позволяют дополнительно структурировать сведения, спонтанно возникающие в Интернет.

Все большее значение в деятельности ФО имеет Интернет (вверху справа на рис. 1). В этой информационной среде информация может быть представлена в различной форме, отличаться степенью структурирования и формализации. Она не всегда является достоверной. Однако в современных условиях Интернет является эффективной инфраструктурой для осуществления различных видов финансовой деятельности и источником сведений о состоянии и тенденциях развития общества. Возможно получение нетривиальной информации о клиентах, использование новых способов увеличения клиентской базы. ФО, более эффективно использующие в своей работе Интернет- информацию, получают дополнительные рыночные преимущества. Большой объем данных имеющихся в Интернет и высокие темпы роста этого объема требуют использования ФО специальных способов работы с ними, в частности интенсивно развивающихся технологий обработки Больших данных (Big Data), а также средств структурирования и оценки достоверности информации.

На основе вышеизложенного можно сделать вывод о необходимости использования в работе ФО новых видов информации, позволяющих повысить эффективность решения стоящих перед ними задач и обеспечить конкурентные преимущества на рынке финансовых услуг. Происходит все большая интеграция финансовой деятельности клиентов и партнеров ФО в автоматизированную информационную инфраструктуру общества и ФО должны учитывать эту тенденцию.

Типовые задачи ФО и критерии эффективности автоматизации

Принятые в таблицах сокращения: АССиОИ – автоматизированная система сбора и обработки информации; АСФО – автоматизированная система ФО; ПО – программное обеспечение; ПУР – принятие управленческих решений.

Существуют различные типы ФО: банки, страховые компании, платежные системы, а также посреднические структуры, участвующие в выполнении платежей и др. организации. Каждая организация решает свой круг задач, работа с которыми автоматизируется.

Текущие задачи, связанные с конкретными финансовыми операциями решаются в транзакционной системе, а аналитические – с использованием информации, собираемой в ХД в виде, удобном для обобщенной обработки.

ХД – основа для принятия решений. С их помощью могут решаться задачи, связанные с анализом доходности услуг ФО, изучением клиентской базы с целью повышения охвата рынка, выявлением востребованных услуг, возможностей оказания дополнительных услуг в регионах, получения необходимых видов отчетности и другие.

В таблице 1 в качестве примера приведен перечень основных аналитических задач, решаемых банком с использованием ХД, а также рассмотрены способы решения этих задач, источники данных и критерии эффективности.

Таблица 1

Основные аналитические задачи, стоящие перед банком, и подходы к их решению

Финансовые задачи	Способы решения	Источники данных	Критерии эффективности
Подготовка отчетности для регулирующих органов	Создание АССиОИ для расчета необходимых показателей и формирования регламентируемых отчетов	АБС, системы учета сделок	Удовлетворение требований регулирующих органов по срокам предоставления и качеству отчетов
Подготовка внутренней управленческой отчетности	Создание ХД и специализированных витрин	АБС, системы учета сделок. Для глубокого анализа – ОД	Эффективность ПУР

В таблице 2 приведены основные свойства рассмотренных выше технологических решений, которые используются для обработки финансовой информации.

Таблица 2

Применение информационных технологий для решения задач ФО,

Заголовок: (1) ИТ для обработки финансовой информации; (2) Краткая характеристика; (3) Ответственный за информационное наполнение и развитие; (4) Примеры решаемых задач; (5) Основные параметры с точки зрения ФО.

1-я строка: (1) Транзакционные системы; (2) Выполнение согласованной и авторизованной обработки финансовой информации в реальном времени; (3) ИТ-служба ФО; (4) Финансовые операции; (5) Время реакции системы. Согласованность. информации. Отказоустойчивость системы.

2-я строка: (1) Хранилища данных; (2) Сводная и ретроспективная информация для принятия управленческих решений; (3) ИТ-служба ФО; (4) Принятие решений на основе анализа данных в интересующих аспектах; (5) Полнота данных. Гибкость запросов. Масштабируемость.

3-я строка: (1) Технологии работы с клиентами ФО; (2) Возможность клиента в автоматизированном режиме решать свои задачи в ФО; (3) ИТ-служба ФО, партнеры, клиент (технические средства, умение использовать клиентское ПО в соответствии с регламентом ФО; (4) Выполнение финансовых операций клиентом. Поиск и оформление услуг; (5) Доля клиентов, использующих ПО ФО для собственного обслуживания. Место ПО ФО среди подобных решений по оценкам экспертов.

4-я строка: (1) СМЭВ; (2) Возможность получения заверенной информации, необходимой для финансовой деятельности от государственных и муниципальных структур; (3) Уполномоченные государственными структурами организации (например, Ростелеком); (4) Получение документов о юридических и физических лицах в электронном формате; (5) Доля запросов к госструктурам, которая выполняется в автоматическом режиме. Отсутствие задержек в получении информации.

5-я строка: (1) Открытые данные; (2) Использование общезначимых данных, сформированных уполномоченными структурами и в результате консенсуса заинтересованных пользователей в решении задач ФО; (3) Заинтересованные государственные, общественные, коммерческие организации. В отдельных случаях, возможно, с участием физических лиц; (4) Общезначимая информация об объектах социально-экономической инфраструктуры; (5) Семантическая однозначность информации. Актуальность и достоверность информации.

6-я строка: (1) ЕСИА; (2) Автоматизированная идентификация и аутентификация пользователей автоматизированных систем – физических лиц и представителей юридических лиц; (3) Уполномоченные государственными структурами организации (например, Ростелеком); (4) Осуществление доступа к автоматизированной системе ФО. Заочное заключение соглашений ФО с клиентом; (5) Защищенность системы. Простота и надежность использования и выявления нарушений в функционировании.

Примечание. Представлена в виде линейных списков.

Ниже мы рассмотрим отдельные наиболее перспективные, по нашему мнению, направления развития систем обработки финансовой информации.

Одним из путей такого развития является использование ОД.

ОД в задачах ФО

В автоматизированных системах, которые внедряются в обществе, появляются данные, которые не имеют ограничений в виде патентов, авторского права и др. по распространению и использованию в любых применениях – ОД [4]. Они позволяют решать на новом качественном уровне ряд задач, стоящих перед ФО.

В литературе имеются различные трактовки понятия ОД. Различия в определении этого понятия не позволяют единообразно сравнивать различные технические и технологические решения в этой области. Обычно ОД возникают в коммерческих системах или как результат внедрения различных государственных программ, таких как «открытое правительство». В отдельных случаях они могут возникать и в результате инициативы юридических и физических лиц.

Использование ОД требует решения ряда концептуальных, технологических и технических проблем. Некоторые вопросы, связанные с использованием ОД в современных системах, рассмотрены в работе [5]. Отмечается, что ОД являются результатом работы автоматизированных систем различных организаций и отражают существенные свойства предметов и явлений, которые могут использоваться непосредственно, а также служить ключами для связывания информации об описываемых этими данными объектах при выполнении различных аналитических операций с финансовыми характеристиками объектов. Например, если в ОД доступна информация о местоположении детских площадок в некотором районе города, а также сведения о стоимости каждой из этих площадок, то с помощью имеющихся средств анализа пространственной информации можно получить стоимость этих площадок на некоторой территории – в парке, районе города и др.

ОД могут быть классифицированы по различным основаниям, что позволяет глубже осмыслить и структурировать это явление.

Известную классификацию, характеризующую степень формализованности и возможность использования ОД в автоматизированных системах, предложил один из разработчиков концепции Всемирной Паутины Тим Бернерс-Ли [6, 7]. Он предложил «пятизвездочную уровневую модель» ОД.

Одна звезда – ОД в любом машиночитаемом формате ХД, т.е., файл с графическим образом документа получит одну звезду. Данные в проприетарном формате, где имеется какое-либо структурирование, допускающее автоматическую обработку, будут оценены в две звезды. При-

мер таких данных – файл электронной таблицы, в формате Excel. Структурированные данные в свободном формате, автоматизированная обработка которых не требует каких-либо лицензий и платежей оцениваются тремя звездами. Это, в частности, данные в структурированном текстовом формате CSV. Использование URL-ссылок в данных оценивается в четыре звезды. Пять звезд в оценке получают данные, предоставляемые в связи с другими данными, например доступные в виде взаимоувязанных таблиц.

Предоставление информации в рамках концепции ОД востребовано и поддерживается рядом федеральных и региональных структур России. Основанием для публикации ОД является подпункт «г» пункта 2 Указа Президента РФ № 601 от 7 мая 2012 года [8]. Так, например, имеются страницы ОД на официальных сайтах Федеральной налоговой службы [9], Росстата [10], Московского правительства [11]. На портале ОД Москвы кроме собственно данных и лицензионных соглашений размещены приложения, разработанные пользователями и организациями для различных компьютерных платформ, в которых используются эти данные.

Однако представленные в имеющихся сервисах ОД в силу ряда концептуальных, организационных и технологических причин структурированы недостаточно. При их формировании и поддержке недостаточно учитывается смысловое содержание ОД, не вполне определено их место среди других видов имеющихся информационных ресурсов.

Для учета особенностей смыслового содержания ОД приведенную классификацию по уровню фор- мализованности информации целесообразно дополнить классификацией по содержанию ОД [5, 12]. В соответствии с ней информация в автоматизированных системах подразделяется на 4 слоя в соответствии с условиями ее формирования и принципам доступа к информации:

Государственный приватный.

Государственный официальный общего применения.

Коммерческий.

Бесплатный.

Государственный приватный уровень закрыт для открытого использования и содержит сведения ограниченного распространения. Это вызвано политическими, военными, экономическими ограничениями, а также соблюдением приватности юридических и физических лиц (персональные данные). Если допустимо раскрытие информации, она попадает в категорию государственной официальной информация общего применения – сведений, государственных структур, которые нет смысла скрывать.

Для развития информационной инфраструктуры в государстве эта категория данных является наиболее важной, поскольку позволяет связывать все остальные категории сведений с использованием общей се-

мантики, ключей данных, обеспечивает совместимость регламентов и ограничений по работе с информацией. Нужно отметить, что современные реализации технологий ведения ОД официальными структурами не обеспечивают достаточного для разработки сложных приложений уровня регламентации и стандартизации данных.

В процессе деятельности коммерческих организаций формируется коммерческий информационный уровень, используемый как для обеспечения собственной деятельности предприятия, так и для продажи или передачи заказчикам и партнерам произведенных в коммерческих целях информационных продуктов. Уровень бесплатной информации – сведения, получаемые безвозмездно в результате функционирования бесплатных сервисов, в качестве хобби или благотворительности. В то же время информация, полученная бесплатно, может использоваться и в коммерческих целях.

ОД востребованы на всех перечисленных информационных уровнях, кроме первого – государственного приватного. На остальных трех уровнях ОД должны использоваться комплексно и согласованно с тем, чтобы получать максимальный эффект, но избегать связанных с их использованием негативных явлений.

Принципиальным вопросом является обеспечение информационной связности информационных уровней на основе государственного официального, который, по сути, является метаинформацией в социальной системе. Наличие общих классификаторов, онтологий, определений данных и таблиц с данными позволяет согласованно использовать информационную инфраструктуру, развивать ее, привязывая новое информационное содержание к общим смыслам, ключам данных и регламентам ведения информация остальных уровней. Государственный официальный уровень информации нуждается в управлении, и целенаправленном развитии. Это позволит получать дополнительные технологические, экономические и социальные результаты.

Наиболее актуальной открытой информацией для решения финансовых задач являются:

– информация о людях в объеме, не противоречащем закону о персональных данных;

– открытая информация о юридических лицах;

– местоположения и значимые свойства различных объектов (здания, транспортная инфраструктура и пр.) на поверхности Земли, используемые в решении финансовых задач;

– форматы и структурированные описания различных документов для их единообразного использования, например в финансовой отчетности;

– финансовая информация в обобщенной или обезличенной форме, которая может применяться для аналитических целей (статистические данные, объемы продаж по видам товаров и услуг и пр.);

– служебная информация, необходимая для согласованного функционирования и обеспечения взаимно предсказуемого поведения сопрягаемых автоматизированных систем, в частности: справочники и классификаторы, регламенты изменений данных и пр.

Использование ОД в России сейчас особенно актуально в связи с автоматизацией предоставления государственных услуг и участием в этом процессе, в том числе, и ФО. Общедоступный официальный информационный сегмент позволит упростить и ускорить внедрение государственных услуг, а также расширить функциональность имеющихся решений за счет информационных ресурсов коммерческих организаций и участия в процессах разработки новых сервисов социально активных граждан.

Выводы раздела. Использование инфраструктуры ОД открывает новые возможности в обработке данных ФО. Эти данные позволяют связывать информацию, получаемую во внутренних автоматизированных системах с данными во внешних информационных источниках. Поэтому развитие инфраструктуры ОД – это один из основных факторов, влияющих на развитие информационных технологий в ближайшем будущем.

В настоящее время инфраструктура ОД находится в стадии становления, однако уже сейчас возможно решение различных пользовательских и аналитических задач с их использованием. ОД являются информационной базой для предложения ФО перспективных сервисов и услуг, предоставляемых в коммерческом или бесплатном режиме.

ХД ФО как базовая среда информационной интеграции

ХД являются основой для реализации технологий интеллектуального принятия решений в современных ФО. Основные принципы построения ХД рассмотрены, например, в работах [1, 13]. Особенности обработки данных финансовой отчетности отражены в монографии [14].

Эффективность и применимость ХД для ФО может быть оценена рядом его существенных характеристик, к которым можно отнести:

– функциональную полноту ХД при решении задач ФО;

– используемые базовые технологии;

– специализированное программное обеспечение, применяемое для решения задач ФО;

– средства интеграции информации ХД с другими источниками данных, содержащими интересующую информацию.

Важной задачей является разработка перспективной архитектуры ХД.

Традиционно ХД предназначено для размещения больших объемов структурированных сведений с целью последующего анализа и должно осуществлять обработку ретроспективных данных ФО с обеспечением

достаточной отказоустойчивости, необходимого уровня защиты от неправомерного использования, включать средства интеллектуального анализа данных.

Однако в настоящее время появляются источники данных – такие как социальные сети, мультимедиа-данные, сведения о местоположении и перемещении объектов, семантически размеченные финансовые сведения. Поэтому становится актуальной задача использования совместно с информацией ХД гетерогенных внешних данных. Полная интеграция разнотипных данных, как мы уже отмечали ранее, часто невозможна, да и не требуется в практике, однако частичное решение этой задачи востребовано и может быть реализовано.

Рассмотрим ХД в связи с задачами интеграции с другими типами данных. Традиционно ХД – тип базы данных, специализированной на объединении данных из различных источников и последующей ее аналитической обработке. Кроме базы данных, оно включает средства обработки информации, используемые в служебных целях и средства аналитической обработки для клиентов.

В ХД включаются данные из различных источников, обеспечивается их целостность и интегрированность. При этом данные, находящиеся в ХД, как правило, являются неизменяемыми и размещаются в хранилище после завершения транзакции в системе оперативной обработки данных (OLTP-системе) ФО. В ХД размещаются ретроспективные сведения, позволяющие анализировать изменения объектов БД во времени. ХД имеют предметный характер – они касаются некоторой предметной области, работа с которой автоматизируется.

В отличие от OLTP систем, где используются небольшие транзакции, данные в ХД загружаются периодически и включают все изменения в исходной системе за необходимый промежуток времени. Поэтому данные в ХД появляются с задержкой, которая допустима для последующего применения этих данных.

ХД должно обеспечивать оптимальную производительность при выполнении наиболее востребованных типов запросов к данным, необходимых, например, для получения финансовых отчетов. Как правило, структура данных при этом отличается от нормализованного до 3 нормальной формы и выше представления сведений в OLTP системах. Используется структура данных «звезда» или «снежинка». В них таблица фактов окружается таблицами измерений, содержащими сведения о параметрах, характеризующих эти факты.

В условиях современной информационной инфраструктуры параметры, характеризующие факты не ограничиваются сведениями, имеющимися в самом хранилище, а используют результаты подключения внешних информационных источников – других автоматизированных систем и Интернет в целом. На рис. 2 представлен пример традицион-

ной для ХД схемы «звезда» для анализа продаж с указанием на ней возможных подключений внешних источников данных разных типов. На схеме имеется таблица фактов – продажи, а также несколько основных измерений: продукты, клиенты, момент продаж и каналы продаж. Схема дополнена примером возможных связей перечисленных сущностей с внешней информационной средой.

Связь 1 продуктов с ОД позволяет получать сведения о свойствах продукта, унифицированным способом идентифицировать продукт и, как следствие, связать информацию о продукте с информацией о нем в сети Интернет – связь 2. Дополнительная информация о клиентах, может быть получена как с использованием ОД (связь 3), так и непосредственно из сети Интернет (связь 4) и использоваться в процессе анализа. Для идентификации клиентов в рамках нормативных ограничений может применяться ЕСИА (связь 5), а при наличии ограничений по продаже продукта – СМЭВ (связь 6), информация из этих систем также может использоваться в аналитических целях.

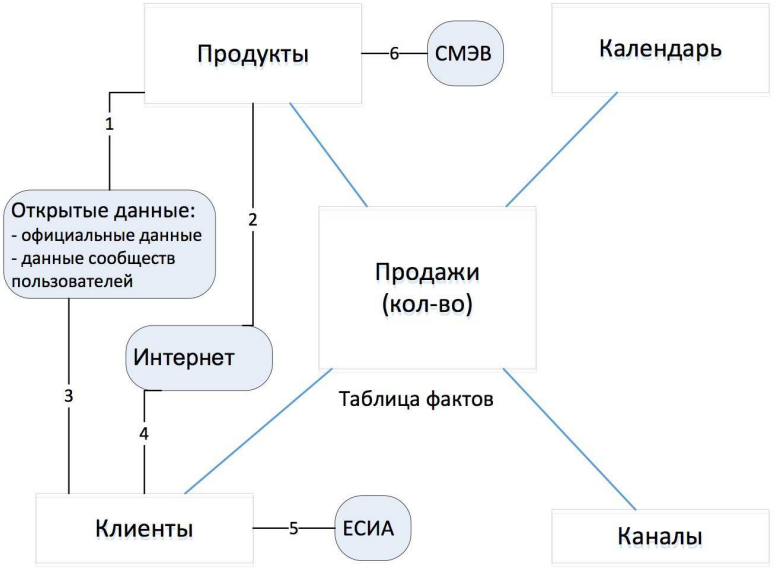


Рис. 2. Типичная схема «звезда» ХД, дополненная внешними источниками данных

Следует отметить, что в случае совместного использования информации ХД и сведений, получаемых динамически из внешних автоматизированных систем, результат анализа оказывается неоднозначным, поскольку изменение внешних данных не всегда регламентировано. Одна-

ко это ограничение для многих аналитических задач является допустимым.

Рассмотрим влияние внешней информации на архитектуру ХД. На рис. 3 представлена типичная базовая архитектура, обеспечивающая функционирование ХД в ФО. Данные, находящиеся в оперативных системах, объединяются по смысловым связям и агрегируются в области предварительной обработки, а затем трансформируются в ХД. При этом формируются метаданные хранилища, в котором сохраняются сведения о формате данных, источниках их возникновения, регламенте обновления и др. Данные в хранилище представляются в виде витрин данных – специализированных баз данных, предназначенных для решения соответствующей задачи. При построении витрин используются внешние данные. Пользователи применяют скомпонованные данные для решения аналитических, отчетных и исследовательских задач.

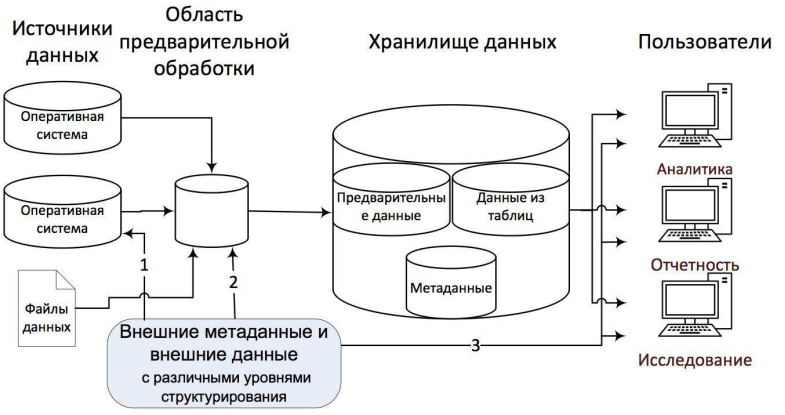


Рис. 3. Влияние внешних информационных ресурсов на базовую архитектуру ХД

Внешняя информация может участвовать в обработке на различных этапах описанного процесса. Ее можно с некоторой долей условности разделить на метаданные и значения данных.

Внешние метаданные в рассматриваемом контексте – это справочники и классификаторы, определения и форматы элементов данных, ключи, связи между сущностями, стандартизованные значения ключевых атрибутов сущностей, описания объектов финансовой инфраструктуры в форме онтологий и другие общепринятые элементы описания данных.

Внешние метаданные могут быть доступны оперативным системам (связь 1 на рис. 3) и таким образом в возможной степени участвовать в

формировании источников данных хранилища. Например, если в различных оперативных системах некоторый продукт описан по единым правилам, то интеграция данных о нем в ХД существенно облегчается.

При предварительной обработке данных для размещения в ХД также возможно подключение внешних метаданных с целью интеграции информации оперативных систем и файлов данных (связь 2 на рис. 3). Однако в этом случае такое преобразование требует использования дополнительных средств согласования информации и в общем случае не является однозначным. На этом этапе возможна также загрузка в ХД внешних данных, которые должны стать его частью и сохраняться постоянно.

Если ХД построено с использованием внешних метаданных, то клиентское программное обеспечение, решающее прикладные задачи может извлекать данные из внешних источников (связь 3 на рис. 3) и они за счет стандартизации метаданных будут структурированы в той степени, с которой это позволяют сделать метаданные, имплицитно встроенные в ХД. Таким образом, клиентское программное обеспечение сможет извлекать меняющееся и не полностью структурированное содержание внешних автоматизированных систем, опираясь на доступное в собственном ХД базовое информационное наполнение.

Выводы раздела

На основании изложенного выше можно сделать вывод о необходимости развития средств для включения в традиционные ХД внешней информации. Эти средства должны предоставлять механизмы для автоматизированного расширения информационного наполнения ХД дополнительной семантической составляющей, получаемой от формирующейся в настоящее время на государственном и коммерческом среды доставки метаинформации заинтересованным потребителям.

Организация обработки новых видов информации ФО и сервисной ИТ-компанией

В деятельности ФО используются сложные и быстро развивающиеся информационные технологии. Как правило, ФО не имеет ресурсов, достаточных для их самостоятельного изучения, кастомизации в процессе решения собственных задач и внедрения. Для реализации этих целей ФО привлекает одну или несколько специализирующихся на таких работах ИТ-компаний. Будем называть такие компании сервисными ИТ-компаниями.

В работе [15] рассмотрены вопросы, связанные с организацией работы сервисной ИТ-компании по разработке, внедрению, развитию и поддержке системы обработки информации ФО

Взаимодействие сервисной ИТ-компании и ФО показано на рис. 4.

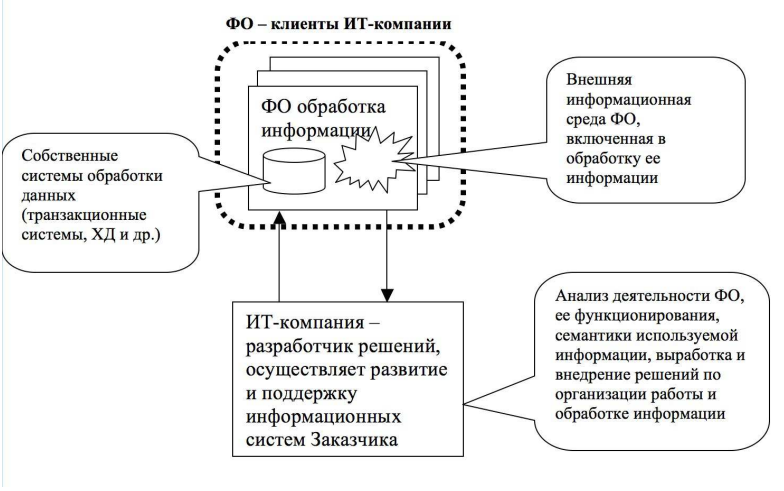


Рис. 4. Сервисная ИТ-компания и ее взаимодействие с ФО

Сервисные ИТ-компании проектируют систему обработки информации ФО, а также осуществляют консалтинг по ее эффективному использованию и развитию. Такие сервисные структуры могут выступать в качестве поставщика решений для информационной интеграции в обслуживаемых ФО для всех используемых ими ИТ-систем. Если отдельные прикладные задачи, бизнес-модели работы, выполняемые ФО не формализуемы и не поддаются полной автоматизации сервисная компания предлагает ФО человеко-машинные подходы к их решению и осуществляет поиск путей более эффективного решения задачи.

Расширяющееся использование внешних по отношению к ФО источников информации требует более формализованного подхода к способам ее получения, проверки полноты и достоверности и использованию в работе. В этих условиях становится особенно актуальным использование модели процессов обработки информации в организации.

ИТ-компания заинтересована в выработке тиражируемого решения, присущего всем ФО и способного настраиваться на особенности работы конкретной организации.

Особенностью деятельности ИТ-компании на финансовом рынке является сложность и «виртуальность» продуктов и услуг предоставляемых обслуживаемыми ФО. Финансовый рынок изменчив: появляются новые концепции работы с информацией, совершенствуется ИТ-инфраструктура, средства разработки информационных систем. В последнее время существенное влияние на деятельность ФО оказывает международная обстановка, поскольку в их деятельности нужно учиты-

вать как изменения в бизнес-процессах, так и риски от использования ИТ-решений, на которые могут распространяться санкции, вызванные политическими причинами.

Как правило, продукты работы ИТ-компании – это инструменты для работы с информацией и информационное наполнение автоматизированных систем в требуемой форме. Информация в автоматизированных системах появляется в соответствии с тем порядком, регламентами, которые приняты в организациях, которые ее создают и используют. Поэтому актуально понимание методов получения и обработки информации, а также оптимизация информационного обмена с целью уменьшения издержек, ускорения и повышения достоверности ее обработки.

Для анализа механизмов функционирования предприятия используются технологии бизнес-моделирования. Информация, полученная в результате этой работы – это метаинформация о функционировании собственного предприятия и контрагентов.

Для решения этой задачи существуют различные средства.

Одним из востребованных в практике финансовых и ИТ-компаний, является процессный подход. Рассмотрим основные организационные, методологические и технологические аспекты внедрения процессного управления в ИТ-компании, которая занимающейся продвижением интеграционных решений на финансовом рынке, а также оказывающей сервисные услуги для ФО. Использование процессного подхода – одна из основных тенденций в совершенствовании управления компаниями в современных условиях [16, 17].

При использовании процессного подхода иерархическая структура управления компанией дополняется моделью. Модель имеет самостоятельное значение для функционирования компании, а также применяется при использовании средств автоматизации для управления компанией. Непосредственное управление заменяется настраиваемым механизмом, поддерживающим выполнение процессов и контролем значений метрик, характеризующих эти процессы. Такой подход позволяет получить дополнительный эффект в работе компании [18, 19]. Можно выделить следующие основные цели моделирования в компании БП:

– Реализация предсказуемости и возможности повторного выполнения всех БП компании.

– Увеличение управляемости компании в процессе ее функционирования.

– Автоматизация процесса управления, оптимизация управленческой структуры и повышение качества управления.

– Реализация измеримости основных процессов работы компании и выполнение регулярной оценки позволяют постоянно совершенствовать текущую работу компании.

– Деятельность каждого сотрудника структурируется и детально оценивается числовыми показателями, что позволяет совершенствовать стратегию развития компании и формировать новые направления ее развития.

– Тиражирование удачных практик работы сотрудников с целью повышения эффективности совместной работы и предсказуемости ее результатов.

В целом использование бизнес-моделирования позволяет решать проблемы формирования компетенций, требуемых для освоения новых направлений деятельности и применения достигнутых компетенций на нужном уровне. Решаются проблемы, связанные с накоплением знаний как компании в целом, так и ее сотрудников в удобной для использования форме.

С помощью процессного управления компания создает собственное «лицо» и позитивный имидж на рынке как надежного и компетентного участника этого рынка. Специализация ИТ-компании на рынке финансовых услуг предполагает создание собственного портфеля востребованных ФО продуктов и услуг, выработку и тиражирование уникальных предложений, предоставления сервисов, востребованных на финансовом рынке. В этом контексте технологию процессного управления во многих случаях целесообразно использовать как внутри ИТ-компании для собственных целей, так и предлагать типовые процессы, обеспечивающие функционирования ФО, в также соответствующие средства автоматизации на рынке как один из ее продуктов.

Технология бизнес-моделирования должна удовлетворять ряду требований, в том числе предоставлять возможность полно и наглядно отражать БП. В процессе формализации БП обычно формируется база данных объектов и их существенных свойств, используемых в моделировании, и составляющих репозиторий системы. Описания объектов поддерживаются ведением классификаторов для всех типов объектов и их элементов для их единообразного понимания, дальнейшего развития, и своевременной модернизации модели.

Все позиции каждого классификатора должна быть подробно описаны, для однозначного восприятия сотрудниками. Как правило, репозиторий включает и позволяет наглядно отображать на соответствующих диаграммах модели следующие базовые классификаторы:

– организационная структура моделируемой компании;

– должности и роли сотрудников компании;

– документы, применяемые в компании;

– продукты, услуги компании, а также их структурированное описание.

Применяемые технологии моделирования должны быть удобны как для восприятия человеком, так и для машинной обработки. Это позво-

ляет обеспечить непрерывное совершенствование способов применения технологии сотрудниками, а также увеличить уровень автоматизации работы с моделью в процессе ее использования.

Средства моделирования БП обычно включают способ наглядного графического отображения (графическую нотацию) процессов удобный для восприятия человеком, позволяющий документировать разработку и обсуждать БП в процессе разработки и использования.

Средства формализованного описания БП обеспечивают однозначность восприятия модели, исключают ее двусмысленность и позволяют применять различные средства автоматизации работы с бизнес-моделью. Имеются технологии, позволяющие с помощью соответствующих программных средств выполнить бизнес-модель и отражать ход выполнения конкретных экземпляров БП, то есть обеспечивающие автоматизированный ввод и последующую обработку получаемой в процессе выполнения БП информации. Такими свойствами обладают, в частности, средства выполнения БП, встроенные в технологии BPM, Workflow.

Рассмотрим коротко распространенные на практике технологии моделирования БП, а также их особенности. Начальным этапом в построении бизнес-модели, как правило, является построение иерархии БП моделируемой организации. Моделирование начинается с формирования иерархического списка БП.

Для каждого элемента списка указывается вход, выход, владелец этого процесса, исполнитель, роль, метрики, документы. На нижнем уровне модели всегда реализуются конкретные операции. На более высоких уровнях иерархии они обобщаются, что позволяет целостно увидеть картину деятельности организации и не пропустить ничего существенного.

В процессе построения иерархии БП бизнес-консультанты выявляют общие и типичные процессы и их элементы, характерные для каждой компании. В результате появляется обобщенная схема декомпозиции БП, в той или иной степени присущая любой компании.

Описанная универсальная схема может использоваться на начальном этапе построения процессной модели любого предприятия с соответствующей сферой деятельности.

Примером стандартизованной реализации такой типовой схемы может служить классификация процессов APQC (American Productivity and Quality Center). Эта структура классификации процессов развивается и в настоящее время, например, она доступна в версии 5.2.0. для предприятий деятельность которых связана с маркетингом и продажами [20]. Классификация отражает универсальную модель процессов, начиная с верхнего уровня. Для ее использования в нашем случае необходимо выполнить модификацию под задачи ИТ-компании, учитывающую

особенности работы на финансовом рынке в ИТ-сфере или отразить в ней задачи моделируемой ФО.

Для этого из стандартизованного классификатора исключаются позиции в иерархии, которые связаны с материальным производством, уточняются и детализируются необходимые пункты иерархии отражающие специфику работы компании, работа которой моделируется. Поскольку структура процессов выверена и обобщена, то, например, для ИТ-компании интересующего нас профиля требуются в основном удаление отдельных неиспользуемых позиций, а также незначительная корректировка названий ряда процессов применительно к особенностям компании. Для построения иерархии БП часто используются электронные таблицы, однако многие средства автоматизации поддержки бизнес-моделирования осуществляют ведение иерархической информации об описываемых БП с использованием собственного репозитория.

Одна из наиболее известных технологий бизнес-моделирования – это SADT (Structured Analysis and Design Technique). Другое название этой технологии – IDEF0 [21].

Исторически это одна из первых широко распространенных технологий моделирования БП, которая возникла в 70-х годах двадцатого века в американской аэрокосмической промышленности. Технология известна на рынке, что является ее преимуществом, поскольку она освоена специалистами-аналитиками, позволяет создавать модели с необходимой степенью детализации.

Однако SADT имеет и недостатки, связанные с трудностью ее восприятия неподготовленным пользователем, а также недостаточными возможностями по автоматизации встраивания бизнес-модели, выполненной в этой нотации в работу компании.

Технология ARIS (Architecture of Integrated InformationSystems) бизнес-моделирования, предложенная ком панией IDS Scheer, является популярной в России технологией. Автором методологии является А. Шеер [22, 23]. В настоящее время технология и программное обеспечение распространяется компанией Software AG [24].

При использовании технологии с помощью поставляемого программного обеспечения возможно построение около 80 видов диаграмм, а также ведение репозитория процессов и объектов, применяемых в диаграммах. Используя информацию репозитория можно сформировать различные нормативные документы и отчеты, выполнять интеграцию с базами данных и приложениями, содержащими соответствующую информацию. Это облегчает поддержку модели организации в процессе эксплуатации.

Технология ARIS встраивается в собственные программные продукты рядом крупных поставщиков ПО, в частности Oracle (Oracle Business Process Analysis Suite) и SAP в поставляемом этой компанией интеграционном решении для продукта SAP R/3.

К недостаткам технологии ARIS можно отнести ее некоторую избыточность и громоздкость, а также сложность для восприятия рядовыми сотрудниками организации. В современных условиях существенным недостаткам системы можно считать проприетарный характер методологии и средств ее автоматизированной поддержки. Однако в практике часто применяется достаточное для решения необходимых задач подмножество этой методологии, что облегчает ее практическое применение.

ARIS также не совсем подходит как средство автоматизации в BPM-системах, для использования в которых модель нужно транслировать в BPMN-нотацию, а также выполнять необходимые ручные доработки.

Унифицированный язык моделирования UML (Unified Modeling Language) является открытым международным стандартом [25]. Он содержит средства моделирования БП наряду со средствами проектирования программного обеспечения и баз данных.Этот язык сложен для освоения и требует специального обучения. В связи с этим его использование для проектирования БП в России не очень распространено. Однако в литературе [26] имеется мнение, что роль его в моделировании БП ввиду открытости и интеграции с популярными средствами разработки ПО будет возрастать. Методология BPMN (Business Process Model and Notation)применяется как специалистами по анализу бизнес-процессов так и бизнес-пользователями,. Так как она стандартизована и формализована, то может применяться для трансляции в исполняемый код на языке BPEL. Одной из целей разработки этой нотации является осуществление связей между описанием БП и автоматизацией их выполнения. Среди специалистов имеется мнение о возможности интеграции в перспективе технологий UML и BPMN, так как они обе разрабатываются консорциумом OMG и решают сходные задачи [27].

Технологии, моделирования БП не исчерпываются перечисленными примерами. Имеется ряд менее известных подходов к решению этой задачи. Часто упрощенные диаграммы БП для решения ограниченных по объему практических задач создаются с помощью универсальных редакторов диаграмм, в частности, MS Visio. Следует отметить, что такие инструменты обеспечивают наглядность отображения БП, однако не обеспечивают ведения репозитория объектов, используемых в модели бизнес-процесса, то есть имеют неполную функциональность. Их можно рекомендовать к использованию на начальных стадиях бизнес-моделирования в качестве средства для удобного редактирования графических диаграмм.

Возникают также собственные нотации моделирования и средства их автоматизации, предлагаемые поставщиками программного обеспечения. Примером является российская разработка Business Studio. В применяемой в этой среде методологии используется собственная удобная графическая среда, которая может применяться как специалистами, так и обычными бизнес-пользователями. Эта технология ориентирова-

на, в том числе, на автоматизированную подготовку регламентирующей документации, необходимой для функционирования компании, а также на создание систем менеджмента качества предприятия. Поскольку система разработана в России, она учитывает особенности организации и документирования работы в нашей стране.

Таким образом, технологии моделирования БП развиваются как средства, обеспечивающие единство языка, взаимодействие сотрудников и позволяющие понимание того, как функционирует предприятие, а при необходимости детализировать нужные БП до операционного уровня. Они также позволяют подготовить регламенты работы предприятия и корпоративные стандарты.

Унифицированные средства бизнес-моделирования позволяют ИТ-предприятию тиражировать опыт обработки бизнес-информации в обслуживаемых ФО, воспринимать обобщенный опыт и использовать типовые бизнес-решения, предлагаемые на рынке.

Эффект от применения технологий моделирования БП для ИТ-компании складывается из повышения эффективности работы собственной компании, а также из обобщения, типизации и автоматизации выполнения БП клиентов, которых она обслуживает. Бизнес-моделирование работы клиентов позволяет полнее выявить имеющиеся у них проблемы, сделать более адресными услуги сервисной ИТ-компании. Оно позволяет повысить качество услуг, а также определить новые потребности ФО и обосновать целесообразность их удовлетворения с использованием средств автоматизации.

Важное значение имеет репозиторий типовых объектов ФО, применяемых в задачах бизнес-моделирования. Такими объектами являются роли и должности сотрудников, справочник подразделений, используемые на предприятии документы и их подробные описания. Рассмотренные выше технологии и инструментальные средства бизнес-моделирования в той или иной степени включают в себя функциональность для разработки и сопровождения репозитория объектов моделирования.

Программные средства для бизнес-моделирования обеспечивают автоматизацию получения и ведения регламентирующих и нормативных документов. С помощью средств автоматизации с использованием модели процессов предприятия и предварительно подготовленных шаблонов осуществляется получение документов, необходимых для работы и сертификации предприятия с целью обеспечения соответствия требованиям стандартов качества, в том числе: должностных инструкций, описаний организационной структуры, ответственности подразделений, регламентов выполнения работ, и других.

Результаты бизнес-моделирования применяются, для решения задач автоматизации при внедрении процессного управления. Поэтому технологии работы с БП часто встраиваются в программные системы,

предназначенные для решения специализированных прикладных задач управления предприятиями [29] таких как CRM, ERP и др.

На основе изучения модели процессов настраиваются и кастомизируются средства автоматизации предприятия. При необходимости оптимизации работы или реорганизации используются результаты ручного, а также автоматизированного реинжиниринга функциональности, автоматизированной на предприятии в процессную модель. При создании модели БП предприятия можно выделить внутренние и сквозные процессы. Разработка и применение каждого из этих типов БП имеет свои особенности.

Внутренние процессы выполняются внутри подразделений, а сквозные – касаются всей компании или проходят через несколько ее подразделений.

Характерные примеры внутренних процессов предприятия: организация продаж; производства; оказание внутренних сервисных услуг ИТ-компании, организованной в виде виртуальных специализированных компаний; учет особенностей рынка, требований к продуктам и способам продаж; использование технологий таких как Jira, Agile в организации труда и командной работы разработчиков.

Сквозные процессы объединяют предприятие, «сшивают» его деятельность. Они объединяют внутренние БП, поэтому их оптимизация значимо влияет на общую эффективность компании. Примерами таких процессов, характерными для большинства предприятий являются:

– стратегическое планирование;

– продажи;

– документооборот;

– разработка новых продуктов и услуг.

Жизненный цикл многих услуг предприятия также является сквозным БП, формирующимся на основе наработанных предприятием типовых элементов, таких как ключевые компетенции, клиентская база и др.

Следует ожидать в перспективе изменений в работе с БП, связанных с новым направлением в развитии ИТ-инфраструктуры – появлением технологий обработки «Открытых данных», которые мы рассматривали ранее. Очевидно, что в будущем в их составе появятся общезначимые сведения по бизнес-моделированию и обобщенным практикам эффективного управления, полученным с их использованием, поскольку у компаний, работающих в сфере высоких технологий формируются типовые БП. В финансовой сфере развитие сообществ, поддерживающих ОД позволит увеличить уровень семантической связности информации, консолидировать сведения о различных аспектах деятельности ФО и финансовых услугах, предлагаемых на рынке.

При наличии соответствующих технологических и экономических условий тиражирование собственных эффективных практик бизнес-

моделирования с использованием инфраструктуры ОД позволит заинтересованным предприятиям реализовать свои конкурентные преимущества, качественно изменить уровень автоматизации бизнеса и повысить эффективность деятельности как собственной компании, так и обслуживаемых ФО в результате тиражирования удачных бизнес-практик.

Пример развития абстрактной компании с использованием процессного подхода к управлению

Рассмотренные выше положения процессного подхода можно применить в деятельности произвольной компании как для организации собственной деятельности, так и в процессах, связанных с обслуживанием ФО. Применительно к компании процессный подход дополнен концепцией виртуальных предприятий, которая предполагает, что основные подразделения компании стали отдельными «предприятиями» имеющими собственный бюджет. Они называются центрами финансовой ответственности (ЦФО). Единые для компании сервисные подразделения, которые сами являются ЦФО, обеспечивают регламентную и бюджетную поддержку других виртуальных компаний. Взаимодействуют ЦФО на внутреннем «рынке», который управляется метриками, имеющими, в том числе, и финансовый характер. Кроме этого учитывается фонд рабочего времени (ФРВ) сотрудников компании.

Процессный подход фактически обеспечивает новый уровень оценки эффективности функционирования подразделений компании.

Результаты работы ЦФО оцениваются с использованием финансовых и нефинансовых показателей. Они характеризуют полный цикл работы ЦФО, включающий, как правило, маркетинг, продажи и производство.

Для взаиморасчетов ЦФО установлены внутренние ставки для оплаты, например, имеется внутренняя ставка по обслуживанию одного сотрудника для услуг бухгалтерии.

Функциональная деятельность ЦФО осуществляется с использованием нефинансовых метрик для управления коммерческими подразделений. Эти метрики позволяют в достаточной степени оценить их основную деятельность. Так, менеджеры по продажам оцениваются, в том числе, по следующим показателям: активность в маркетинговой деятельности; оценка эффективности участия в мероприятиях; количество встреч с заказчиками; использование ФРВ сотрудниками.

Финансовые показатели в целом позволяют сформировать полностью прозрачный механизм, раскрывающий получение и расход средств компании.

Значения показателей для каждого ЦФО вводятся в автоматизированную систему сервисными службами компании. Массив значений получаемых показателей позволяет комплексно анализировать эффективность работы каждого ЦФО и компании в целом.

На рисунке 5 показано взаимодействие между участниками БП компании. Они действуют в рамках ограничений (регламентов, взаимных соглашений), которые обеспечивают выполнение обусловленных процессной моделью управление коммуникаций между подразделениями компании, а также осуществляют внешние связи с клиентами и партнерами.

Использование рассмотренного подхода позволяет осуществлять оперативное управление и анализ состояния компании, адаптировать имеющиеся в компании сервисы в соответствии с требованиями по мере изменения стоящих перед нею задач.

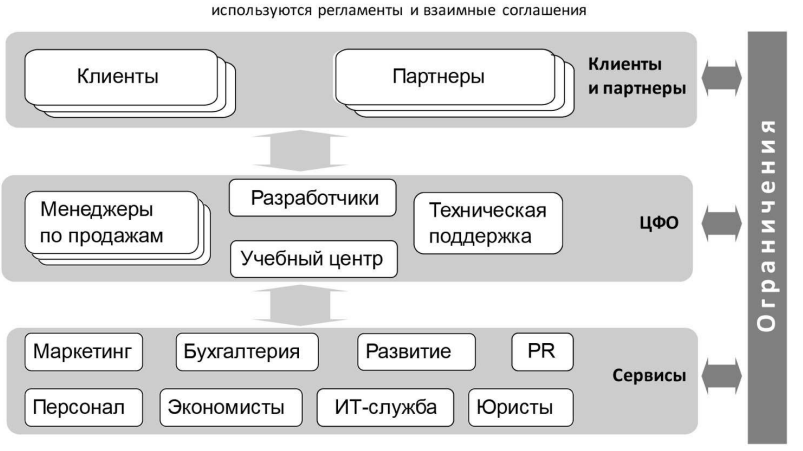


Рис. 5. Взаимодействие между участниками БП компании

Выводы раздела. Использование в работе ИТ-компании процессного подхода позволяет создать в ней эффективные механизмы мониторинга состояния и управления, сократить издержки, повысить управляемость и адаптивность к возникающим изменениям, а накопленный опыт использовать в работе со своими Заказчиками.

Процессный подход позволяет организации адаптироваться к изменениям структуры информации на рынке, включить в регламент обработки финансовой информации внешние информационные ресурсы, а также сориентировать всех участников коммерческой деятельности на достижение конечного результата. Риски в работе распределяются на всех участников коммерческой деятельности.

Горизонтальные связи, формируемые в процессной модели управления, работают более эффективно, чем вертикальные при их правильном применении и настройке. Наличие процессных механизмов управления позволяет использовать новые аналитические возможности на основе измеримых показателей работы.

Заключение

Перспективным направлением автоматизации обработки финансовой информации является комплексное использование всех имеющихся в распоряжении ФО сведений для решения финансовых задач. Нами рассмотрены методы интеграции информации, и способы организации работы ФО и ИТ-компании, решающих задачи финансовой автоматизации. Эти методы могут применяться с использованием доступных средств автоматизации обработки новых видов информации – открытых данных, больших данных, а также сведений, получаемых во взаимодействии с государственными автоматизированными системами, обеспечивающими функционирование систем автоматизации финансовой деятельности.

Использование процессного подхода в процессе обработки финансовой информации позволяет реализовать дополнительные механизмы, обеспечивающие мониторинг работы организации, сориентировать всех участников деятельности компании на достижение конечного результата.

Рассмотренные в работе положения могут применяться на различных этапах разработки и функционирования финансовых автоматизированных систем, а также при выполнении исследований в области методов обработки и использования финансовой информации.

Список литературы

1. Спирли, Э. Корпоративные хранилища данных. Планирование, разработка, реализация. Т. 1: / Э. Спирли; пр. с англ. – М.: «Вильямс», 2001. – 400 с.
2. Волков, Д. Консолидатор 2.0. / Д. Волков // Открытые системы. – 2014. – №9. – С. 1.
3. Гроув, Э. Выживают только параноики: как использовать кризисные периоды, с которыми сталкивается любая компания / Э. Гроув. – М.: Бизнеском, 2011. – 231 с.
4. Волков, А.И. Открытые данные: проблемы и решения / А.И. Волков, Л.А. Рейнгольд // Прикладная информатика. – 2014. – № 3 (51). – С. 5–12
5. Демидов, М. Открытые данные: Россия стоит на низком старте. CNews.ru: статьи / М. Демидов [Электронный ресурс]. URL: http://www.cnews.ru/reviews/index.shtml72013/03/27/523707
6. Открытые государственные данные: российский и зарубежный опыт. Информационный обзор. Серия «Развитие информационного общества и электронного правительства» // Центр технологий электронного правительства НИУ ИТМО. 2012. Вып. 3. 7 с. URL: http://egov.ifmo.ru/files/reviews/eGov_Review_ 2012_03_opendata.pdf
7. Указ Президента Российской Федерации от 7 мая 2012 года № 601 «Об основных направлениях совершенствования системы госу-

дарственного управления» [Электронный ресурс]. URL: http://www.rg.ru/2012/05/09/gosupravlenie- dok.html

8. Сайт Федеральной налоговой службы РФ. Открытые данные. http://www.nalog.ru/opendata/

9. Сайт Федеральной службы государственной статистики РФ. Открытые данные. http://www.gks.ru/opendata/

10. Портал открытых данных Правительства Москвы http://data.mos.ru/

11. Рейнгольд, Л.А. Интеграция информации в социально-экономической системе как основа инновационного развития государства / Е.А. Рейнгольд, Е.А. Рейнгольд, О.А. Славин // Труды ИСА РАН: Методы и модели системного анализа. Оценка эффективности и инвестиционных проектов. Системная диагностика социально-экономических процессов: Т. 61. Вып. 3. – М.: URSS, 2011. – С.76-83

12. Перминов, Г.И. Системы интеллектуального анализа данных (Business Intelligence): учеб.-метод. комплекс / Г.И. Перминов. – М.: ГУ-ВШЭ, 2007 – 121 с.

13. Исаев, Д.В. Автоматизированные системы формирования консолидированной финансовой отчетности: учеб. пособие / Д.В. Исаев, Т.К. Кравченко. – М.: 2006., 370с.

14. Волков, А.И. Методологические и программно-технологические аспекты внедрения процессного управления в ИТ-компании / А.И. Волков // Прикладная информатика. – 2014. – № 2 (50). – С. 6–13

15. Собакарева, А.В. Процессный подход и мероприятия по устранению проблем его внедрения на российских предприятиях / А.В. Собакарева // Вестник МГТУ. – 2008. – Т. 11, № 18. – С. 279–283

16. Комиссарова, М.А. Возможности использования реинжиниринга как основного инструмента управления компаниями с позиций процессного подхода / М.А. Комиссарова // Креативная экономика. – 2011. – № 7 (55). – С. 10–16 URL: http://www.creativeconomy.ru/articles/41

17. Коновалов, С.Н. Моделирование процессов СМК и управление изменениями / С.Н. Коновалов // Материалы с официального сайта «Менеджмент качества» ISO 9000. URL: http://quality.eup.ru/DOCUM4/modelsmk. htm

18. Курьян, А.Г. Реализация процессного подхода в рамках систем менеджмента качества на основе методологии функционального моделирования IDEF0 / А.Г. Курьян, П.С. Серенков // Автоматизация в промышленности. – 2003. – № 3. – С.26-35

19. American Productivity and Quality Center. URL: http://www.apqc.org/

20. Верников, Г. Основные методологии обследования организаций. Стандарт IDEF0 / Г. Верников [Электронный ресурс]. URL: http://www. cfin.ru/vernikov/idef/idef0.shtml

21. Каменова, М. Моделирование бизнеса. Методология ARIS: практическое руководство / М. Каменова, А. Громов, М. Ферапонтов, А. Шматалюк. – М.: «Весть Метатехнология», 2001

22. Шеер, А.В. Бизнес-процессы. Основные понятия. Теория. Методы / А.В. Шеер; пер с англ. – М., 2000. URL: http://www.softwareag.com/ru/

23. Буч, Г. UML. Руководство пользователя / Г. Буч, Дж. Рамбо, А. Джекобсон; пер. с англ. – 2-е изд., стер. – М.: ДМК Пресс; СПб.: Питер, 2004. – 432 с.

24. Репин, В.В. Процессный подход к управлению. Моделирование процессов / В.В. Репин, В.Г. Елиферов. – М.: Стандарты и качество, 2004. – 408 с.

25. Репин, В.В. Обзор практики управления проектами внедрения процессного подхода в российских компаниях / В.В. Репин, А.Ю. Сооляттэ. – М.: Финэкс- перт. ру, 2005. – С.19

26. Система бизнес-моделирования Business Studio. URL: http://www.businessstudio.ru/

27. Панин, А.В. Управление предприятием через идентификацию и классификацию процессов / А.В. Панин // Экономический вестник Ростовского государственного университета. – 2007. – С. 224–229

Кригер А.Б.

МОНИТОРИНГ И АНАЛИЗ ЭКОНОМИЧЕСКОЙ ДЕЯТЕЛЬНОСТИ В ГОСУДАРСТВЕННЫХ УЧРЕЖДЕНИЯХ ИСПОЛНИТЕЛЬСКОГО ИСКУССТВА

«Экономика и социум»: Современные технологии управления организацией №3(12)-2014 г.
Режим доступа:
http://iupr.ru/sovremennye_tehnologii_upravleniya_organizaciyay__3_12__2014_g_/

(статья приводится в сокращении)

Аннотация: *В работе обсуждаются результаты анализа экономической деятельности государственного учреждения исполнительского искусства. Предложен подход совершенствованию системы мониторинга экономической деятельности на основе создания хранилища данных.*

Ключевые слова: *учреждения исполнительского искусства, экономические показатели, индексы Баумоля, эконометрическая модель, структуры данных.*

Споры о том, могут ли учреждения культуры являться экономически эффективными, продолжаются не один десяток лет. Исторический опыт показывает, что характер действия традиционных рыночных регуляторов для сферы культуры и искусства носят ограниченный характер [1].

Для анализа результатов экономической деятельности учреждений культуры и искусства важны два основных элемента производственных отношений: отношений собственности с одной стороны, способов и источников получения доходов [2] с другой стороны.

На сегодняшний день в российской практике для учреждений культуры сложились следующие отношения собственности:

– организации государственной и ведомственной собственности;

– коммерческие и некоммерческие организации, собственниками которых являются учредители.

Если во втором случае эффективность деятельности учреждения оценивают собственники исходя из установленных ими же критериев, то в случае с государственными учреждениями осуществляется контроль за целевым использованием бюджетных средств и мониторинг посещаемости учреждений.

Полагаю, что будет справедливым утверждение, что в российских регионах предприниматели вкладывают свой капитал в учреждения искусства, дающие достаточно быструю окупаемость инвестиций: кинотеатры, танцевальные студии, галереи изобразительного и прикладного искусства и т.п.

Учреждения исполнительского искусства в подавляющем большинстве случаев являются государственными учреждениями, и как следствие финансируются из бюджетов различных уровней. Однако зрелищные учреждения получают собственные доходы от основной (исполнительской) деятельности, от иных разрешенных видов деятельности.

Целью данного исследования является анализ закономерностей для экономических показателей деятельности государственного учреждения исполнительского искусства, выявление факторов влияющих на экономические показатели и характеристик связанные с получением дохода от всех видов деятельности.

Анализ проводился на примере данных регионального государственного драматического театра (Приморский край). В качестве исходных данных использовались открытые отчеты федерального статистического наблюдения (форма П-4, форма № 9-НК). Кроме того, использовались данные статистического наблюдения по региональной экономике.

Так как основным методом исследования является метод прикладной статистики, то для моделирования использовались предварительно

обработанные исходные данные: темпы изменения показателей, индексы и т.п. Таким образом, непосредственно данные отчетов в работе не приводятся и не анализируются.

Следует оговориться, что непосредственно не ставилась цель анализировать влияние объемов бюджетного финансирования на результаты деятельности театрального предприятия.

Отчеты статистического федерального наблюдения являются сводными регистрами мониторинга деятельности учреждений культуры. Указанные отчеты дают в распоряжения аналитика следующие показатели:

– общую характеристику эксплуатируемых зданий и сооружений, включая количество площадок, их вместимость, эксплуатационное состояние;

– число постановок, включая новые и возобновленные постановки;

– статистические данные по персоналу, включая численность, загрузку (в человеко-часах), размер фонда заработной платы по разным типам персонала;

– показатели, характеризующие основную деятельность, включая число мероприятий, численность зрителей, поступления (выручка) от мероприятий;

– размер бюджетного финансирования и структуру расходов.

Данные статистического наблюдения позволяют сформировать динамические ряды экономических показателей деятельности учреждения культуры. Для целей анализа в данной работе использованы: динамика численности персонала, динамика заработной платы, динамика цен на билеты, динамика доходов театра, динамика дефицита доходов театра, динамика производительности труда в театре, динамика числа зрителей. Дополнительно рассматривались следующие показатели: число зрителей, структура репертуара, структура персонала.

Задачами исследования динамики данной задачи являются: оценка устойчивости тенденции, оценка присутствия сезонных колебаний, анализ характера колебаний показателей, сравнение динамики экономических показателей для театра и для региональной экономики.

Анализируемые динамические ряды представляют собой базисные изменения и месячные темпы изменения экономических показателей. Используются наблюдения за три календарных года.

Расчет автокорреляционной функции (далее ACF) темпов динамики показали следующие результаты.

ACF темпов доходов театрального предприятия указывает на отсутствие устойчивой тенденции (быстрый спад автокорреляционной функции) и случайный характер колебаний. Анализ месячного темпа доходов подтверждает выводы. ACF месячного темпа доходов однозначно указывает на случайность колебаний показателя. ACF темпа заработной платы указывает на отсутствие тенденции (значения ACF ≈ 0)

и случайный характер колебаний. ACF темпа изменения производительности труда указывает на неустойчивой тенденции (быстрый спад автокорреляционной функции) и слабовыраженную сезонность колебаний.

Из представленных результатов следует однозначный вывод: экономические показатели деятельности театрального предприятия не имеют выраженной устойчивой тенденции изменения. Колебания практически всех показателей (исключая текущие расходы) являются случайными.

В то же время сравнение с данными по региональной экономике указывает, что динамика показателей экономической деятельности театрального предприятия имеет принципиально иной характер. Для региональной экономики характерна устойчивая тенденция изменения аналогичных показателей, слабые колебания, отсутствие влияния сезонного фактора. Пример – данные по динамике средней заработной платы по региональной экономике и в учреждении культуры (рис. 7)

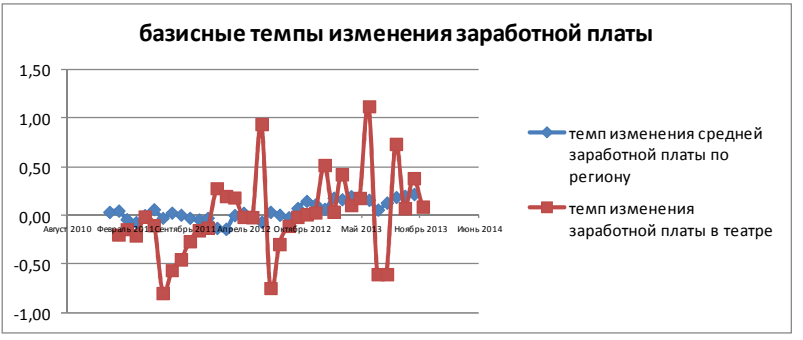


Рис. 7[1] .Темп изменения уровня оплаты труда в региональном учреждении культуры и по региональной экономике в целом

С точки зрения дальнейшего моделирования случайность колебаний экономических показателей является положительным фактом. В данном случае, динамический ряд можно использовать для построения модели регрессии, т.к. статистическая связь между факторами однозначно не является следствием существующей тенденции. Однако анализ матрицы коэффициентов корреляции показывает, что доступные для анализа закономерностей показатели слабо связаны результативным признаком (темпом изменения производительности), см. таблица 1.

Попытка провести анализ с дополнительными показателями, взятыми за год, а именно объемом коммерческой выручки, бюджетного

[1] Нумерация соответствует полному тексту статьи

финансирования, числа зрителей, числа спектаклей по типам, не дали утешительного результата. Факторы либо слабо связаны с признаком (объемом коммерческой выручки), либо сильно коррелированы друг с другом. Более всего в полученных результатах озадачивает отсутствие корреляционной связи между числом зрителей и объемом коммерческой выручки ($\rho_{yx} = 0{,}14$).

Таблица 1

Матрица коэффициентов корреляции

Показатель	Динамика производительности	Динамика расходов	Динамика среднемесячной зар / платы	Количество спектаклей	Количество зрителей, тыс. чел.
Динамика Производительности	1,00				
Динамика расходов	0,35	1,00			
динамика Среднемесячная зарплаты	0,31	-0,07	1,00		
Количество спектаклей, шт.	0,31	0,16	-0,13	1,00	
Количество зрителей, тыс. чел.	0,29	0,10	-0,18	0,82	1,00

Полученные неубедительные результаты привели автора исследования к идее анализа соотношения производительности государственного театра и средней производительности труда в экономики.

Теоретической основой проведенного моделирования являются исследования А.Я. Рубинштейна [2, 3] для российских учреждений культуры. В основу указанных исследований положена идея классической работы В. Баумоля и В. Боуэна, так называемой «болезни цен». Суть явления состоит отставание производительности труда в организациях исполнительских искусств от динамики средней производительности в экономике.

Оценка эконометрической модели «болезни цен», построенная в [3], подтвердила отставание производительности труда в учреждениях искусства в целом по Российской Федерации.

Задачей автора данной работы была проверить, насколько закономерность подтверждается на региональном уровне, эффективна ли единая система мониторинга деятельности исполнительских учреждений.

Рассмотрим экзогенные и эндогенные переменные модели. Эндогенной переменной является общий индекс Баумоля B, измеряющий темп годового (при наличии данных, месячного) прироста дефицита дохода на одно посещение. При этом отрицательные значения общего индекса Баумоля $B(B<0)$ указывают на сокращение удельного дефицита дохода и наоборот его положительные значения ($B>0$) свидетельствуют об увеличении негативных последствий «болезни цен».

Экзогенные переменные индексы B_1, B_2, B_3, характеризуют показатели экономической деятельности. «Индекс Баумоля» B_1, равен темпу годового (месячного) прироста отношения производительностей в учреждении исполнительского искусства к средней производительности труда в экономике региона. «Индекс Баумоля» B_2, соответствует темпу годового прироста отношения среднемесячной заработной платы в организациях искусства к средней заработной плате в экономике. При этом положительные значения B_2 ($B_2>0$) указывают на опережающий рост заработной платы в театре, отрицательные значения ($B_2<0$) – на ее отставание. «Индекс Баумоля» B_3, измеряет темп годового (месячного) прироста отношения текущих цен театра к общему уровню инфляции. Положительные значения этого индекса ($B_3>0$) свидетельствуют о сверхинфляционной динамике цен в театре, отрицательные значения ($B_3<0$) – об их отставании от роста цен в экономике.

Указанные индексы Баумоля B_1, B_2, B_3 являются достаточно удобными динамическими показателями, позволяющими «мониторить» деятельность театра, анализировать закономерности изменения дефицита доходов. Вид модели $B=\omega_0+\omega_1 B_1+\omega_2 B_2+\omega_3 B_3+\xi$.

Рассмотрим результаты оценивания модели. Проводим предварительное тестирование с помощью теста Дики-Фулера для индексов Баумоля. Тест Дики-Фулера – тест на единичные корни. Сущность теста – проверяется гипотеза о том, что динамический ряд является стационарным в широком смысле, не является «случайным блужданием». Проведенные расчеты показали (табл. 2), что гипотеза на единичные корни может быть отвергнута для всех временных рядов. Только для частного индекса Баумоля B3 уровень значимости составляет 6%. Но данный уровень не является критичным для практических задач. Кроме того, при рассмотрении в тесте модели с константой уровень значимости улучшается до 1%.

Таблица 2

Результаты проверки на стационарность временных рядов

Показатель	Отвержение нулевой гипотезы о существовании единичного корня Augmented Dickey-Fuller test (Prob)
Общий индекс Баумоля *B*	0,001
Частный индекс Баумоля» B_1	0,012
Частный индекс Баумоля» B_2	0,034
Частный индекс Баумоля» B_3	0,06

Следующим шагом является оценка эконометрической модели. Оценивание модели реализовано в трех вариантах. Первый вариант – модель, построенная на исходных данных без предварительной обработки. Вторая модель – из выборок исключены наблюдения соответствующие точкам с уровнем доходов выходящих за пределы доверительного интервала. В третьем случае, из значений доходов исключена дополнительная субсидия из регионального бюджета. Результаты моделирования представлены в табл. 3.

Таблица 3

Оценка регрессионной зависимости удельного дефицита доходов

Переменные уравнения регрессии	Коэффициенты регрессии /значимость	Коэффициенты регрессии /значимость	Коэффициенты регрессии /значимость
Общий индекс Баумоля *B*	Модель 1	Модель 2 (урезанная выборка)	Модель (корректированные данные)
Частный индекс Баумоля B_1	5,01 / 0,06	3,26 / 0,30	4,73 / 0,07
Частный индекс Баумоля B_2	-3,49 / 0,42	-1,20 / 0,80	-3,19 / 0,45
Частный индекс Баумоля B_3	-9,47 / 0,05	-14,18 / 0,01	-10,06 / 0,03
Свободный член	-0,81 / 0,65	0,05 / 0,98	-0,70 / 0,69
Число наблюдений	24	21	24
R^2	32	40	33
Adjusted R^2	22	30	23

Положительные значения коэффициента для частного индекса B_1 подтверждает снижение удельного дефицита дохода театрального учреждения при условии снижения отставания производительности театра от производительности в экономике.

Отрицательные значения коэффициента для переменной B_2 (значения $B_2>0$) приводят к уменьшению удельного дефицита дохода театра при отставание роста заработной платы в театре, по отношению к средней заработной плате по экономике.

Отрицательные значения коэффициента для B_3 говорят о том, что рост цен на билеты выше уровня инфляции приводит к снижению удельного дефицита доходов театрального предприятия.

Полученные результаты по уровню статистической значимости и по уровню коэффициента детерминации сопоставимы с результатами исследования А.Я. Рубинштейна [3]. Однако закономерности существенно отличаются от общероссийских.

Проведенные по всем трем вариантам расчеты позволяют сделать вывод о подтверждении предположения модели. Однако темп годового прироста заработной платы не влияет на темп изменения дефицита доходов в театре. Данный вывод существенно отличается от результатов моделирования на общероссийских данных [3, 4]. При этом очевидно, что корректировка данных существенного улучшения модели не дает.

Полученные результаты моделирования в большей степени порождают новые вопросы, нежели дают рекомендации по совершенствованию экономической деятельности зрелищных учреждений.

Укажем на наиболее актуальные. Непонятными остаются причины резких изменений темпов экономических показателей театра. Исходные данные не позволяют получить объяснение, почему статистически не связаны численность зрителей и коммерческая выручка. Здесь можно предположить, что реально на выручку влияет заполняемость залов. Эконометрическая модель указывает на положительное влияние опережения уровня цен на билеты от уровней цен по экономике. Однако, практически постоянная численность зрителей в «сезон», наводит на мысль, что «потребителями» культурных благ является достаточно узкая группа населения. В этом случае увеличение уровней цен на билеты едва ли приведёт к росту посещаемости, а, следовательно, к росту доходов.

Располагаемые данные статистических наблюдений не позволяют разрешить указанные вопросы. Очевидна непрозрачность и недостаточность данных мониторинга для выявления закономерностей экономического функционирования.

С точки зрения автора целесообразно изменить систему мониторинга. Для этого достаточно создать структуру данных, которая, основываясь на принятой системе статистических наблюдений, давала бы

возможность актуализации наблюдений, визуализации в удобной форме, расчетов дополнительных показателей.

Такой структурой может быть реляционное хранилище данных. Хранилища данных, в отличии от баз данных, позволяют консолидировать наблюдения из различных источников. Это и создает условия для актуализации и оперативной обработки. Вариант логической модели реляционного хранилища данных представлен на рис. 8.

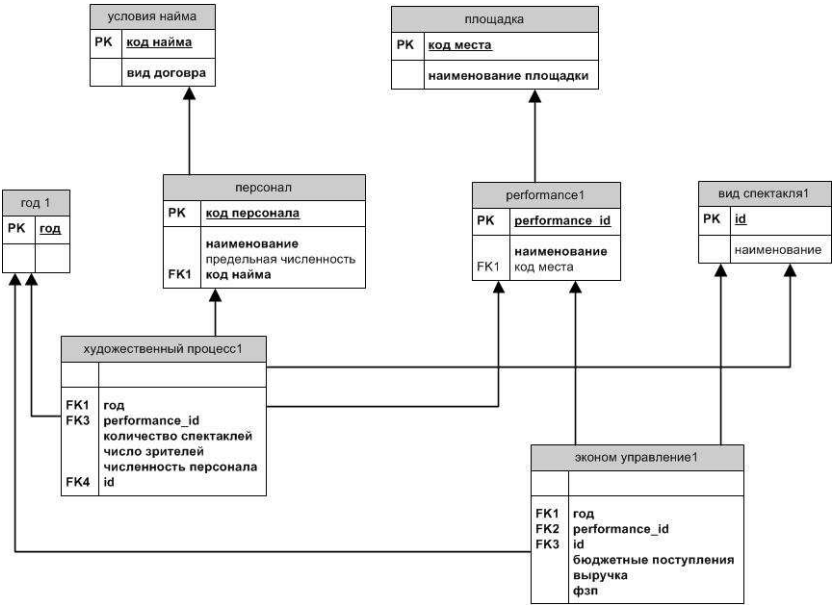


Рис. 8. Логическая модель хранилища данных, архитектура «снежинка»

Практическая реализация хранилища данных в более простом варианте (архитектура «звезда») выполненная выпускницей Школы «Экономики и менеджмента» ДВФУ 2014 г. М.О. Рязановой показала удобство использования такой структуры данных для практических специалистов.

Список литературы

1. Тарханова, Е.Г. Методы оценки эффективности деятельности некоммерческих организаций / Е.Г. Тарханова // Известия ИГЭА. – 2011. – № 4 (78) – С. 110–114

2. Рубинштейн, А.Я. К теории рынков опекаемых благ. Ст. 1. Опекаемые блага и их место в экономической теории / А.Я. Рубинштейн // Общественные науки и современность. – 2009. – № 1. – С. 139–153.

3. Рубинштейн, А.Я. «Болезнь Баумоля» в сфере культуры: Опыт эконометрического исследования / А.Я. Рубинштейн // Художественная культура. – 2012. – №3. – С. 8–10.

4. Рубинштейн А.Я. Эконометрический анализ опекаемых благ в сфере культуры / А.Я. Рубинштейн // Отчет XIII Апрельской международной научной конференции по проблемам развития экономики и общества. – М.: ГУ-ВШЭ, 2012.

4. АРХИТЕКТУРЫ ХРАНИЛИЩ ДАННЫХ

Сабир Асадуллаев

АРХИТЕКТУРЫ ХРАНИЛИЩ ДАННЫХ – 1

Об авторе: Сабир Асадуллаев, исполнительный архитектор
SWG IBM EE/A, IBM,
Режим доступа:
https://www.ibm.com/developerworks/ru/library/sabir/axd_1/index.html#ico
mments

Эта работа открывает цикл из трех статей, посвященных архитектурам хранилищ данных (ХД) и их предшественников. Обилие различных подходов, методов и рекомендаций приводят к некоторой путанице понятий, достоинств, недостатков и границ применимости тех или иных архитектурных решений. В первой статье рассмотрены эволюция понимания места OLAP, компоненты архитектуры ХД, виртуальные ХД и независимые витрины данных. Вторая публикация 1 посвящена таким архитектурам, как централизованное ХД с системой извлечения, преобразования и загрузки данных (ETL), хранилище данных с системой извлечения, загрузки и преобразования данных (ELT), ЦХД с оперативным складом данных (ОСД), расширенная модель с витринами данных (ВД). В третьей статье 2 рассмотрены хранилище с накоплением данных в ВД, централизованная ETL с параллельными ХД и ВД, ХД с интеграционной шиной, а также рекомендованная архитектура хранилища данных.

OLAP и OLTP

Любая транзакционная система, как правило, содержит два типа таблиц. Один из них отвечает за быстрые транзакции. Например, при продаже билетов необходимо обеспечить работу большого числа кассиров, которые обмениваются с системой короткими сообщениями. Действительно, вводимая и распечатываемая информация, касающаяся фамилии пассажира, даты вылета, рейса, места, пункта назначения, может быть оценена в 1000 байт. Таким образом, для обслуживания пассажиров необходима быстрая обработка коротких записей. Другой тип таблиц содержит итоговые данные о продажах за указанный срок, по направлениям, по категориям пассажиров. Эти таблицы используются аналитиками и финансовыми специалистами раз в месяц, или в конце года, когда необходимо подвести итоги деятельности компании. И если количество аналитиков в десятки раз меньше числа кассиров, то объемы данных, необходимых для анализа, превышают размер средней транзакции на несколько порядков величины. Естественно, что во время вы-

полнения аналитических работ время отклика системы на запрос о наличии билета увеличивается. Создание систем с резервом вычислительной мощности может сгладить негативное воздействие аналитической нагрузки на транзакционную активность, но приводит к значительному удорожанию комплекса, при том, что избыточная мощность большую часть времени остается невостребованной. Вторым фактором, приведшим к разделению аналитических и транзакционных систем, являются разные требования, которые предъявляют аналитические и транзакционные системы к вычислительным комплексам.

История OLAP начинается в 1993, когда была опубликована статья 3 «Обеспечение OLAP (оперативной аналитической обработки) для пользователей – аналитиков». Первоначально казалось, что разделения транзакционных и аналитических систем (OLTP – OLAP) вполне достаточно.

Однако вскоре выяснилось, что OLAP – системы очень плохо справляются с ролью посредника между различными транзакционными системами – источниками данных и клиентскими приложениями.

Стало ясно, что необходима среда хранения аналитических данных. И поначалу на эту роль претендовали единые базы данных, в которые предлагалось копировать исходную информацию из источников данных. Эта идея оказалась не вполне жизнеспособной, поскольку транзакционные системы разрабатывались, как правило, без единого плана, и содержали противоречивую и несогласованную информацию.

Рис. 1. Эволюция понимания места OLAP в архитектуре

Так появились хранилища данных, предназначенные для надежного хранения информации, и системы извлечения, очистки и загрузки данных. OLAP-системы работали поверх хранилищ данных.

Вскоре выяснилось, что хранилища данных накапливают настолько важную для организации информацию, что всякий несанкционированный доступ в хранилище чреват серьезными финансовыми потерями. Кроме того, ориентированные на надежное хранение форматы данных плохо сочетаются с требованиями быстрого информационного обслуживания. Территориальная распределенность и организационная структура предприятия также требуют специфического подхода к информационного обслуживания каждого подразделения. Решением является витрины данных, которые содержат необходимое подмножество информации из хранилища. Наполнение витрин из хранилища может происходить в часы спада активности пользователей. В случае сбоя информация может быть легко восстановлена из хранилища с минимальными потерями.

Витрины данных могут обслуживать задачи отчетности, статистического анализа, планирования, сценарных расчетов, и, в том числе, многомерного анализа (OLAP). Таким образом, системы OLAP, первоначально претендовавшие на роль чуть ли не половины вычислительного мира (отдавая вторую половину OLTP системам), в настоящее время занимают место аналитических средств уровня рабочих групп.

Шесть уровней архитектур хранилища данных

Архитектура хранилищ данных иногда напоминает детскую игру в кубики. Как их ни сложи, все равно получается нечто, что можно встретить в реальной жизни. Иной раз в организации можно обнаружить наличие нескольких корпоративных хранилищ данных, каждое из которых позиционируется как единый и единственный источник непротиворечивой информации.

Еще забавнее многослойные витрины данных при наличии единого хранилища. Почему нельзя построить новую витрину поверх хранилища? Видите ли, пользователям захотелось объединить некоторые данные из двух витрин в третью. Это, может быть, имело бы смысл, если бы в витринах содержалась информация, которой нет в хранилище, например, если бы пользователи обогащали витрину своими расчетами и данными. Даже если так, то какова ценность этих обогащенных данных по сравнению с теми, что прошли через сито очистки в соответствии с корпоративными правилами? Кто отвечает за качество этих данных? Как они появились в системе? Никто не знает, но всем хочется получить доступ к информации, которой нет в хранилище.

Хранилища данных чем-то похожи на системы очистки воды. Вода собирается из разных источников с различным химическим составом. Поэтому в каждом конкретном случае применяются свои методы очистки и обеззараживания воды. Вода, отвечающая строгим стандартам качества, поступает к потребителям. И как бы мы ни жаловались на качество воды, именно такой подход предотвращает распространение эпидемий в боль-

шом городе. И никому не приходит в голову (я очень на это надеюсь) очищенную воду обогащать водой из ближайшей лужи. Но в ИТ свои законы.

В дальнейшем будут рассмотрены различные архитектуры хранилищ данных, кроме совсем экзотичных вариантов.

Мы будем рассматривать архитектуры корпоративного хранилища данных на шести уровнях, так как, несмотря на то, что сами компоненты могут отсутствовать, уровни в том или ином виде сохраняются.

Рис. 2. Шесть уровней архитектуры хранилища данных

Первый уровень представлен источниками данных, в качестве которых выступают транзакционные и унаследованные системы, архивы, разрозненные файлы известных форматов, документы MS Office, а также любые иные источники структурированных данных.

На втором уровне размещается система извлечения, преобразования и загрузки данных (ETL – Extract, Transformation and Load). Основная задача ETL – извлечь данные из разных систем, привести их к согласованному виду и загрузить в хранилище. Программно-аппаратный комплекс, на котором реализована система ETL, должен обладать значительной пропускной способностью. Но еще важнее для него – это высокая вычислительная производительность. Поэтому лучшие из систем ETL способны обеспечивать высокую степень параллелизма вычислений, и даже работать с кластерами и вычислительными гридами.

Роль следующего уровня – надежное, защищенное от несанкционированного доступа, хранение данных. В соответствии с предлагаемой тройной стратегией 4, мы полагаем, что на этом уровне должны размещаться также системы ведения метаданных и нормативно-справочной

информации (НСИ). Оперативный склад данных (Operational Data Store) необходим тогда, когда требуется как можно более оперативный доступ к пусть неполным, не до конца согласованным данным, доступным с наименьшей возможной задержкой. Зоны временного хранения (Staging area) нужны для реализации специфического бизнес – процесса, например, когда перед загрузкой данных контролер данных должен просмотреть их и дать разрешение на их загрузку в хранилище.

Иногда зонами временного хранения называют буферные базы данных, необходимые для выполнения внутренних технологических операций, например, ETL выбирает данные из источника, записывает их во внутреннюю БД, обрабатывает и предает в хранилище. В данной работе под зонами временного хранения понимаются области хранения данных, предназначенные для выполнения операций внешними пользователями или системами в соответствии с бизнес требованиями обработки данных. Выделение зон временного хранения в отдельный компонент ХД необходим, так как для этих зон требуется создание дополнительных средств администрирования, мониторинга, обеспечения безопасности и аудита.

Информационные системы на уровне распределения данных все еще не имеют общепринятого названия. Они могут называться просто ETL, так же, как и система извлечения, преобразования и загрузки данных на втором уровне. Или, чтобы подчеркнуть отличия от ETL, их иногда называют ETL-2. При этом системы уровня распределения данных выполняют задачи, значительно отличающиеся от задач ETL, а именно, выборку реструктуризацию и доставку данных (SRD – Sample, Restructure, Deliver) ETL извлекает данные из множества внешних систем. SRD выполняет выборку из единого хранилища данных. ETL получает несогласованные данные, которые надо преобразовать к единому формату. SRD имеет дело с очищенными данными, структуры которых должны быть приведены в соответствие с требованиями различных приложений. ETL загружает данные в центральное хранилище. SRD должно доставить данные в различные витрины в соответствии с правами доступа, графиком доставки и требованиями к составу информации.

Уровень предоставления данных предназначен для разделения функций хранения и функций обслуживания различных задач. Витрины данных должны имеет структуры данных, максимально отвечающие потребностям обслуживаемых задач. Поскольку не существует универсальных структур данных, оптимальных для любой задачи, витрины данных следует группировать по территориальным, тематическим, организационным, прикладным, функциональным и иным признакам.

Уровень бизнес-приложений представлен сценарными расчетами и статистическим анализом, многомерным анализом, средствами планирования и подготовки отчетности. Естественно, что список бизнес-приложений этим не исчерпывается.

Виртуальные хранилища данных

Виртуальные хранилища данных остались в той романтической эпохе, когда казалось, что можно реализовать все, что ни измыслит мозг человеческий. О них уже никто не помнит, и потому вновь и вновь изобретают их, правда, на новом уровне. Поэтому просто необходимо начать с того, чего уже давно нет, но пытается возродиться в новом обличье. Идея создания виртуальных хранилищ основывалась на нескольких возвышенно-прекрасных идеях.

Первая идея – это сокращение расходов. Нет необходимости тратить средства на дорогостоящее оборудование для центрального хранилища данных. Не надо содержать высококвалифицированный персонал, обслуживающий это хранилище. Не нужны серверные помещения с дорогостоящим оборудованием систем охлаждения, пожаротушения и мониторинга.

Вторая идея – надо работать с самыми свежими данными. Аналитические системы должны напрямую работать с источниками данных, минуя всех посредников. Посредники – это зло, это все знают. У наших экспертов нет доверия к программам-посредникам. Эксперты всегда работали напрямую с данными исходных систем.

Третья идея – мы сами все напишем. Все, что нужно – это рабочая станция и доступ к источникам данных. И еще компилятор. Наши программисты все равно сидят без дела. Они разработают программу, которая по запросу пользователя будет сама обращаться ко всем источникам, сама будет доставлять данные на пользовательский компьютер, сама будет преобразовывать несовпадающие форматы, сама будет выполнять анализ данных, и сама же покажет все на экране.

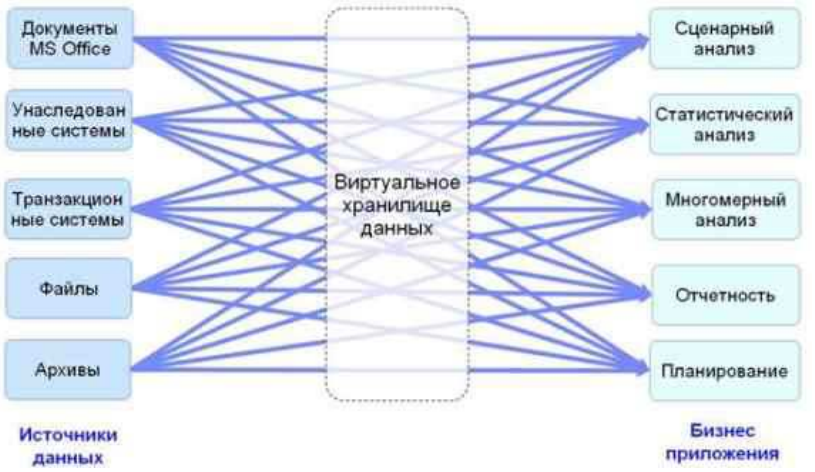


Рис. 3

В компании много разных пользователей с разными нуждами? Ничего страшного, мы модифицируем нашу универсальную программу под столько вариантов, сколько требуется.

Появился новый источник данных? Все замечательно. Мы перепишем все наши программы с учетом особенностей этого источника.

Изменился формат данных? Прекрасно. Мы перепишем все наши программы с учетом нового формата.

Все хорошо, все при деле, надо расширять отдел программирования.

Да, еще пользователи систем-источников данных жалуются, что с некоторых пор их системы очень медленно работают по той причине, что при каждом, даже повторном запросе наше универсальное клиентское приложение снова и снова обращается к источнику данных. Поэтому надо приобрести новые, более мощные серверы. Где сокращение расходов? Его нет. Наоборот, расходы только растут. Нужно больше разработчиков, больше серверов, больше электричества, больше площадей под серверные помещения.

Может, все же есть выгода от такой архитектуры?

Мы получили жесткую связь между источниками данных и аналитическими приложениями. Любое изменение в источнике данных должно согласовываться с разработчиками универсального клиента с тем, чтобы избежать передачи искаженных и неверно интерпретируемых данных в аналитические программы. На каждом рабочем месте необходимо поддерживать набор интерфейсов доступа к различным системам – источникам.

Иногда говорят, что все это очевидно и не стоит тратить время на разъяснение того, что и так всем понятно. Но почему эти же разработчики на запрос пользователя «Мне нужны данные из витрин А, Б и В» пишут клиентское приложение, которое обращается к сразу к нескольким витринам, вновь и вновь воспроизводя умершую архитектуру виртуального хранилища данных?

Независимые витрины данных

Независимые витрины данных появились как физическая реализация понимания того, что транзакционная и аналитическая обработка данных плохо уживаются на одной ЭВМ. Причины несовместимости заключаются в следующем:

Для транзакционной обработки характерно большое количество чтений и записей в базу данных. Аналитическая обработка может потребовать всего несколько обращений к БД.

Длина записей в OLTP обычно не превышает 1000 символов. Аналитический запрос может потребовать мегабайты данных за одно обращение для анализа.

Количество пользователей транзакционной системы может достигать несколько тысяч человек. Число аналитиков обычно в пределах нескольких десятков.

Характерными требованиями для транзакционных систем является круглосуточная бесперебойная работа 365 дней в году (24 х 365). Аналитическая обработка не выдвигает столь четко сформулированных требований к готовности аналитических комплексов, но не подготовленная в срок отчетность может привести к серьезным неприятностям, как для аналитиков, так и для предприятия.

Нагрузка на транзакционные системы распределяется более или менее равномерно во времени. Нагрузка на аналитические системы, как правило, максимальна в конце отчетных периодов (месяца, квартала, года).

Транзакционная обработка осуществляется, в основном над текущими данными. Аналитические вычисления производятся над историческими данными.

Данные в транзакционных системах могут обновляться, тогда, как в аналитических системах данные должны только добавляться, и попытка внесения изменений задним числом должна вызывать, по меньшей мере, настороженность.

Таким образом, транзакционные и аналитические системы выдвигают разные требования к программно-аппаратному комплексу в части производительности, пропускной способности, доступности комплекса, моделям данных, организации хранения данных, способов доступа к данным, пиковым нагрузкам, объемам обрабатываемы данных и методам обработки.

Создание независимых витрин было первой реакцией на необходимость разделения аналитической и транзакционных систем. В те времена это был большой шаг вперед, упростивший проектирование и эксплуатацию программно-аппаратных комплексов, так как не надо было пытаться удовлетворить взаимоисключающим требованиям аналитических и транзакционных систем.

Преимуществом создания независимых витрин является легкость и простота их организации, так как каждая из них оперирует с данными одной задачи, и поэтому не возникает проблем с метаданными и НСИ. Нет никакой необходимости в сложных системах извлечения, преобразования и загрузки данных (ETL). Данные просто копируются на регулярной основе из транзакционной системы в витрину данных. Одно приложение – одна витрина. Поэтому независимые витрины данных часто называют прикладными витринами данных. Но что делать, если пользователям нужно использовать информацию из нескольких витрин одновременно? Разработка сложных клиентских приложений, способных обращаться ко многим витринам и на лету преобразовывать данные, уже была скомпрометирована виртуальными хранилищами данных.

Рис. 4. Независимые витрины данных

Значит, нужен единый репозиторий – хранилище данных. Но информация в витринах не согласована. Каждая витрина унаследовала от транзакционной системы свою терминологию, свою модель данных, свою нормативно-справочную информацию, в том числе, кодировку данных. Например, в одной системе дата выполнения операции может быть закодирована в российском формате ДД.ММ.ГГГГ (день, месяц, год), а в другой в американском формате ММ.ДД. ГГГГ (месяц, день, год). Значит, при слиянии данных необходимо понимать, что означает дата 06.05.2009 – это 5 июня, или 6 мая. Итак, нам нужна система извлечения, преобразования и загрузки данных.

Таким образом, все преимущества независимых витрин данных исчезают при первом же требовании пользователей работать с данными из нескольких витрин.

Заключение

В статье рассмотрены эволюция понимания места OLAP, компоненты архитектуры ХД, виртуальные ХД и независимые витрины данных. В двух следующих публикациях будут обсуждены достоинства и ограничения следующих архитектур: централизованное ХД с системой извлечения, преобразования и загрузки данных (ETL), хранилище данных с системой извлечения, загрузки и преобразования данных (ELT), ЦХД с оперативным складом данных (ОСД), расширенная модель с витринами данных (ВД), хранилище с накоплением данных в ВД, централизованная ETL с параллельными ХД И ВД и рекомендованная архитектура хранилища данных.

Сабир Асадуллаев

АРХИТЕКТУРЫ ХРАНИЛИЩ ДАННЫХ – 2

Об авторе: Сабир Асадуллаев, исполнительный архитектор SWG IBM EE/A, IBM,

Режим доступа:
https://www.ibm.com/developerworks/ru/library/sabir/axd_1/index.html#icomments

(статья воспроизводится в сокращении)

Вторая работа продолжает цикл из трех статей, посвященных архитектурам хранилищ данных (ХД) и их предшественников. Обилие различных подходов, методов и рекомендаций приводят к некоторой путанице понятий, достоинств, недостатков и границ применимости тех или иных архитектурных решений. В первой статье [1] рассмотрены эволюция понимания места OLAP, компоненты архитектуры ХД, виртуальные ХД и независимые витрины данных. Эта публикация посвящена таким архитектурам, как централизованное ХД с системой извлечения, преобразования и загрузки данных (ETL), хранилище данных с системой извлечения, загрузки и преобразования данных (ELT), ЦХД с оперативным складом данных (ОСД), расширенная модель с витринами данных (ВД).

Централизованное хранилище данных с ETL

Виртуальные хранилища данных и независимые витрины показали, что для эффективной работы аналитических систем необходим единый репозитарий данных. Для наполнения этого репозитория необходимо извлечь, согласовать разнородные данные из различных источников и загрузить эти данные в репозиторий.

Средства извлечения, преобразования и загрузки данных (ETL) должны знать все об источниках данных: структуры хранящихся данных и их форматы, различия в алгоритмах обработки данных, смысл хранящихся данных, график выполнения обработки информации в транзакционных системах. Игнорирование этих данных о данных (метаданных) неизбежно приводит к ухудшению качества информации, загружаемой в хранилище. В результате пользователи теряют доверие к хранилищу данных, стараются получать информацию напрямую из источников, что приводит к неоправданным временным затратам специалистов, эксплуатирующих системы – источники данных.

Таким образом, информация об источниках данных должна использоваться средствами ETL. Поэтому средства ETL должны работать в тесной связке со средствами ведения метаданных.

При обработке извлеченных данных необходимо преобразовать их к единому виду. Поскольку основные данные хранятся в реляционных базах данных, нужно учесть различие в кодировке данных. Даты могут кодироваться в разных формата; адреса могут использовать различные сокращения; кодировка продуктов может следовать различным номенклатурам. Первоначально информация о нормативно справочной информации (НСИ) заносилась в алгоритмы преобразования данных ETL. По мере роста числа источников данных объема обрабатываемых данных (он может достигать терабайтов в сутки), стало ясно, что необходимо отделить средства управления НСИ от средств ETL, и обеспечить их эффективное взаимодействие.

Таким образом, средства ETL извлекают данные из источников, во взаимодействии со средствами ведения метаданных и НСИ преобразуют их к требуемым форматам и загружают в репозиторий данных. В качестве репозитория чаще всего выступает репозиторий хранилища данных, но также может быть и оперативный склад данных (ОСД), и зоны временного хранения, и даже витрины данных. Поэтому одним из ключевых требований к средствам ETL является их способность взаимодействовать с различными системами.

Рис. 1. Централизованное хранилище данных с ETL

Необходимость повышения оперативности предоставляемой аналитической информации и рост объемов обрабатываемых данных выставляют повышенные требования к производительности средств ETL и их масштабируемости. Поэтому средства ETL должны использовать раз-

личные схемы параллельных вычислений и уметь работать на высокопроизводительных системах различных архитектур.

Как видно, к средствам ETL предъявляются самые разные требования:

– Необходимо собрать данные от разных систем – источников, даже если одна или несколько систем в результате сбоя не смогли в срок завершить свою работу и предоставить необходимые данные.

– Полученная информация должна быть распознана и преобразована в соответствии с алгоритмами преобразования, а также с помощью систем ведения НСИ и метаданных.

– Преобразованная информация должна быть загружена в зоны временного хранения, в хранилище данных, в ОСД, в витрины данных, как того требует производственный процесс.

– Средства ETL должны иметь высокую пропускную способность с тем, чтобы собирать и выгружать все возрастающие объемы данных.

– Средства ETL должны обладать высокими вычислительными возможностями и масштабируемостью для сокращения времени обработки данных для уменьшения задержек в предоставлении данных для аналитических работ.

– Средства ETL должны предоставлять разнообразные инструменты извлечения данных в различных режимах работы – от пакетного сбора для систем, некритичных к временным задержкам, до инкрементальной обработки в режиме, близком к реальному времени.

В связи с этими, зачастую взаимоисключающими требованиями, проектирование и разработка средств ETL превращается в сложную задачу даже тогда, когда используются решения, предлагаемые на рынке.

Централизованное хранилище данных с ELT

Традиционную систему извлечения, преобразования и загрузки данных (ETL) нередко упрекают в низкой производительности и высокой стоимости из-за необходимости создания выделенного программно-аппаратного комплекса. В качестве альтернативы предлагаются средства извлечения, загрузки и преобразования данных (ELT), которым приписываются высокая производительность и эффективное использование оборудования.

С тем, чтобы понять, каковы сравнительные преимущества и недостатки систем ETL и ELT, обратимся к трем основным функциям корпоративного хранилища данных (КХД):

1) полный и своевременный сбор и обработка информации от источников данных;

2) надежное и защищенное хранение данных;

3) предоставление данных для аналитических работ.

На вход систем ETL / ELT поступают разнородные данные, которые необходимо сравнить, очистить, привести к единым форматам, обработать по требуемым вычислительным алгоритмам. С одной стороны, в системах ETL / ELT данные практически не задерживаются, с другой – через эти системы в хранилище втекает основной поток информации. Поэтому требования к обеспечению защиты информации могут быть умеренными.

Рис. 2. Централизованное хранилище данных с ELT

Центральное хранилище данных (ЦХД), как правило, содержит такой объем информации, что ее полное раскрытие может привести к серьезным потерям для компании. В этом случае ЦХД требует создания вокруг себя надежного периметра информационной безопасности. Структуры данных в хранилище должны быть оптимизированы под требования долговременного, надежного и защищенного хранения. Применение схемы ELT означает, что ЦХД должно осуществлять и трансформацию данных.

Предоставление данных для аналитических работ требует реорганизации структур данных под каждую специфическую задачу. Многомерный анализу необходимы кубы данных; статистический анализ, как правило, работает с рядами данных; сценарный анализ и моделирование могут использовать файлы MS Excel. В рассматриваемой архитектуре бизнес – приложения используют данные непосредственно из ЦХД. В такой архитектуре в ЦХД должны храниться данные в структурах, оптимизированных как под текущие, так и под будущие бизнес – прило-

жения. Более того, подобный прямой доступ повышает вероятность несанкционированного доступа ко всем данным в хранилище.

Таким образом, мы видим, что в данной архитектуре на ЦХД возложены функции трансформации данных и обслуживания аналитических приложений. Обе эти функции несвойственны ЦХД, которое в таком виде превращается в устройство «все в одном», в котором, как правило, составляющие компоненты хуже, чем если бы они были реализованы отдельно (например, фотоаппарат в мобильном телефоне).

Как решается вопрос разделения функций хранения данных и предоставления данных для аналитических приложений, мы рассмотрим позже.

Применение схемы ETL позволяет полностью разнести функции обработки и хранения данных. Схема ELT нагружает центральное хранилище данных несвойственными ей функциями преобразования данных. В результате переноса функциональности от ETL в ЦХД нам необходимо не только обеспечить ту же вычислительную производительность, но и спроектировать универсальную платформу, способную равно эффективно обрабатывать данные и хранить их. Этот подход, может быть, применим для сегмента SOHO, но для корпоративных решений требуются профессиональные устройства.

Несмотря на декларируемые преимущества производительности схемы ELT, на практике выясняется, что:

1. Качество данных влияет на время их загрузки. Например, ETL при очистке и преобразовании данных может отбрасывать до 90% повторяющихся данных. ELT в этом случае загрузит все данные в ЦХД, где и будет происходить очистка.

2. Скорость преобразования данных в хранилище сильно зависит от алгоритмов обработки и структур данных. В некоторых случаях более эффективна SQL – обработка внутри базы данных хранилища, в других – быстрее будут работать внешние программы, извлекающие данные для обработки и загружающие результаты обработки в хранилище.

3. Некоторые алгоритмы очень сложно реализовать, используя средства SQL. Это накладывает ограничения на использование схемы ELT, тогда как ETL может использовать более эффективные инструменты обработки данных

4. ETL является единой областью, где сконцентрированы правила извлечения, обработки и загрузки данных, что упрощает эксплуатацию, доработку и тестирование алгоритмов. ELT, напротив, разносит алгоритмы сбора и загрузки с алгоритмами преобразования данных. То есть, для тестирования новых алгоритмов преобразования нужно либо рисковать целостностью данных в хранилище, находящемся в промышленном

производстве, либо создавать тестовую копию хранилища, что является весьма дорогостоящим мероприятием.

Таким образом, сравнивая ETL и ELT, мы видим, что преимущества при загрузке и преобразовании данных неочевидны, что ELT сталкивается с ограничениями SQL при преобразовании данных, и что экономия на программно – аппаратном комплексе ELT приводит к финансовым затратам на создание программно-аппаратной тестовой копии ЦХД.

Применение ELT, возможно, оправдано, если:

1. Нет жестких требований к надежности, производительности и защищенности хранилища.

2. Бюджетные ограничения вынуждают идти на риск утраты данных.

3. Хранилище данных и источники данных взаимодействуют через сервисную шину (SOA).

Последний случай наиболее экзотичен, но и он имеет право на существование в определенных условиях. В этом случае на шину возложена интеграция источников с ХД на уровне обмена сообщениями, и минимальное (по меркам хранилища) преобразование данных и их загрузка хранилище.

Централизованное ХД с ОСД

Процессы извлечения, преобразования и загрузки данных, безусловно, требуют некоторого времени для завершения своей работы. Дополнительная задержка вызвана необходимостью проверки загруженных в хранилище данных на непротиворечивость с уже имеющимися данными, на консолидацию данных, на перевычисления итоговых значений с учетом новых данных.

Оперативный склад данных (ОСД) был предложен в 1998 г. 3 с тем, чтобы сократить время задержки между поступлением информации из ETL и аналитическими системами. Операционный склад данных располагает менее точной информацией из-за отсутствия внутренних проверок, и более детальными данными из-за отсутствия этапа консолидации данных. Поэтому данные из ОСД предназначены для принятия тактических решений, тогда как информация из центрального хранилища данных (ЦХД) лучше подходит для решения стратегических задач 4.

Действительно, если в результате неполных и неточных данных, содержащихся в ОСД, будет привезена лишняя бутылка воды, то это не приведет к серьезным потерям. Однако ошибка планирования из-за низкого качества данных в ЦХД может оказать негативное воздействие на принятие решения о номенклатуре и объемах оптовых закупок.

Требования к защите информации в ОСД и ЦХД также различаются. В нашем примере в ОСД размещаются данные с горизонтом времени, не превышающим нескольких часов. В ЦХД может храниться историческая информация, охватывающая период в несколько лет для более надежного прогнозирования необходимых объемов закупок. И эта историческая информация может представлять значительный коммерческий интерес для конкурентов. Поэтому аналитики-тактики могут напрямую работать с ОСД, тогда как аналитики-стратеги должны работать с ЦХД через витрины данных для разграничения ответственности. Отсутствие витрин данных помогает тактикам быстрее получить доступ к данным. Наличие витрин данных не препятствует стратегическому анализу, так как такой анализ осуществляется ежемесячно, или даже ежеквартально.

Рис. 3. Централизованное ХД с ОСД

Архитектура, представленная на рис. 3, предполагает прямую работу бизнес-приложений с ЦХД. Разбор достоинств и ограничений подобного подхода будет выполнен в разделе «Расширенная модель ХД с витринами данных». Сейчас необходимо отметить, что при последовательном перемещении данных ОСД фактически выполняет еще одну роль зоны временного хранения. Аналитики-тактики, работая с данными из ОСД, вольно или невольно выявляют ошибки и противоречия в данных, тем самым, повышая их качество. Исправленные данные из ОСД в данной схеме перегружаются в ЦХД. Однако возможны и иные схемы, например, когда данные из ETL посту-

пают одновременно в ОСД и ЦХД. После использования в ОСД ненужные данные просто стираются. Эта схема применима в тех случаях, когда человеческое вмешательство в данные может только исказить их, вольно, или невольно.

Расширенная модель ХД с витринами данных

Прямая работа пользовательских программ с корпоративным хранилищем данных (КХД), допустима, если пользовательские запросы не препятствуют нормальному функционированию КХД, если между пользователями и КХД имеются высокоскоростные линии связи, или если случайный доступ ко всем данным не ведет к серьезным потерям. Администрирование прямого доступа пользователей к КХД представляет собой чрезвычайно сложную задачу. Например, пользователь одного подразделения имеет право доступа к данным другого подразделения только через 10 дней после получения этих данных. Или пользователь может видеть только агрегированные показатели, но не детальные данные. Существуют и другие, еще более запутанные правила доступа. Их ведение, учет и изменение приводит к неизбежным ошибкам, вызванным сочетанием сложных условий доступа.

Витрины данных, содержащие информацию, предназначенную для выделенной группы пользователей, значительно снижают риски нарушения требования информационной безопасности.

Рис. 4. Расширенная модель с витринами данных

До сих пор серьезной проблемой для территориально распределенных организаций является качество линий связи. В случае обрыва или

недостаточной пропускной способности удаленные пользователи лишаются доступа к информации, содержащейся в КХД. Решением являются удаленные витрины данных, которые заполняются либо в нерабочее время, либо инкрементально, по мере поступления информации, с использованием транспорта с гарантированной доставкой.

Разные пользовательские приложения нуждаются в различных форматах данных: многомерные кубы, ряды данных, двумерные массивы, реляционные таблицы, файлы в формате MS Excel, текстовые файлы с разделителями, XML-файлы и т.д. Никакая структура данных в КХД не может удовлетворить этим требованиям. Выходом является создание витрин, чьи структуры данных оптимизированы под специфические требования отдельных приложений.

Еще одной причиной необходимости создания витрин данных является требование к надежности КХД, которое часто определяется, как пять или четыре девятки. Это означает, что КХД может простаивать не более 5 минут в год (99,999%) или не более часа в год (99,99%). Создание комплекса с такими характеристиками является сложной и весьма недешевой инженерной задачей. Требования к защите от терактов, саботажа и стихийных бедствий еще более усложняют построение программно-технического комплекса и осуществление соответствующих организационных мероприятий. Чем сложнее такой комплекс, чем больше данных он хранит, тем выше его стоимость и сложнее его поддержка. Наличие витрин данных резко снижает нагрузку на КХД, как по количеству пользователей, так и по объему данных в хранилище, так как эти данные могут быть оптимизированы под хранение, а не под обслуживание запросов.

Если витрины наполняются напрямую из КХД, то фактическое количество пользователей снижается с сотен и тысяч до десятков витрин, которые и являются пользователями КХД. При использовании средств SRD (Sample, Restructure, Delivery – выборка, реструктуризация, доставка) количество пользователей сокращается до 1. В этом случае вся логика информационного снабжения витрин сосредотачивается в SRD. Витрины могут быть оптимизированы под обслуживание пользовательских запросов. Программно-технический комплекс КХД может быть оптимизирован исключительно под надежное, защищенное хранение данных.

Средства SRD также смягчают нагрузку на КХД за счет того, что разные витрины могут обращаться к одним и тем же данным, тогда как SRD извлекает данные один раз, преобразует к различным форматам и доставляет в разные витрины данных.

Заключение

В статье рассмотрены такие архитектуры, как централизованное ХД с системой извлечения, преобразования и загрузки данных (ETL), хранилище данных с системой извлечения, загрузки и преобразования данных (ELT), ЦХД с оперативным складом данных (ОСД), расширенная модель с витринами данных (ВД). В завершающей работе будут обсуждены достоинства и ограничения следующих архитектур: ХД с накоплением данных в ВД, централизованная ETL с параллельными ХД и ВД, ХД с интеграционной шиной, а также рекомендованная архитектура хранилища данных.

Сабир Асадуллаев

АРХИТЕКТУРЫ ХРАНИЛИЩ ДАННЫХ – 3

Об авторе: Сабир Асадуллаев, архитектор решений
SWG IBM EE/A, IBM
Режим доступа:
https://www.ibm.com/developerworks/ru/library/sabir/axd_1/index.html#icomments

Третья работа завершает цикл из трех статей, посвященных архитектурам хранилищ данных (ХД) и их предшественников. Обилие различных подходов, методов и рекомендаций приводят к некоторой путанице понятий, достоинств, недостатков и границ применимости тех или иных архитектурных решений. В первой статье [1] рассмотрены эволюция понимания места OLAP, компоненты архитектуры ХД, виртуальные ХД и независимые витрины данных. Вторая публикация [2]посвящена таким архитектурам, как централизованное ХД с системой извлечения, преобразования и загрузки данных (ETL), хранилище данных с системой извлечения, загрузки и преобразования данных (ELT), ЦХД с оперативным складом данных (ОСД), расширенная модель с витринами данных (ВД). В этой статье рассмотрены хранилище с накоплением данных в ВД, централизованная ETL с параллельными ХД и ВД, ХД с интеграционной шиной, а также рекомендованная архитектура хранилища данных

Централизованная ETL с параллельными хранилищами и витринами данных

В данном случае система извлечения, преобразования и загрузки данных (ETL) является центром, вокруг которого строится вся архитектура КХД. Информация из разнородных источников поступает в ETL, которая загружает очищенные и согласованные данные в центральное хранилище данных (ЦХД), в оперативный склад данных (ОСД), если он

есть, и, при необходимости, в зоны временного хранения. Это обычная практика для КХД. Необычным является загрузка данных из ETL напрямую в витрины данных.

На практике такая архитектура возникает из-за требований скорейшего, без временных задержек, доступа к аналитическим данным. Использование оперативного склада данных не решает задачи, так как пользователи могут находиться в другом регионе, и им требуется территориальная витрина данных. Другой причиной может стать запрет на размещение разнотипной информации в ОСД по соображениям безопасности.

По тем или иным причинам, подобные архитектуры встречаются, и одной из проблем их эксплуатации являются сложности с восстановлением данных после краха витрин, напрямую снабжающихся из ETL. Дело в том, что средства ETL не предназначены для долговременного хранения извлеченных и очищенных данных. Транзакционные системы, как правило, ориентированы на выполнение текущих операций. Поэтому при потере данных в витринах, связанных с ETL, приходится либо поднимать информацию из средств резервного копирования (backup) транзакционных систем, либо организовывать исторические архивы систем – источников данных. Подобные архивы не только требуют средств на свое создание и поддержку в эксплуатации, но и являются, с корпоративной точки зрения, избыточными, так как дублируют функции корпоративного хранилища, но предназначены для ограниченного количества витрин данных.

Еще одним решением является двойное подключение подобных витрин – напрямую к средствам ETL и к хранилищу данных, что приводит к недоразумениям и рассогласованиям результатов аналитических работ. Причина кроется в том, что данные, поступающие в хранилище, как правило, проходят дополнительные проверки на непротиворечивость с уже загруженными данными. Например, может прийти финансовый документ с реквизитами, почти совпадающими с документом, поступившим в ЦХД ранее. Система ETL, не обладая памятью обо всех загруженных данных, не может выявить, является ли новый документ законным исправлением существующего, или это ошибка.

Средства верификации данных могут выявить подобные ситуации, действуя внутри хранилища данных. В случае выявления ошибки новые данные будут отброшены. Если же это регламентированное исправление, то изменения коснутся не только данных цифр, но и агрегированных показателей, составленных при участии исправляемых данных.

Таким образом, информация, попавшая в витрину данных напрямую из ETL, может противоречить данным, поступившим из ЦХД. В качестве решения иногда в витрине реализуют те же алгоритмы верификации данных, что и в ЦХД. Недостатком является не-

обходимость поддержки и синхронизации одних и тех же алгоритмов в ЦХД и в витринах, питающихся непосредственно от ETL.

Рис. 1. Централизованная ETL с параллельными ХД И ВД

Подытоживая, можно сказать, что параллельные витрины данных приводят к повторной обработке данных, к созданию избыточных операционных архивов, к поддержке дублирующих приложений и децентрализации обработки данных, что, как правило, является причиной рассогласования информации.

Тем не менее, параллельные витрины имеют право на существование в тех случаях, когда оперативность доступа к аналитической информации важнее недостатков этой архитектуры.

Хранилище с накоплением данных в витринах

Основанием для появления этой архитектуры явились следующие предпосылки.

Некоторые компании до сих пор внедряют и эксплуатируют разрозненные прикладные витрины данных. Качество данных в этих витринах удовлетворяет аналитиков, работающих с витринами.

В некоторых компаниях сложилось мнение, что создание корпоративного хранилища данных (КХД) подобно смертельному трюку с непредсказуемыми последствиями. Несмотря на то, что трудности создания и внедрения КХД, прежде всего, связаны не с технологическими вопросами, а с плохой организаций проекта и недостаточным вовлечением экспертов – будущих пользователей КХД, тем не менее, возникает желание пойти легким путем.

Требование быстрых результатов. Необходимость отчитываться ежеквартально вызывает потребность в быстрых осязаемых результатах. В результате появляется стремление сделать и внедрить какое-нибудь ограниченное решение без связи с остальными задачам.

Вольно или невольно следуя этим принципам, компании сначала внедряют разрозненные независимые витрины, в надежде, что содержащиеся в них данные будут легко, просто и быстро объединены. В реальности все гораздо сложнее. Качество данных в витринах может удовлетворять экспертов, работающих с ними, но эти информация не согласована с данными из других витрин, поэтому на стол руководству ложатся отчеты, которые нельзя привести к единому виду.

Одни и те же показатели могут вычисляться по разным алгоритмам, на основании разного набора данных, за разные сроки. Показатели с одинаковыми названиями могут скрывать разные сущности, и наоборот, одинаковые сущности могут иметь разные наименования.

Рис. 2. Хранилище с накоплением данных в витринах

Диагноз – пользователи независимых прикладных витрин говорят на разных языках бизнеса, и каждая витрина содержит собственные метаданные.

Другая проблема заключается в различии нормативно-справочной информации (НСИ), используемых в независимых витринах данных. Разница в кодировке данных, в используемых кодификаторах, справочниках, классификаторах, идентификаторах, нормативах и словарях делает невозможным объединение этих данных

без серьезного анализа, проектирования и разработки средств ведения НСИ.

Однако в организации уже существуют планы, бюджет и сроки создания КХД на основе независимых витрин данных. Руководство ожидает получить результат быстро и недорого. Разработчики, обещавшие экономию ресурсов, вынуждены сделать хоть что-нибудь. Так создаются хранилища несогласованных отчетов, что в корне противоречит самой идее создания хранилищ данных как единого и единственного источника очищенных, согласованных и непротиворечивых исторических данных.

Понятно, что ни руководство, ни пользователи подобного хранилища не склонны доверять информации, содержащейся в нем. Поэтому на следующем этапе встает необходимость радикальной переработки, а по сути, создания заново, хранилища, ориентированного на хранение не отчетов, а показателей, из которых будут собираться отчеты.

Эта работа невозможна без использования средств ведения метаданных и НСИ, область действия которых будет распространяться только на центральное хранилище (ЦХД), так как независимые витрины данных содержат свои метаданные и НСИ.

В результате руководство и эксперты могут получить согласованные и непротиворечивые отчеты, но они не смогут проследить происхождение данных сквозным образом, так как между независимыми витринами и ЦХД есть разрыв в ведении метаданных.

Таким образом, стремление к достижению сиюминутных результатов и к демонстрации быстрых успехов приводит к отказу от единого, сквозного управления метаданными и НСИ. Итогом такого подхода является наличие семантических островов, где сотрудники говорят на разных бизнес-языках.

Тем не менее, эта архитектура имеет право на существование, там, где единая модель данных или не нужна, или невозможна, и где в ЦХД передается сравнительно небольшой объем данных без необходимости детализации их происхождения и исходных составляющих. Например, если компания, оперирующая в разных странах, уже внедрила национальные хранилища данных, которые следуют локальным требованиям законодательства и правилам ведения бизнеса и финансового учета. Центральное хранилище данных может забирать из национальных ХД только часть информации для корпоративной отчетности. Создавать единую модель данных нет необходимости, поскольку она не будет востребована на национальном уровне.

Естественно, что такая схема требует высокой степени доверия к национальным данным, и может быть использована, если умышленное или неумышленное искажение этих данных не приведет к тяжелым финансовым последствиям для всей организации.

Хранилище данных с интеграционной шиной

Широкое распространение сервис – ориентированной архитектуры (СОА) [3] привело к желанию использовать ее в решениях для корпоративных хранилищ данных (КХД) вместо средств извлечения, преобразования и загрузки данных (ETL) в центральное хранилище (ЦХД) и вместо средств выборки, реструктуризации и доставки данных (SRD) в витрины данных.

Интеграционная шина, которая лежит в основе СОА, предназначена для интеграции веб-сервисов и приложений и выполняет следующие задачи:

1. Определяет сервис, соответствующий запросу от источника, и направляет запрос к сервису.

2. Преобразует транспортные протоколы между источником запроса и сервисом.

3. Преобразует форматы сообщений между источником запроса и сервисом.

4. Управляет бизнес-событиями различных источников.

Рис. 3. Хранилище данных с интеграционной шиной

На первый взгляд функциональные возможности СОА позволяют применить ее для замены ETL и SRD. Действительно, ETL выполняет посреднические функции между ЦХД и источниками данных, а SRD является посредником между ЦХД и витринами данных. Если заменить ETL и SRD на интеграционную шину, то, казалось бы, можно воспользоваться гибкостью, предоставляемой шиной для интеграции приложений. Представим себе, что ЦХД, оперативный склад данных (ОСД), зо-

ны временного хранения, системы ведения метаданных и НСИ обращаются к шине как независимые приложения с запросами к источникам данных на обновление данных.

Прежде всего, в разы возрастет нагрузка на системы-источники данных, так как одна и та же информация будет многократно передаваться по запросам в ЦХД, ОСД, зоны временного хранения и системы управления метаданными и НСИ. Очевидное решение – создать собственное хранилище данных при шине для кеширования запросов.

Во-вторых, регламент сбора информации, ранее централизованный в ETL, теперь рассеян по приложениям, запрашивающим данных. Рано или поздно возникнут рассогласования в регламентах сбора данных для ЦХД, ОСД, систем ведения НСИ и метаданных. Данные, собранные по разным методикам, в разные отрезки времени, обработанные по разным алгоритмам, будут не согласованны друг с другом. Тем самым будет разрушена основная цель создания ЦХД как единого источника согласованных непротиворечивых данных.

В случае замены SRD на интеграционную шину последствия не столь драматичны. Но для того, чтобы ЦХД могло отвечать на запросы витрин данных, направленных через шину, оно должно быть преобразовано в сервис. Это значит, что хранилище должно соответствовать наиболее распространенному стилю web – сервисов, и поддерживать протоколы HTTP/ HTTPS и SOAP и XML – формат сообщений. Такой подход работает для коротких сообщений, но в витрины необходимо передавать большой объем данных, что может быть решено с помощью передачи двоичных объектов. Необходимая реструктуризация данных не может быть возложена на шину, и должна выполняться либо в ЦХД, либо в витрине. Эта функция может быть решена с помощью сервиса-посредника, принимающего данные, и передающего их в витрины данных после реструктуризации. То есть, мы возвращаемся к идее средства SRD с шинным интерфейсом.

Таким образом, интеграционная шина может быть использована в архитектуре КХД как транспортная среда между источниками данных и ETL и между SRD и витринами данных в тех случаях, когда компоненты КХД территориально разнесены и находятся за межсетевыми экранами в соответствии со строгими требованиями к защите информации. В этом случае для обеспечения взаимодействия достаточно, чтобы был разрешен обмен по протоколам HTTP/ HTTPS. Вся логика сбора и преобразования информации должна быть по-прежнему сосредоточена в ETL и SRD.

Рекомендованная архитектура КХД

Архитектура корпоративного хранилища данных (КХД) должна удовлетворять многим функциональным и нефункциональным требованиям, которые зависят от конкретных задач, решаемых КХД. Как нет универсального банка, авиакомпании, или нефтяного концерна, так нет и единого

решения КХД, пригодного на все случаи жизни. Но основные принципы, которым должно следовать КХД, все же можно сформулировать.

Прежде всего, это качество данных, которое можно понимать, как полные, точные и воспроизводимые данные, доставленные в срок туда, где они нужны. Качество данных трудно измерить напрямую, но о нем можно судить по принимаемым решениям. То есть, качество данных требует инвестиций, но и само способно приносить прибыль.

Во-вторых, это защищенность и надежность хранения данных. Ценность информации, накопленной в КХД, может быть сравнима с рыночной стоимостью компании. Несанкционированный доступ к КХД чреват серьезными последствиями, поэтому должны быть приняты меры, адекватные ценности данных.

В-третьих, данные должны быть доступны сотрудникам в объеме, необходимом и достаточном для выполнения своих функциональных обязанностей.

В-четвертых, сотрудники должны иметь единое понимание данных, то есть должно быть установлено единое смысловое пространство.

В-пятых, необходимо, по возможности, устранить конфликты в кодировках данных в системах источниках.

Рис. 4. Рекомендованная архитектура КХД

Предлагаемая архитектура следует проверенным принципам модульного конструирования «непотопляемых отсеков». Стратегия «Разделяй и властвуй» применима не только в политике. Разделяя архитектуру на модули, мы одновременно концентрируем в них определенную функциональность, получая власть над неуправляемой ИТ стихией. Средства ETL обеспечивают полный, надежный, точный сбор информации из источников данных благодаря сосредоточенной в ETL логике

сбора, обработки и преобразования данных и взаимодействию с системами ведения метаданных и НСИ.

Система ведения метаданных является главным "хранителем мудрости", к которому можно обратиться за советом. Система ведения метаданных поддерживает актуальность бизнес-метаданных, технических, операционных и проектных метаданных.

Система ведения НСИ является третейским судьей при разрешении конфликтов кодировок данных.

Центральное хранилище данных (ЦХД) несет только нагрузку по надежному защищенному хранению данных. В зависимости от поставленных задач, надежность программно-технического комплекса (ПТК) ЦХД может достигать 99,999%, то есть обеспечивать бесперебойную работу с простоем не более 5 мин в год. ПТК ЦХД может обеспечивать защиту данных от несанкционированного доступа, саботажа и стихийных бедствий. Структура данных в ЦХД оптимизирована исключительно с целью обеспечения эффективного хранения данных.

Средства выборки, реструктуризации и доставки данных (SRD) в такой архитектуре являются единственным пользователем ЦХД, беря на себя всю работу по заполнению витрин данных и, тем самым, снижая нагрузку на ЦХД по обслуживанию запросов пользователей.

Витрины данных содержат данные в структурах и форматах, оптимальных для решения задач пользователей данной витрины. В настоящее время, когда даже ноутбук может быть оснащен терабайтным диском, проблемы, связанные с многократным повторением данных в витринах, не имеют значения. Главное преимущество этой архитектуры – предоставление доступа для удобной работы пользователей с необходимым объемом данных, возможность быстрого восстановления содержимого витрин из ЦХД при сбое витрины, обеспечение работы пользователей при отсутствии связи с ЦХД.

Достоинство этой архитектуры заключается в возможности раздельного проектирования, создания, эксплуатации и доработки отдельных компонентов без радикальной перестройки всей системы. Это означает, что начало работ по созданию КХД не требует сверхусилий или сверхинвестиций. Достаточно начать с ограниченного по своим возможностям программно-технического комплекса, и следуя предложенным принципам, создать работающий и действительно полезный для пользователей прототип. Далее необходимо выявить узкие места и развивать соответствующие компоненты.

Применение этой архитектуры вместе с тройной стратегией интеграции данных, метаданных и НСИ [4], позволяет сократить сроки и бюджет проекта внедрения КХД и развивать его в соответствии с изменяющимися требованиями бизнеса.

Заключение

В статье обсуждаются достоинства и ограничения следующих архитектур: ХД с накоплением данных в ВД, централизованная ETL с параллельными ХД и ВД, ХД с интеграционной шиной, а также рекомендованная архитектура хранилища данных.

Рекомендованная архитектура корпоративного хранилища данных позволяет создать в короткие сроки и с минимальными инвестициями работоспособный прототип, полезный для бизнес-пользователей. Ключевым моментом этой архитектуры, обеспечивающим эволюционное развитие КХД, является внедрение на ранних этапах систем ведения метаданных и НСИ.

ПРИЛОЖЕНИЯ

Диаграммы и иллюстрации

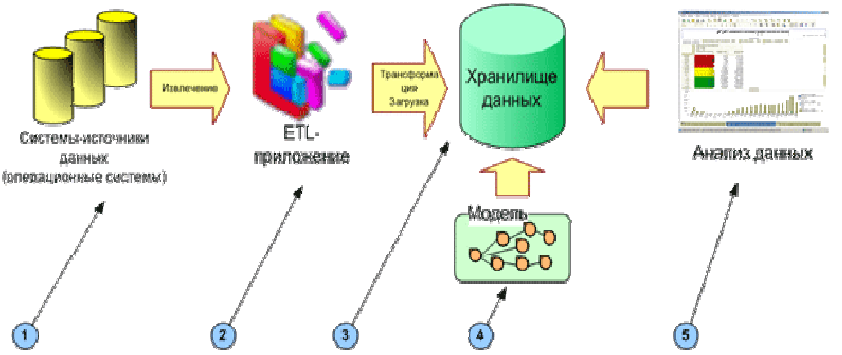


Рис. 1. Компоненты BI-системы

Источник: Э.Э. Акимкина, А.Э. Аббасов Анализ инструментальных средств информационных систем для обработки многомерных данных // Информационно-технологический вестник. 2016. Т, 2. С. 61–75

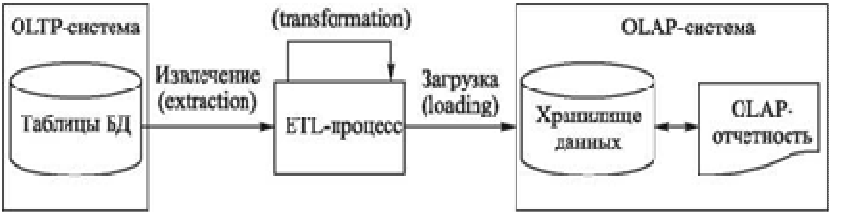


Рис. 2. ETL-процесс

Источник: С.А. Тоноян, В.А. Высочанский методика проектирование корпоративного хранилища данных на базе платформы SAP NET WEAVER BUSINESS WAREHOUSE / Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2016. № 4

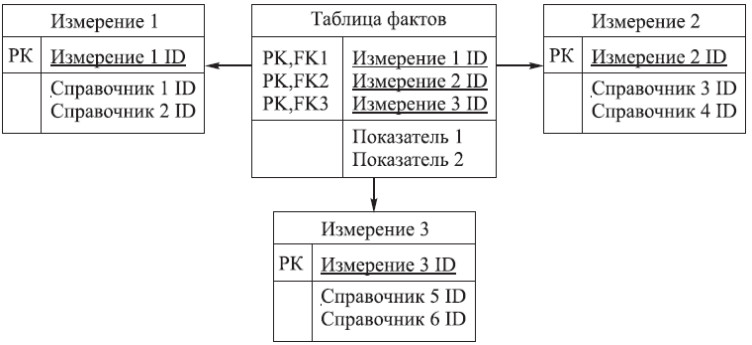


Рис. 3. Классическая многомерная модель OLAP-куба

Источник: С.А. Тоноян, В.А. Высочанский методика проектирование корпоративного хранилища данных на базе платформы SAP NET WEAVER BUSINESS WAREHOUSE // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2016. № 4

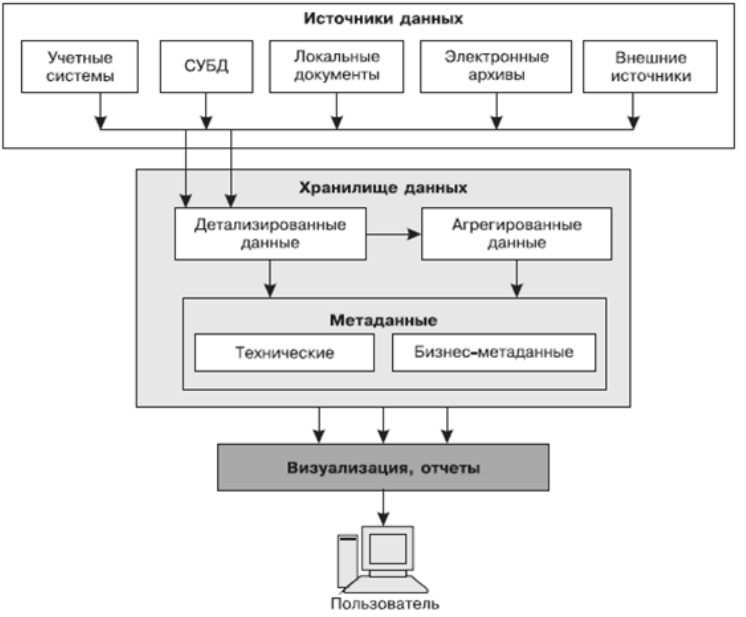
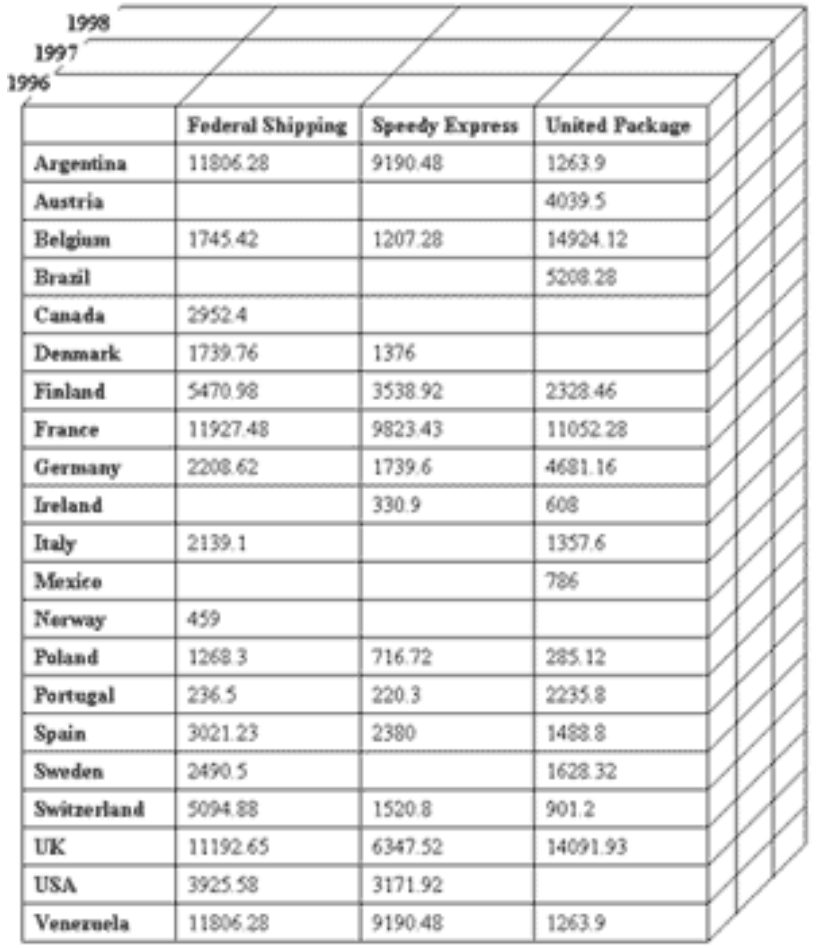


Рис. 4. Концептуальная схема ХД

Источник: Э.Э. Акимкина, А.Э. Аббасов Анализ инструментальных средств информационных систем для обработки многомерных данных / Информационно-технологический вестник. 2016. Т. 2. С. 61–75



	Federal Shipping	Speedy Express	United Package
Argentina	11806.28	9190.48	1263.9
Austria			4039.5
Belgium	1745.42	1207.28	14924.12
Brazil			5208.28
Canada	2952.4		
Denmark	1739.76	1376	
Finland	5470.98	3538.92	2328.46
France	11927.48	9823.43	11052.28
Germany	2208.62	1739.6	4681.16
Ireland		330.9	608
Italy	2139.1		1357.6
Mexico			786
Norway	459		
Poland	1268.3	716.72	285.12
Portugal	236.5	220.3	2235.8
Spain	3021.23	2380	1488.8
Sweden	2490.5		1628.32
Switzerland	5094.88	1520.8	901.2
UK	11192.65	6347.52	14091.93
USA	3925.58	3171.92	
Venezuela	11806.28	9190.48	1263.9

Рис. 5. Трехмерный набор агрегатных данных – OLAP-куб

Источник: Введение в OLAP: часть 1. Основы OLAP. Режим доступа: http://olap.ru/basic/OLAP_intro1.asp

Рис. 6. Иерархия измерений, связанном с географическим положением

Источник: Дмитрий Лобач. Основы OLAP. Режим доступа: http://www.softkey.info/reviews/review.php?ID=465

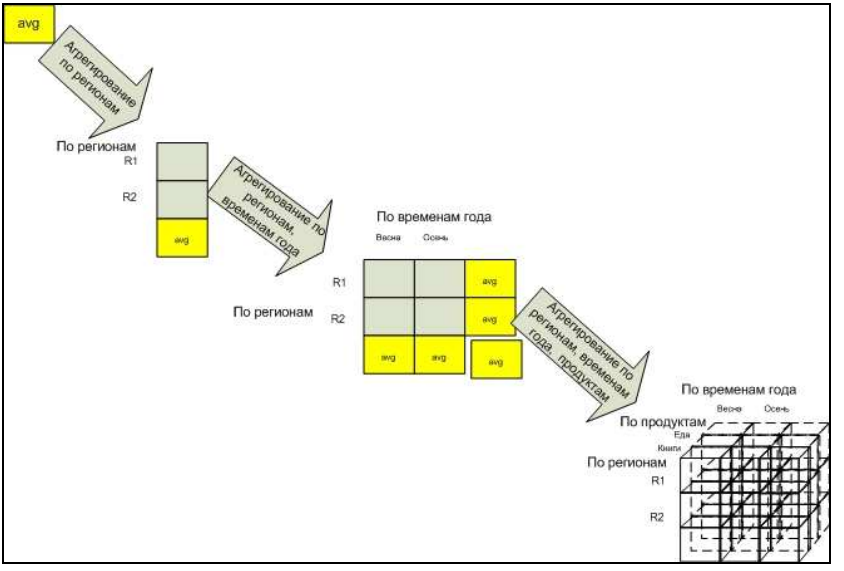


Рис. 7. Схема агрегирования данных для формирования куба

Источник: Математическая модель olap-кубов

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 4

1. АРХИТЕКТУРЫ ДАННЫХ. ПОНЯТИЕ АРХИТЕКТУРЫ ДАННЫХ. РАЗВИТИЕ СИСТЕМ ХРАНЕНИЯ И ОБРАБОТКИ ДАННЫХ 6

Левичев А. POWER BL OT MICROSOFT: СЕРВИС БИЗНЕС-АНАЛИТИКИ ДЛЯ КОМПАНИЙ 6

Акимкина Э.Э., Аббасов А.Э. АНАЛИЗ ИНСТРУМЕНТАЛЬНЫХ СРЕДСТВ ИНФОРМАЦИОННЫХ СИСТЕМ ДЛЯ ОБРАБОТКИ МНОГОМЕРНЫХ ДАННЫХ 8

2. МНОГОМЕРНЫЕ ДАННЫЕ. OLAP-ТЕХНОЛОГИЯ, КАК КЛЮЧЕВОЙ КОМПОНЕНТ ХРАНИЛИЩ ДАННЫХ 23

Лобач Дм. ОСНОВЫ OLAP 23

Федечкин С. ХРАНИЛИЩЕ ДАННЫХ: ВОПРОСЫ И ОТВЕТЫ 28

Горохов М.М., Переведенцев Д.А. ФОРМИРОВАНИЕ МНОГОМЕРНОЙ МОДЕЛИ ДАННЫХ ДЛЯ ЦЕЛЕЙ OLAP-АНАЛИЗА В ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЕ УПРАВЛЕНИЯ НАУЧНЫМИ ПРОЕКТАМИ 35

Кузнецов Д., Кудрявцев Ю.А. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ OLAP-КУБОВ 43

3. КОНЦЕПЦИЯ ХРАНИЛИЩ ДАННЫХ 50

Волков А.И. ПРОБЛЕМЫ ИНТЕГРАЦИИ ХРАНИЛИЩ ДАННЫХ С ОТКРЫТЫМИ И БОЛЬШИМИ ДАННЫМИ И ПОДХОДЫ К ИХ РЕШЕНИЮ 50

Кригер А.Б. 77

Кригер А.Б. МОНИТОРИНГ И АНАЛИЗ ЭКОНОМИЧЕСКОЙ ДЕЯТЕЛЬНОСТИ В ГОСУДАРСТВЕННЫХ УЧРЕЖДЕНИЯХ ИСПОЛНИТЕЛЬСКОГО ИСКУССТВА 77

4. АРХИТЕКТУРЫ ХРАНИЛИЩ ДАННЫХ 87

Сабир Асадуллаев. Архитектуры хранилищ данных – 1 87

Сабир Асадуллаев. Архитектуры хранилищ данных – 2 96

Сабир Асадуллаев. Архитектуры хранилищ данных – 3 105

ПРИЛОЖЕНИЯ 115

Диаграммы и иллюстрации 115

Учебное издание

ХРАНИЛИЩА ДАННЫХ И ИХ ИСПОЛЬЗОВАНИЕ

Хрестоматия

Составитель
Кригер Александра Борисовна

В авторской редакции
Компьютерная верстка М.А. Портновой

120

Подписано в печать 12.09.2017. Формат 60×84/16.
Бумага типографская. Печать офсетная. Усл. печ. л. 10,0
Тираж 200 экз. Заказ

Издательство Владивостокского государственного университета
экономики и сервиса
690014, Владивосток, ул. Гоголя, 41
Отпечатано во множительном участке ВГУЭС
690014, Владивосток, ул. Гоголя 41