

03, март 2016

УДК 004.622

Проблемы интеграции данных и способы их решения на примере служб SQL Server Integration Services

Бутылева Ю.М., студент

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

Научный руководитель: Ревунков Г.И., к.т.н., доцент

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

revunkov@bmstu.ru

Введение

Практически во всех сферах деятельности необходимо использование баз данных. В частности, крупные корпорации и другие организации обладают несколькими базами данных, связанными с различными видами деятельности. В таких базах могут храниться несогласованные данные или данные с разными представлениями. Если у корпорации или крупной компании появится необходимость осуществления анализа всех хранящихся данных, то их объединение вызовет ошибки.

Так, в одном из интервью Михаил Корнаухов, генеральный директор компании «ИнтэлЛекс» [1], сообщил, что для управления важна согласованность данных в разных системах. Иначе это приводит к значительным убыткам и ухудшает результаты деятельности компаний и организаций в целом. Отсюда возникает одна из важнейших проблем баз данных – интеграция данных. В данной статье рассматриваются основные проблемы при интеграции данных и способы их решения на примере SQL Server Integration Services.

Определение интеграции данных

Под интеграцией данных понимают объединение и представление данных в единообразном виде. При этом это могут быть не только разнородные данные, но и содержащие данные системы. Потребность в применении интеграции возрастает с ростом объема данных.

Существуют следующие уровни интеграции данных: физический, логический и семантический. На физическом уровне происходит замена представления данных из различных источников в единый физический формат. Доступ к данным, содержащимся в разных источниках, но представленным в единой схеме, осуществляется на логическом уровне. Роль семантического уровня заключается в обеспечении представления данных с помощью концептуальной схемы предметной области.

Основные проблемы интеграции данных

Согласованность данных

Основные проблемы при интеграции данных возникают при согласовании многочисленных источников с различными форматами данных, вследствие отсутствия единой системы управления данными. Кроме того, данные могут быть представлены в различном виде и предоставляться разными источниками в различные промежутки времени, что также затрудняет интеграцию. Отсюда необходимо понимать, какого качества будут данные, их форматы, и уметь их преобразовать в понятный формат. Вдобавок перед выполнением процессов интеграции важно уделить внимание анализу и очистке данных [2].

Например, во многих информационных системах компаний может не учитываться качество данных, из-за этого получается система со своим набором данных. Может получиться, что данные являются разными по представлению (например, “Koln” и «Köln»).

В крупных компаниях или корпорациях происходит обмен огромными объемами данных в реальном времени. Поэтому появляется необходимость постоянной интеграции данных, которая удовлетворяет быстрой загрузке данных и загрузке во множество систем. Также потребность в дополнительной обработке данных может возрасти, например, при использовании в системах ETL (Extract (извлечение), Transform (преобразование), and Load (загрузка)) различных источников данных. Но при применении ETL для решения задач могут появиться серьёзные ограничения.

Технические трудности

Важным аспектом при интеграции данных является процесс её организации. Так как ошибочная организация будет становиться неконтролируемой, и цена её исправления и использования возрастет.

В начале процесса организации подбирается платформа интеграции данных. Затем выбираются те аспекты, которые необходимы для решения проблемы. При интеграции не существует определенного подхода для всех задач. Следует понимать, что не стоит использовать дорогие решения для организации простых задач, и иметь четкое представление об используемой платформе во избежание неправильного её использования [2].

Для решения этих и других проблем существует множество продуктов от таких компаний, как Oracle, Microsoft, SAP, Business Objects и др. Каждый продукт имеет свои достоинства и недостатки и выбирается в зависимости от поставленной задачи и имеющейся инфраструктуры. По сравнению с остальными решениями наиболее простым в использовании и эффективным можно выделить решение от Microsoft. Кроме того, оно обеспечивает быструю интеграцию в инфраструктуру компаний и может обойтись дешевле других продуктов. Далее в статье будут рассмотрены службы Integration Services и способы решения проблем интеграции с их помощью.

SQL Server Integration Services

Службы MS SQL Server Integration Services (SSIS) представляют собой платформу для принятия решений по интеграции и преобразованию данных. Это эффективное средство для выборки, преобразования, загрузки, а так же выгрузки и проверки данных из разнородных источников разными адресатами. Integration Services предоставляет различные инструментальные средства, включающие средства для построения и выполнения пакетов. Кроме того, службы включают в себя графические инструменты, облегчающие работу.

В предыдущих версиях SQL Server (например, SQL Server 2000) использовалась служба преобразования данных (Data Transformation Services, DTS), в то время как на современных платформах, начиная с 2005, применяются Integration Services. Различия в этих двух службах очень значительны, так как в службах Integration Services ничего не осталось от DTS. А именно: изменились инструменты для создания и администрирования пакетов, их форматы и многое другое.

Архитектура SQL Server Integration Services

Для управления операциями задач потока данных в службах SSIS задействуется конвейер с буферами памяти. Применение такого способа управления позволяет осуществлять преобразование данных не только как в ETL системах, с применением

промежуточных результатов, но и без их применения. Службы SSIS состоят из четырех компонентов, показанных на рис. 1.: служба SSIS, объектная модель SSIS, механизм данных и компоненты данных SSIS, рабочий механизм и рабочие компоненты SSIS.

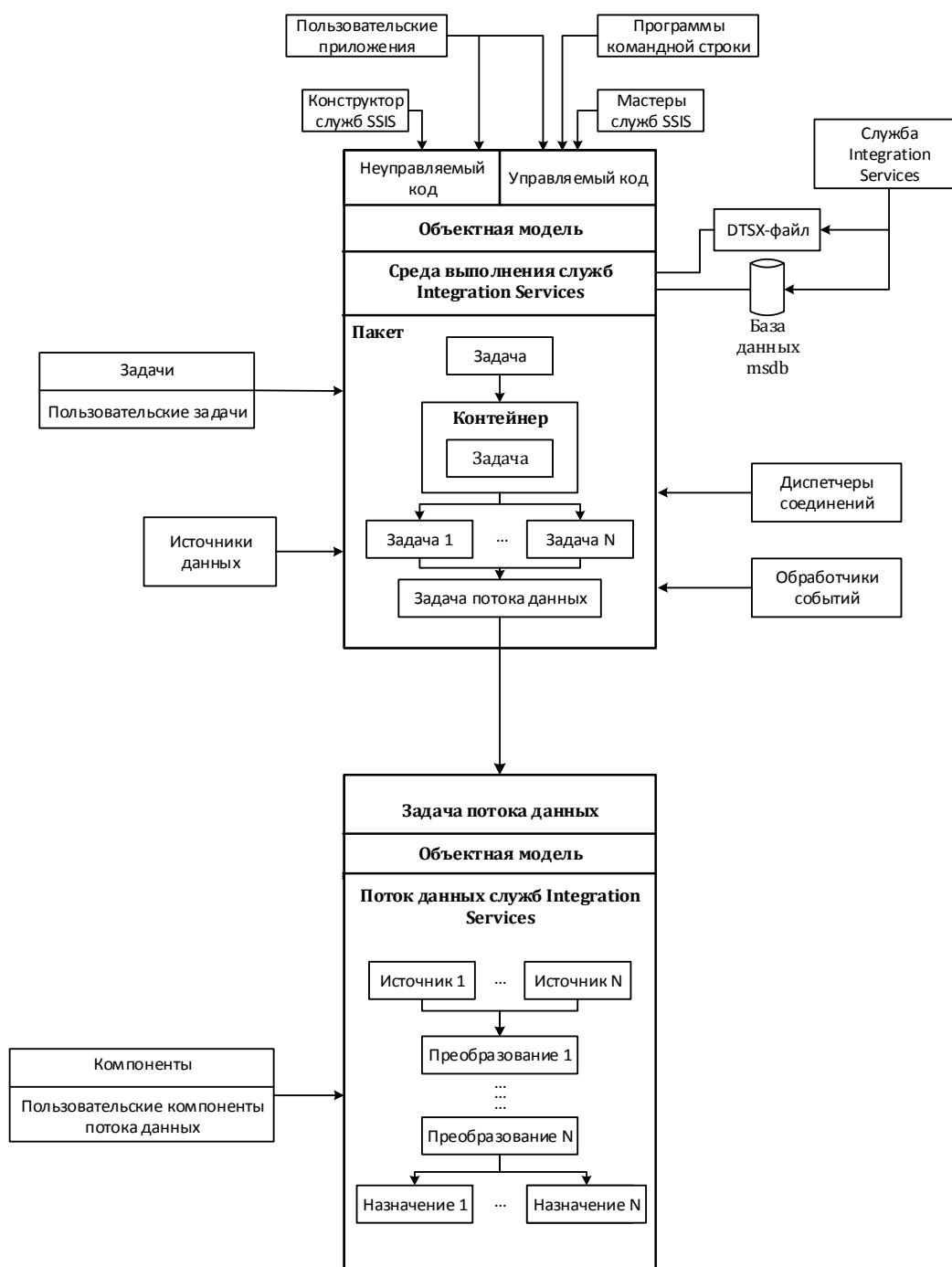


Рис. 1. Архитектура служб SSIS

Основным компонентом служб Integration Services является объектная модель. Она предоставляет доступ к средствам служб SSIS, утилитах командной строки и разработанным приложениям по средствам управляемого программного интерфейса.

Составление пакета, выполнение транзакций, запуск и обеспечение функциональности гарантирует рабочий механизм и рабочие компоненты служб Integration Services.

За обеспечение переноса данных в памяти от источника к адресату отвечает механизм и компоненты потока данных. Вместе с тем, он выполняет преобразования данных. Для добавления в пакет потока данных необходимо, чтобы поток управления пакета включал задачу потока данных. Задача создает, упорядочивает и запускает поток данных и является исполняемым объектом пакета. Механизм потока данных состоит из источников, преобразований и назначений (целевые объекты). Источники предоставляют соединение и точные данные из всех видов источников (таблицы, представления, файлы, XML-файлы и т.д.). Преобразования перед отправкой адресату осуществляют преобразование исходных данных, используя поиск, слияние, манипулирование т.д. Назначение – один из компонент потока данных, записывающий данные из потока в хранилище и создающий набор данных в памяти [3].

Пакеты SSIS

Пакетом SSIS называется объект, выполняющий функциональность служб SSIS. Пакет создается с помощью SQL Server Data Tools. После создания пакет представляется пустым объектом, который можно расширить с помощью добавления в него хотя бы одного элемента потока управления и необязательных элементов потоков данных.

Элементы потока управления осуществляют поддержку структуры элементов, управление последовательностью их выполнения. Существуют следующие три типа элементов потока управления:

- контейнеры (обеспечивающие структуру пакетов)
- задачи (отвечают за функциональность)
- элементы управления очередностью (соединяют выполняемые элементы в упорядоченный поток управления)

На рис. 2. представлен пример потока управления с одним контейнером, тремя задачами пакетного уровня и одной задачей уровня контейнера.

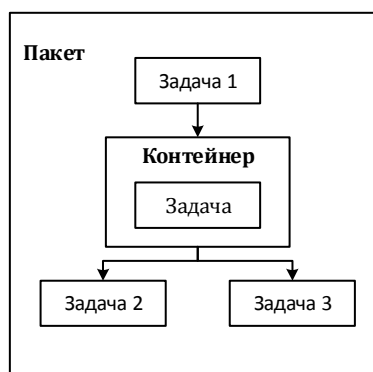


Рис.2. Пример потока управления

Элементы потока управления данных занимаются извлечением, загрузкой и изменением данных в источники данных [3].

Решение проблем интеграции с помощью SQL Server Integration Services

Одним из популярных способов использования служб Integration Services является способ, при котором предполагается совместное использование данных, полученных из различных источников (OLE DB, ODBC, Excel, XML и др.) и хранящихся в различных системах хранения. Для этого в SSIS есть набор компонент, называемых адаптерами. Кроме того, SSIS также поддерживает индивидуальные адаптеры, созданные пользователями. Благодаря этому данные загружаются в источник данных, который включается в поток данных служб SSIS. При совместном использовании данных, полученных из различных источников, происходит преобразование их типа, а так же ищется вспомогательная информация.

Начиная с версии SQL Server 2008 R2 стали поддерживаться высокоэффективные коннекторы данных для переноса данных между различными платформами [4]. Благодаря коннекторам служба Integration Services может передавать данные на высоких скоростях и создавать пакеты, которые будут обладать большей гибкостью и лучшей производительностью.

Для борьбы с некачественными данными могут использоваться преобразования Fuzzy Lookup (Нечеткий Поиск) и Fuzzy Grouping (Нечеткие Группировки). Преобразование нечеткого поиска «очищает» данные, как, например, исправляет данные, осуществляет приведение данных или решение проблем пропущенных значений. Это преобразование находит соответствия между чистыми записями и записями входных данных. Преобразование нечетких группировок исследует данные из источника и группирует соответствующие друг другу строки. Оно стандартизирует записи из группы с

повторами и выбирает наиболее правильную запись. Основными преимуществами преобразований нечеткого поиска и группировок являются улучшение качества данных, повышение качества процессов очистки данных и уменьшение времени на разработку.

С ростом потребности в обнаружении закономерностей в данных и осуществлении прогнозирования на их основе, возникла потребность в Data Mining. Начиная с версии Enterprise, SQL Server может осуществить несколько стратегий поиска закономерностей в данных. Это позволило не только крупным компаниям, но и любым организациям, использовать интеллектуальный анализ.

Заключение

Интеграция данных играет важную роль в компаниях любого уровня. Очень важно правильно выбрать стратегию и платформу интеграции данных. Компания Microsoft предложила наиболее простой и эффективный и дешевый продукт для осуществления интеграции. Он способен решить основные проблемы и не требует глубоких знаний в использовании.

Список литературы

- [1]. Корнаухов М. Рынок движется к необходимости организации единых информационных ресурсов для транспортных компаний. Режим доступа: <http://www.cnews.ru/reviews/free/transport2006/int/organiz/> (дата обращения 10.11.2015).
- [2]. Ковтун М.В. Из практики интеграции информационных систем и данных. Избегаем трудностей. Режим доступа: http://www.prj-exp.ru/integration/integration_practice.php (дата обращения 11.11.2015).
- [3]. Knight B., Veerman E., Dickinson G., Hinson D., Herbold D. Professional SQL Server 2008 Integration Services. Indianapolis, Wiley Publ., 2008. 1010 p.
- [4]. Попов В. Интеграция данных. Режим доступа: <http://itband.ru/2010/09/integrationservices/#comments> (дата обращения 13.11.2015).