

УДК681.3.07

Интеграция данных в информационных системах

*Белошицкий Д.А., аспирант,
Кафедра «Компьютерные системы и сети»
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана*

*Научный руководитель: Брешинов А.В., д.т.н., профессор
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
v.suzev@bmstu.ru*

Введение

Под интеграцией данных в информационных системах понимается обеспечение единого унифицированного интерфейса для некоторой совокупности, вообще говоря, неоднородных независимых источников данных. Такими источниками данных могут быть традиционные системы баз данных, поддерживающие различные модели данных (реляционные, объектные, объектно-реляционные, графовые и т.п.), разнообразные унаследованные системы, репозитории, Web-сайты, файлы структурированных данных. Иными словами, речь идет о поддержке представления совокупности данных из множества независимых источников в терминах единой модели данных. Важно заметить, что состав множества источников может быть наперед заданным или динамически пополняемым, источники данных могут обладать неизменным или обновляемым содержанием.

По определению аналитиков Gartner, интеграция данных охватывает практики, архитектурные подходы и программные инструменты для обеспечения согласованного доступа и доставки данных для всего спектра приложений и бизнес-процессов компании (Magic quadrant for data integration tools, Gartner, 2006). Как свидетельствуют исследования, затраты на программные средства интеграции данных сегодня неуклонно растут в самых разных индустриях и географических регионах.

Интеграция данных включает объединение данных, находящихся в различных источниках и предоставление данных пользователям в унифицированном виде. Этот процесс становится существенным как в коммерческих задачах (когда двум похожим компаниям необходимо объединить их базы данных), так и в научных (комбинирование результатов исследования из различных биоинформационных репозитариев). Роль интеграции данных возрастает, когда увеличивается объем и необходимость совместного

использования данных. Это стало фокусом обширной теоретической работы, а многочисленные проблемы остаются нерешёнными.

Области использования систем интеграции данных

Интеграция схем данных

Так как схемы баз данных разрабатываются независимо друг от друга, они зачастую имеют различную структуру и терминологию. Например, базы данных приложений для автоматизации работы бухгалтерии и управления взаимоотношениями с клиентами (CRM-системы) моделируют разные предметные области, понятийные аппараты которых пересекаются. В итоге возникают проблемы омонимии, синонимии, абсолютного несоответствия понятий. Первым шагом на пути к интеграции данных является нахождение соответствий между двумя схемами данных. После нахождения соответствий совпадающие элементы могут быть объединены в единую обобщенную схему (в случае консолидации) или обобщенное представление, если выбрана архитектура федерализованных баз данных.

Хранилища данных

Хранилища данных – это одна из первых технологий интеграции данных. В настоящее время существует большое количество хранилищ данных, которые составляют основу систем поддержки принятий решений. Таким образом, хранилища данных – это неотъемлемый элемент архитектуры корпоративной информационной системы современного предприятия, доказавший свою эффективность и активно используемый аналитическими подразделениями. Основным источником данных для корпоративного хранилища являются транзакционные системы и другие репозитории (например, файловые архивы). Данные источников имеют принципиально отличную от хранилищ структуру и поэтому оригинальные данные необходимо трансформировать. Для этого используется технология ETL (Extract, Transform, Load). Существуют различные ETL-инструменты и пакеты типа MS Integration Services, но они предполагают по большей части ручную работу, что сильно снижает производительность труда программиста, разрабатывающего ETL-процедуры. В этом случае использование мэтчинга помогло бы исправить ситуацию.

Так как существует достаточно обширный набор ETL-преобразований, то перспективным также представляется решение следующей задачи: зная отображение $S \Rightarrow W$, где S - схема реляционной базы данных-источника, а W -схема данных хранилища, построить отображение $S' \Rightarrow W$. Такая постановка задачи целесообразна, когда схемы S и

S' похожи друг на друга вычислительная сложность $Match(S, S')$ меньше, чем $Match(S', W)$.

Электронная коммерция

В последние годы электронная коммерция обусловила ещё одну область применения мэтчинга схем данных: преобразование сообщений. Дело в том, что электронные торговцы активно используют электронные сообщения определенного формата, описывающие финансовые транзакции. Проблема состоит в том, что существуют несколько форматов данных, например, EDI, XML, или какие-то кастомизированные структуры данных. При интеграции с биржами и системами других электронных торговцев часто возникает задача трансформации сообщений в нужный формат, а это отчасти задача мэтчинга схем данных, только в данном случае понятие «схема данных» вырождается в понятие «формат сообщения».

Семантическая обработка запросов

Предыдущие три примера характерны тем, что они выполняются на этапе проектирование результирующей схемы данных или, по крайней мере, этап ее доработки по причине присоединения нового источника. Совершенно другой сценарий наблюдается во время обработки запросов непосредственно во время их исполнения (run-time). Пользователь явно указывает ожидаемый результат, который должен быть получен запросом, а система должна определить наиболее достоверный и актуальный источник информации для этого запроса, преобразовать названия столбцов (в случае реляционной таблицы) и граничных условий в адекватный для выбранного источника формат, получить результат и вернуть его в терминах пользователя. Например, пользователь вводит запрос:

```
SELECT param1, param2
```

```
FROM [global_database].[global_table1]
```

```
WHERE param1 > 11;
```

В этом случае система на первом шаге находит наиболее подходящий источник информации ([local_database_1].[local_table_1]), преобразует названия параметров запроса (local_param1, local_param2) и изменяет шкалу измерения (пользователь оперирует тысячами рублей, а в БД-источнике информация о local_param1 хранится в рублях). Итоговый запрос к БД local_database_1 таков:

```
SELECT local_param1, local_param2
```

```
FROM [local_database].[local_table1]
```

```
WHERE local_param1 > 11000;
```

После получения данных столбцы результирующей таблицы должны называться param1 и param2, то есть результат формируется в терминах пользователя.

Также операция интеграции может быть полезна для Semantic Web, для преобразования потоков информации между медиаторами.

Уровни интеграции данных

Проблема интеграции данных чрезвычайно многоаспектна и многообразна. Сложность и характер используемых методов ее решения существенным образом зависят от уровня интеграции, который необходимо обеспечить, свойств отдельных источников данных и всего множества источников в целом, требуемых способов интеграции. Системы интеграции данных могут обеспечивать интеграцию данных на физическом, логическом и семантическом уровне. Интеграция данных на физическом уровне с теоретической точки зрения является наиболее простой задачей и сводится к конверсии данных из различных источников в требуемый единый формат их физического представления. Интеграция данных на логическом уровне предусматривает возможность доступа к данным, содержащимся в различных источниках, в терминах единой глобальной схемы, которая описывает их совместное представление с учетом структурных и, возможно, поведенческих (при использовании объектных моделей) свойств данных. Семантические свойства данных при этом не учитываются. Поддержку единого представления данных с учетом их семантических свойств в контексте единой онтологии предметной области обеспечивает интеграция данных на семантическом уровне.

Семантическая интеграция основывается на знании и учете природы данных. Разумеется, хранение данных вместе с метаданными создает дополнительные сложности, но обеспечивает большее удобство работы. Понимание этой истины пришло совсем недавно – ему мешало то, что много лет назад создатели первых компьютеров нашли удачное, как им казалось, решение, отделив данные от контекста и получив возможность работы с данными в чистом виде. В итоге сложилось своеобразное компьютерное «раздвоение сознания»: данные – в машине, а контекст – в человеке.

Наиболее распространенный подход к семантической интеграции данных основан на использовании семантических посредников (Mediators). Средствами посредников поддерживаются унифицированные метаописания интегрируемых источников данных. Как правило, семантические посредники разрабатываются для конкретной узкой предметной области. Механизмы посредников опираются на онтологические спецификации источников.

В последние годы появился ряд работ, посвященных решению проблемы семантической интеграции данных из множества источников, в которых для представления глобальной схемы в системе интеграции данных предлагается использовать аппарат дескриптивных логик, воплощенный в языке описания онтологий OWL (Ontology Web Language).

Дескрипционные логики представляют собой семейство языков, позволяющих формально и однозначно описывать понятия в какой-либо предметной области. Каждый класс («концепт») может быть соотнесен с другим подобным ему концептом путем добавления тэгов метаданных, указывающих на свойства, общие черты, различия и т.д. Расширение моделей тэгами позволяет создавать такие структуры, которых раньше не могло быть. В семантической модели любая информационная единица представляется графом, что упрощает ее модернизацию; например, слияние двух моделей сводится к объединению их графов. Информационная единица может быть представлена идентификатором Uniform Resource Identifier (URI), посредством которого могут быть установлены отношения между двумя или большим числом информационных единиц.

На рисунке 1 ниже показана семантическая модель информации.

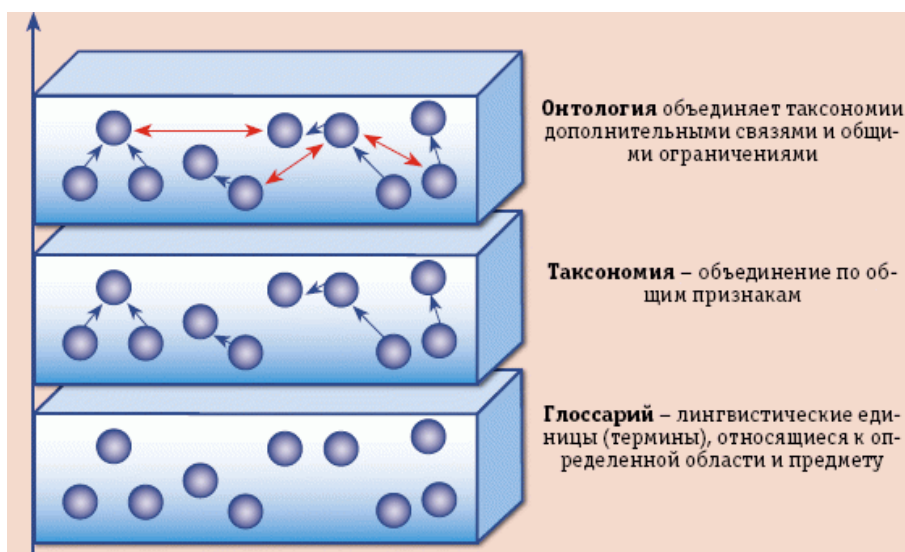


Рис. 1. Семантическая модель информации

На нижнем, атомарном уровне представления данных располагается глоссарий, далее идут таксономии, представляющие собой аналог схемы, в отличие от глоссария, образованного простой иерархией элементов данных. Каждый элемент контента должен представлять свою принадлежность к определенной таксономии. Следующий уровень – онтология (глоссарий и таксономия), представляет собой структуру, выражающую одновременно иерархию и набор отношений между элементами словника в этой иерархии.

Семантическое множество объединяет онтологии, таксономии и глоссарии, входящие в состав корпоративной системы.

Современные подходы к интеграции данных

Существует несколько основных подходов к интеграции данных в организациях:

- Консолидация данных;
- Федерализация данных;
- Распространение данных;
- Гибридный подход;
- Сервисный подход.

В соответствии с классификацией К. Диттриха, интеграцию данных возможно производить на нескольких уровнях архитектуры ИС организации:

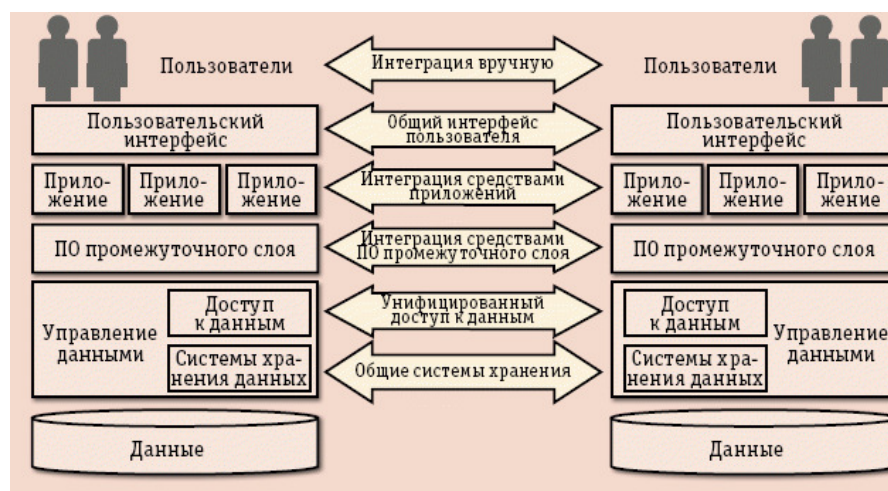


Рис. 2. Уровни архитектуры информационной системы организации, на которых производится интеграция

Каждый из перечисленных выше подходов подразумевает проведение операций, затрагивающих все или часть уровней интеграции по К. Диттриху.

Консолидация данных

При использовании этого метода данные собираются из нескольких первичных систем и интегрируются в одно постоянное место хранения. Такое место хранения может быть использовано для подготовки отчетности и проведения анализа, как в случае Хранилища данных, или как источник данных для других приложений, как в случае операционного склада данных. Процесс заполнения Хранилища состоит из трех фаз — извлечение, преобразование, загрузка (Extract, Transformation, Loading, ETL). Во многих случаях именно ETL понимают под термином «интеграция данных». Еще одна

распространенная технология консолидации данных — управление содержанием корпорации (Enterprise Content Management, ECM). Большинство решений ECM направлены на консолидацию и управление неструктурированными данными, такими как документы, отчеты и web-страницы. Схема консолидации данных показана ниже на рисунке 3.



Рис. 3. Консолидация данных

При использовании этого метода обычно существует некоторая задержка между моментом обновления информации в первичных системах и временем, когда данные изменения появляются в конечном месте хранения. В зависимости от потребностей бизнеса такое отставание может оставлять несколько секунд, часов или много дней. Термин "режим, приближенный к реальному времени" часто используется для описания конечных данных, обновление которых отстает от источника на несколько секунд, минут или часов. Данные, не отстающие от источника, считаются данными в режиме реального времени, но это трудно достижимо при использовании метода консолидации данных.

Конечные места хранения данных, содержащие данные с большими временами отставания (например, более одного дня), создаются с помощью пакетных приложений интеграции данных, которые извлекают данные из первичных систем с определенными, заранее заданными интервалами. Такой подход использует запросы к данным, которые получают периодические "мгновенные снимки" первичных данных. Хотя подобные запросы получают текущие данные, они не отражают тех изменений, которые произошли между двумя последовательными запросами. А за это время данные могли обновляться несколько раз.

Конечные места хранения данных с небольшим отставанием обновляются с помощью оперативных приложений интеграции данных, которые постоянно отслеживают и передают изменения данных из первичных систем в конечные места хранения. Такой подход требует от приложений консолидации данных, чтобы они могли идентифицировать те изменения данных, которые необходимо зафиксировать для

консолидации. Для этого обычно используются определенные формы метода захвата изменений данных. В данном случае в результате выполнения задачи по захвату изменений будут получены все изменения, которые произошли в первичных данных.

Методы извлечения и передачи могут использоваться вместе. Например, оперативное приложение передачи данных может накапливать изменения данных в какой-то области промежуточного хранения, а пакетное приложение извлечения данных может обращаться к нему через определенные интервалы. При этом важно понимать, что метод передачи зависит от того, происходят ли определенные события, а метод извлечения работает по требованию.

Бизнес-приложения, которые обрабатывают консолидированный склад данных, могут генерировать запросы к этим данным, создавать отчеты на их основании и проводить анализ данных. Как правило, эти приложения не могут производить обновления консолидированных данных из-за проблем, связанных с синхронизацией подобных обновлений с первичными системами данных. Тем не менее, некоторые программные продукты для интеграции данных все же предлагают возможности записи, обеспечивая средства решения конфликтов данных, которые могут иметь место между обновленными данными в консолидированном складе и первичными системами.

Некоторые приложения производят обновления консолидированного склада данных и передают эти изменения назад в первичные системы. Примером такой системы является конечный склад данных, который используется для создания еженедельной модели ценообразования. Модель может быть оптимизирована и обновлена в течение недели, а затем снова загружена в первичную систему в начале следующей недели.

Преимуществом консолидации данных является то, что этот подход позволяет осуществлять трансформацию значительных объемов данных (реструктуризацию, согласование, очистку и/или агрегирование) в процессе их передачи от первичных систем к конечным местам хранения. Некоторые сложности, связанные с данным подходом, - это значительные вычислительные ресурсы, которые требуются для поддержки процесса консолидации данных, а также существенные ресурсы памяти, необходимые для поддержки конечного места хранения. Но с учетом постоянно совершенствования аппаратных средств это не проблема.

Федерализация данных

Федерализация данных обеспечивает единую виртуальную картину одного или нескольких первичных файлов данных. Если бизнес-приложение генерирует запрос к этой виртуальной картине, то процессор федерализации данных извлекает данные из соответствующих первичных складов данных, интегрирует их таким образом, чтобы они

отвечали виртуальной картине и требованиям запроса, и отправляет результаты бизнес-приложению, от которого пришел запрос. По определению, процесс федерализации данных всегда заключается в извлечении данных из первичных систем на основании внешних требований. Все необходимые преобразования данных осуществляются при их извлечении из первичных файлов. Интеграция корпоративной информации (Enterprise Information Integration, ЕИ) - это пример технологии, которая поддерживает федеративный подход к интеграции данных.

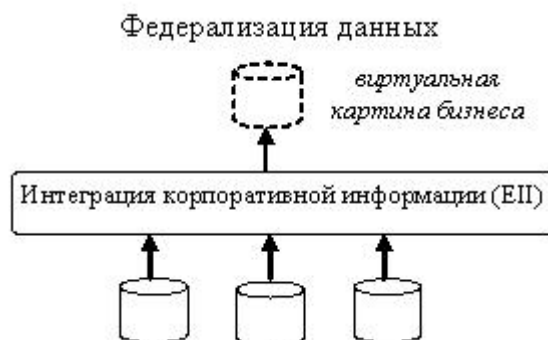


Рис. 4. Федерализация данных

Один из ключевых элементов федеративной системы - это метаданные, которые используются процессором федерализации данных для доступа к первичным данным. В некоторых случаях эти метаданные могут состоять исключительно из определений виртуальной картины, которые ставятся в соответствие ("мэппируются") первичным файлам. В более передовых решениях метаданные также могут содержать детальную информацию о количестве данных, находящихся в первичных системах, а также о путях доступа к ним. Такая расширенная информация может помочь федеративному решению оптимизировать доступ к первичным системам.

Некоторые федеративные решения могут обеспечивать дополнительные бизнес-метаданные, которые отражают семантические связи между элементами данных в первичных системах. Примером таких данных являются данные о потребителях. Метаданные могут содержать общий индикатор потребителя, который ставится в соответствие ("мэппируется") различным ключевым элементам данных о потребителе в первичных системах.

Считается, что основное преимущество федеративного подхода - тот факт, что он обеспечивает доступ к текущим данным и избавляет от необходимости консолидировать первичные данные в новом складе данных. Но следует помнить, что федерализация данных не очень хорошо подходит для извлечения и согласования больших массивов данных или для тех приложений, где существуют серьезные проблемы с качеством

данных в первичных системах. Еще один существенный фактор - потенциальное влияние на производительность и дополнительные затраты на доступ к многочисленным источникам данных во время выполнения программы.

Федерализацию данных возможно использовать в тех случаях, когда стоимость консолидации данных перевешивает бизнес-преимущества, которые она предоставляет. Оперативная обработка запросов и подготовка отчетов могла бы служить примером подобной ситуации. Федерализация данных также, вероятно, оказалась бы полезной в тех случаях, когда политика безопасности данных и лицензионные ограничения запрещают копирование данных первичных систем. Обычно в эту категорию попадают синдикаты данных. Помимо этого, федерализация могла бы использоваться как кратковременное решение для интеграции данных после приобретения или слияния компаний. Но в целом, как показывает опыт, даже в названных выше ситуациях консолидация данных часто оказывается более приемлемым решением, чем федерализация.

Изучение и профилирование первичных данных, необходимые для федерализации, несильно отличаются от аналогичных процедур, требуемых для консолидации. Таким образом, организациям стоит использовать такие продукты для интеграции данных, которые поддерживают как федерализацию, так и консолидацию, или, по крайней мере, продукты, которые могут обеспечивать совместное использование метаданных, необходимых для обоих подходов.

Федеративная архитектура очень полезна для крупных транснациональных корпораций и является весьма удобным подходом для поддержания баланса между необходимостью автономии местных подразделений компании и их гибкости, с одной стороны, и стандартизации и централизованного контроля, которые осуществляет центральный офис, - с другой. При этом под федеративным Хранилищем может иметься в виду как единое физическое федеративное Хранилище, так и федерация более мелких специализированных Хранилищ данных.

Необходимо отметить, что в англоязычной литературе термин *federated data warehouse* сейчас используется в двух разных значениях. Часть специалистов подразумевает под федеративным Хранилищем создание виртуальной структуры, оперирующей с выборками данных. Другие называют федеративным Хранилищем единый физический репозиторий, работающий с копиями данных, который другими словами может быть назван распределенным Хранилищем.

Распространение данных

Приложения распространения данных осуществляют копирование данных из одного места в другое. Эти приложения обычно работают в оперативном режиме и

производят перемещение данных к местам назначения, т.е. зависят от определенных событий. Обновления в первичной системе могут передаваться в конечную систему синхронно или асинхронно. Синхронная передача требует, чтобы обновления в обеих системах происходили во время одной и той же физической транзакции. Независимо от используемого типа синхронизации, метод распространения гарантирует доставку данных в систему назначения. Такая гарантия - это ключевой отличительный признак распространения данных. Большинство технологий синхронного распространения данных поддерживают двусторонний обмен данными между первичными и конечными системами. Примерами технологий, поддерживающих распространение данных, являются интеграция корпоративных приложений (Enterprise application integration, EAI) и тиражирование корпоративных данных (Enterprise data replication, EDR).



Рис. 5. Распространение данных

Большим преимуществом метода распространения данных является то, что он может быть использован для перемещения данных в режиме реального времени или близком к нему. Другие достоинства включают гарантированную доставку данных и двустороннее распространение данных. Доступность многих из этих удобств зависит от конкретного продукта. Метод распространения данных может также использоваться для уравнивания рабочей нагрузки, создания резервных копий и восстановления данных, в том числе в случае чрезвычайных ситуаций.

Практическое применение этого метода отличается достаточно большим разнообразием как в плане производительности, так и в отношении возможностей реструктуризации и очистки данных. Некоторые корпоративные продукты распространения данных могут поддерживать перемещение и реструктуризацию крупных массивов данных, тогда как продукты EAI часто имеют ограниченные возможности передвижения большого количества данных и их реструктуризации. Одна из причин

подобного различия - тот факт, что в центре архитектуры тиражирования корпоративных данных лежат данные, а в центре технологии EAI - сообщения или транзакции.

Гибридный подход

Методы, используемые приложениями интеграции данных, зависят как от нужд бизнеса, так и от технологических требований. Достаточно часто приложение интеграции данных использует так называемый гибридный подход, который включает несколько методов интеграции. Хороший пример такого подхода - интеграция данных о клиентах (customer data integration, *сокр.* CDI), целью которой является обеспечение согласованной картины информации о клиентах.

Самый простой подход к CDI - это создание консолидированного склада данных о клиентах, который содержит данные, полученные из первичных систем. Отставание информации в консолидированном складе будет зависеть от режима консолидации данных (оперативный или пакетный) и от частоты обновления этой информации.

Другой подход к CDI - это федерализация данных, когда определяются виртуальные бизнес-представления данных о клиентах в первичных системах. Эти представления используются бизнес-приложениями для доступа к текущей информации о клиентах в первичных системах. При федеративном подходе также может использоваться справочный файл метаданных для связи информации о клиентах на основе общих ключевых элементов.

Гибридный подход, использующий как консолидацию, так и федерализацию данных, также может иметь место. Общие данные о клиентах (имя, адрес и т.д.) могут быть консолидированы в одном складе, а данные, которые относятся к определенному первичному приложению (например, заказы), могут быть федерализованы. Такой гибридный подход может быть расширен за счет распространения данных. Если клиент обновляет свое имя и адрес во время транзакции в Интернет-магазине, то эти изменения могут быть отправлены в консолидированный склад данных, а оттуда распространены в другие первичные системы, такие как база данных о клиентах розничного магазина.

Пример гибридного подхода показан ниже на рисунке 6.

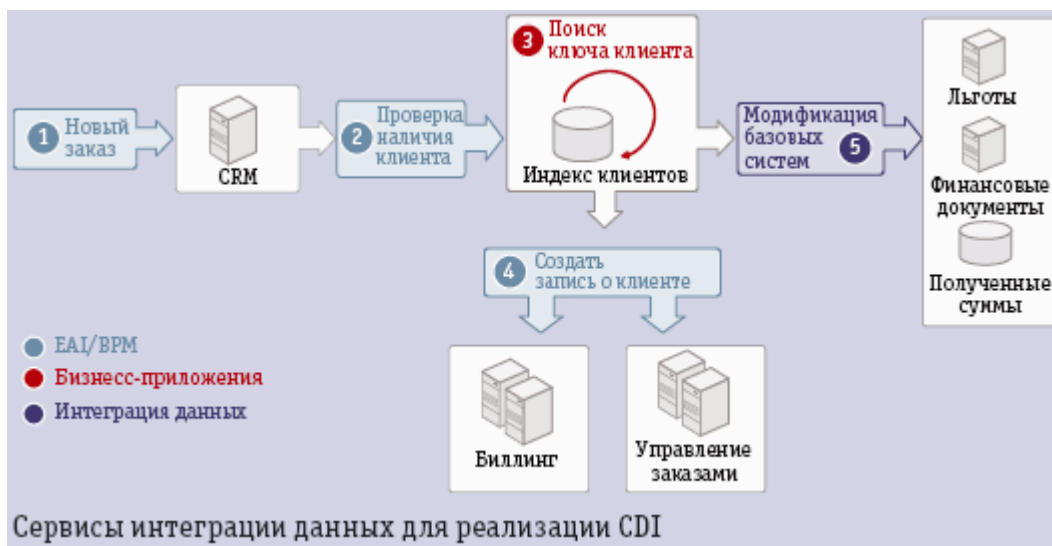


Рис. 6. Распространение данных

Из рисунка видно, что для построения системы интеграции применяются различные технологии, такие, как EAI, BPM, SOA и т.д.

SOA – Service Oriented Architecture

Поддержка XML-преобразований, Web-сервисов, связи с источниками данных с использованием таких спецификаций, как JDBC или JMS, подготовили технологии интеграции данных к реализации на базе сервисной архитектуры. Появилось понятие сервисов данных, которые объединяются в отдельный уровень, абстрагирующий бизнес-логику от технологии доставки данных приложениям из различных источников и преобразования данных. Если в SOA традиционные прикладные сервисы инкапсулируют логику бизнес-приложений, позволяя использовать ее многократно в различных ситуациях, то, аналогично, уровень сервисов данных позволяет инкапсулировать все рассмотренные функциональные возможности технологий интеграции данных в компоненты, доступные для многократного использования различными приложениями и в разных сценариях интеграции данных. Наличие уровня сервиса данных дает возможность развертывать в среде выполнения все функции интеграции данных в виде сервисов, управлять их публикацией и тестированием, обеспечить взаимодействие с реестром/репозиторием сервисов приложений, реализовать сервисные принципы в организации сред разработки и администрирования процессов интеграции данных.

Концепция сервисного подхода к управлению данными, включая задачу их интеграции, в ряде источников получила название Information as a Service, а в отношении единой платформы управления данными на базе сервисов можно встретить такие обозначения, как «информационная фабрика» (Information Fabric) или «информационный сервер» (Information Server). Полноценной реализации подобного сервера, как отмечают

аналитики Forrester Research, пока нет, но решения таких компаний, как BEA Systems, IBM и Oracle, эволюционируют именно в этом направлении.

В идеале сервисная архитектура интеграции данных должна включать в себя три основных компонента: среду универсального доступа к данным, репозиторий и сервисы метаданных, а также сервисы интеграции. Общая концепция показана на рисунке 7.

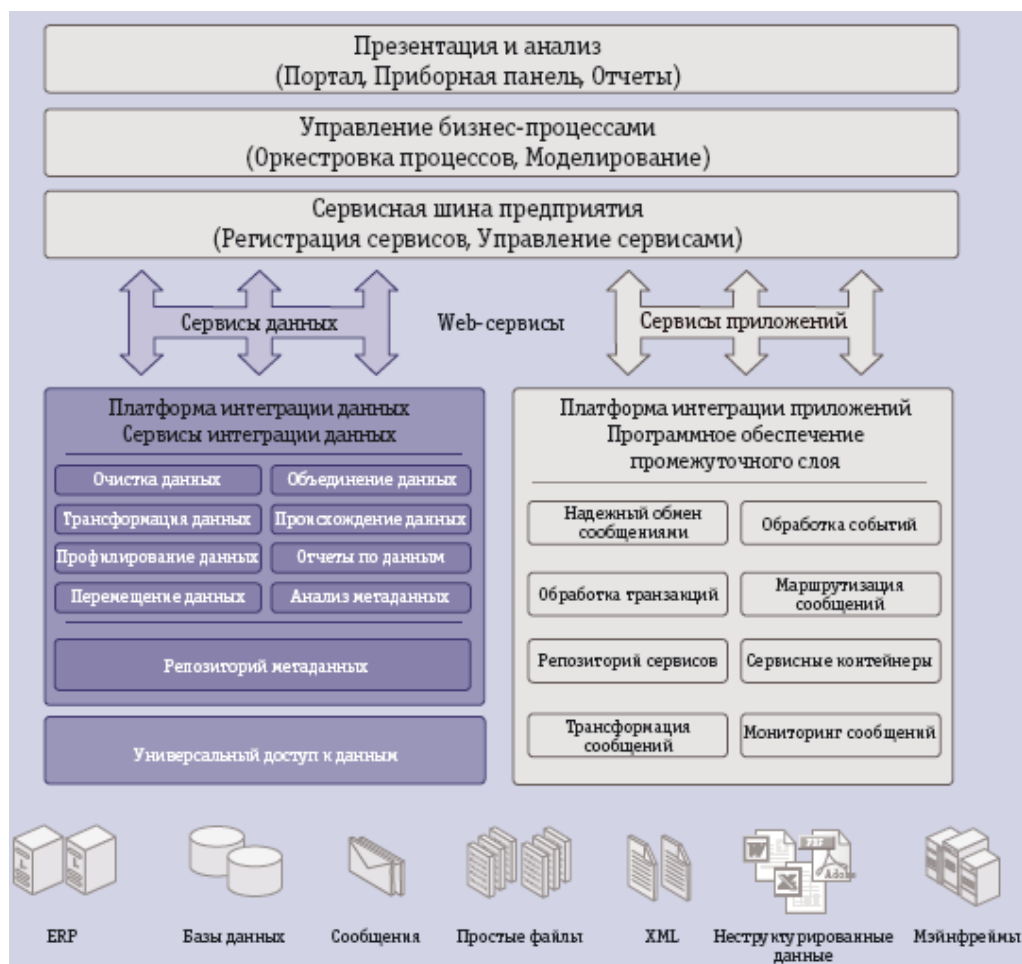


Рис. 7. Service Oriented Architecture

Среда универсального доступа к данным реализует связь с произвольными источниками данных и предоставляет бизнес-сервисам данные различных типов в нужных режимах доставки. Универсальность доступа подразумевает поддержку в организации полного спектра возможных источников данных, включая готовые бизнес-системы, унаследованные решения на платформе мейнфреймов, реляционные базы данных, неструктурированные данные.

Центральную роль в сервисной системе интеграции данных играет среда метаданных. Собранные в едином корпоративном репозитории, метаданные позволяют задать общий язык определения данных и добиться единого понимания данных на

различных уровнях их использования. И хотя любая из систем, скажем, СУБД или информационный портал, имеет свое хранилище метаданных, речь идет об общем представлении для всех таких решений, позволяющем определить на едином языке (например, XML) семантику и правила использования корпоративных данных в целом.

Репозиторий метаданных включает в себя определения данных, правила преобразования в ETL-системах, отображения из источника данных в целевое хранилище, логические модели данных и другие характеристики, описывающие, как искать, осуществлять доступ и использовать данные. Репозиторий позволяет задать единую семантику и правила ко всем процессам организации, отслеживания и управления данными в компании и, подобно универсальному брокеру, связывает высокоуровневые сервисы данных и более детальные определения и зависимости данных. Основная ценность такого репозитория — в абстрагировании логических моделей данных от фактической реализации технологий интеграции данных. Сервисы метаданных позволяют согласовывать семантику данных между различными системами, выявлять происхождение и историю данных для отчетности и аудита, отслеживать факт внесения изменений, обнаруживать связанные с данными слабые места в общей информационной архитектуре. В целом среда метаданных призвана обеспечить надлежащую структуру и качество данных, прежде чем они будут переданы соответствующему бизнес-сервису.

Платформа интеграции позволяет реализовать в виде сервисов интеграции данных основные функции интеграции — режимы доставки, способы трансформации данных, средства профилирования и очистки данных и т.д. К необходимым свойствам такой платформы относится также разбиение данных для интеграции на высокопроизводительных многопроцессорных или распределенных платформах, обеспечение восстановления при сбоях и безопасность работы с данными, например, аутентификация и авторизация, протоколы управления учетными записями и шифрование.

Существующие решения

Аналитики Gartner среди участников рынка систем интеграции данных выделяют такие компании, как Oracle, SAP, Microsoft, Business Objects, SAS Institute, Sybase, Cognos и Tibco Software. Некоторые из них относятся к так называемым нишевым игрокам; их предложения не отличаются универсальностью, не реализуют всех ключевых аспектов интеграции данных и не могут быть причислены к решениям корпоративного уровня. В число лидеров, предлагающих реальную конвергенцию отдельных инструментов в единую платформу интеграции данных, с отрывом от остальных, попали только IBM и Informatica.

В IBM одними из первых сделали ставку на технологии федеративного объединения данных в системе DB2 Information Integrator (сегодня Federation Server). После приобретения в 2005 году компании Ascential гигант вышел в лидеры ETL-решений. Группа продуктов по интеграции данных от IBM относится к семейству WebSphere и включает в себя Federation Server, ETL-систему DataStage, системы Information Analyzer и Quality Stage для профилирования и очистки данных, сервер метаданных Metadata Server. Недавно все эти продукты вошли в состав нового пакета IBM Information Server, реализующего единую, построенную на принципах многократного использования сервисов архитектуру интеграции данных и управления разнородной информацией. Ядром этой архитектуры является среда управления метаданными, пользователям предоставляется возможность многофункциональной работы с данными посредством ролевых интерфейсов, имеются мощные механизмы параллельной обработки больших объемов данных.

Informatica, также ветеран рынка интеграции данных, является крупнейшим поставщиком, специализирующимся только на решениях этого класса. Функции интеграции сосредоточены в системе Informatica PowerCenter, которая изначально была разработана для реализации возможностей ETL, но со временем (с приобретением других компаний, лицензированием и разработкой новых технологий) пополнилась возможностями федеративного объединения данных (Data Federation Option), профилирования и обеспечения качества данных (Data Profiling Option, Data Cleanse and Match Option), поддержки неструктурированных источников (Unstructured Data Option) и управления метаданными (Metadata Exchange Option и Metadata Manager).

В версии системы PowerCenter 8.5 реализована единая платформа интеграции данных с централизованной инфраструктурой метаданных и функциями интеграции в качестве сервисов. PowerCenter обеспечивает доступ к данным реляционных баз, файлов, унаследованных систем на мэйнфреймах, источникам неструктурированных данных и очередей сообщений, поддерживая пакетную доставку данных, доставку в реальном времени и в режиме извлечения только модифицированных данных. Модуль Metadata Exchange Option координирует технические и бизнес-метаданные из инструментария моделирования данных, систем бизнес-анализа, каталогов исходных и целевых баз данных и репозиториях самой системы PowerCenter. Решение Metadata Manager, входящее в состав расширенной версии PowerCenter Advanced Edition, обеспечивает сбор метаданных из различных источников, а также средства анализа метаданных и отчетности по ним.

Для масштабной интеграции данных, при которой требуется обработка большого объема данных, в том числе в реальном времени, PowerCenter включает опцию поддержки

корпоративных grid-инфраструктур, построенных их многоузловых кластеров на базе стандартных Intel-серверов или серверов-лезвий. Enterprise Grid Option включает в себя консоль управления grid-средой с возможностью ее конфигурирования, контроля и миграции узлов и сервисов данных, сложный алгоритм адаптивной балансировки нагрузки и средства динамического разбиения данных для оптимальной параллельной интеграции.

Аналитики отмечают, что компания Informatica делает упор на интеграцию данных в процессе межкорпоративного взаимодействия, что пока не свойственно большинству участников рынка.

Список литературы

1. Брешенков А.В. Базы данных. Проектирование баз данных на основе информации табличного вида. LAP LAMBERT Academic Publishing GmbH & Co. KG Dudweiler, rbr, 66123 Saarbrücken, Germany, 2011, 394 с.
2. М.Р. Когаловский. Методы интеграции данных в информационных системах - Сб. трудов Третьей Всероссийской конференции "Стандарты в проектах современных информационных систем". Москва, 23-24 апреля 2003 г.
3. Н. Дубова. Краткий курс интеграции данных – Открытые системы, 9, 2007.
4. Л. Черняк. Интеграция данных: синтаксис и семантика – Открытые системы, 10, 2009.