А.А. Кресов, В.В.Уваров

ПРИНЦИПЫ ИНТЕГРАЦИИ ДАННЫХ В СФЕРЕ НЕДРОПОЛЬЗОВАНИЯ

Приведен анализ трех основных подходов к интеграции разнородных информационных ресурсов: консолидация, федерализация, распространение. Предложены принципы разработки системы интеграции данных в сфере недропользования.

Недропользование, данные, интеграция, консолидация, федерализация, распространение, хранение, согласование данных, ETL, EII, SOA.

1. Состояние проблемы интеграции данных.

Значимость интеграции разнородных данных на текущий момент трудно переоценить. Для принятия адекватных и обоснованных управленческих решений необходимо информационное обеспечение этого процесса средствами хранения, накопления, анализа и интерпретации всех необходимых данных. Вместе с тем интегрированность, согласованность, непротиворечивость данных может быть эффективно обеспечена только при наличии единого централизированного источника информации вне зависимости от архитектуры и реализации информационной инфраструктуры организации. Построение такой инфраструтуры является длительным и трудоемким процессом, сложность которого напрямую зависит не столько от объемов накопленных ретроспективных данных, сколько от количества и разнообразия источников информации и приложений. Решающее значение в этом процессе, безусловно, играет правильный выбор методов и средств интеграции данных.

Можно выделить три основных подхода к интеграции данных [5]:

- консолидация (централизация хранения),
- федерализация (унификация доступа),
- распространение данных.

Для обоснованного выбора принципов интеграции данных в сфере недропользования проведем анализ этих подходов.

Консолидация данных. При использовании этого подхода данные извлекаются из нескольких источников и интегрируются в одно постоянное место хранения (хранилище данных). Хранилище данных может быть использовано для подготовки отчетности и проведения анализа с целью выявления закономерностей в данных или прогнозирования будущих результатов. Для данного подхода характерно пакетное извлечение данных из источника, преобразование между исходным и целевым форматом хранения и последующая загрузка в хранилище (рис. 1), что приводит к некоторой задержке обновления данных. Однако задержка в данном случае не столь критична, поскольку средства анализа и прогноза оперируют гораздо большими временными интервалами. Технология, применяемая при таком подходе, носит название ETL (Extract-Transform-Load).

Процесс интеграции проходит в три этапа: извлечение данных из источника (extract), преобразование между исходным и целевым форматом или структурой (transform) и загрузка в целевое место хранения (load). Наибольшую сложность представляет реализация второго этапа — преобразования. Под преобразованием (transform) современные ETL-системы подразумевают не только конвертацию из исходного формата в целевой, но и согласование, очи-

стку, проверку корректности данных, включая обращение к удаленным сервисам.

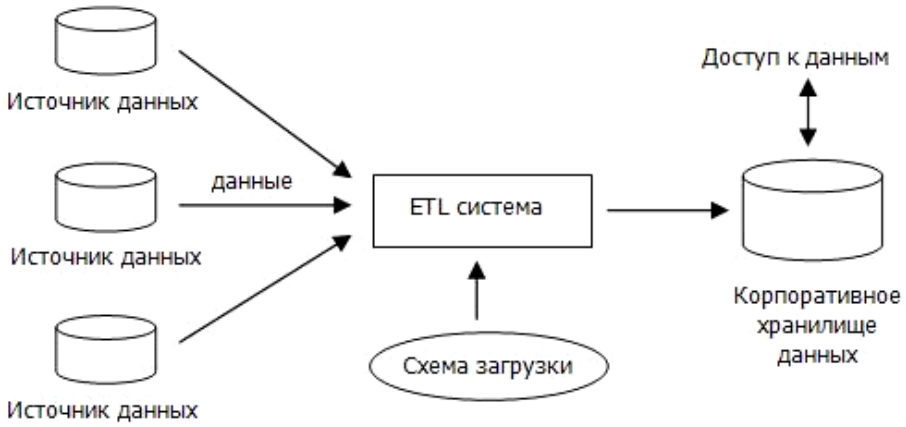

Рис. 1. Схема интеграции данных на основе хранилища

Для загрузки ETL-разработчик формирует некоторую схему, в которой устанавливаются соответствие между полями информационных сущностей источника и приемника данных (mapping), указываются применяемые к этим полям функции (в том числе и функции проверки данных, если таковые требуются), описываются дополнительные сценарии пре- и постпроцессинга. На основе данной схемы ETL-система формирует набор запросов на извлечение и вставку данных или высокоуровневый программный код, состоящий из вызова функций компонентов системы. Частным случаем такой схемы является специально разработанный модуль преобразования данных (конвертор), в котором соответствия между полями информационных сущностей и применяемые функции заложены в программном коде. Проектирование и разработка средств ETL превращается в сложную задачу даже тогда, когда используются решения, предлагаемые на рынке [1].

Основным недостатком описанного подхода является то, что данные необходимо копировать из источника в хранилище, осуществляя при этом сложные операции преобразования и согласования. Однако с ростом производительности вычислительной техники, развитием возможностей и гибкости ETL-систем временные затраты будут менее и менее существенными. *Вторым недостатком* является задержка в обновлении данных, поскольку загрузка в хранилище осуществляется пакетно, с некоторой периодичностью. Важнейшие *достоинства подхода* — высокая скорость обработки запросов, качество хранимых данных (прошедших согласование и очистку) простота и удобство процесса анализа данных.

Федерализация данных. Федеративный подход обеспечивает единое виртуальное видение разнородных источников данных. При этом данные фактически хранятся в разных по составу и структуре источниках (плоские файлы, веб-сайты, базы данных), информация в которых может частично дублироваться. Источники остаются полностью автономными. Интеграция данных сводится к интеграции схем хранения и созданию программного компонента (процессора федерализации), обеспечивающего прозрачный доступ к физически распределенным данным. Процессор федерализации анализирует поступающие от приложений запросы, переадресовывает их соответствующим

источникам, в которых эти данные могут храниться, интегрирует полученные данные в соответствии с виртуальной картиной и возвращает ответ приложению (рис. 2). Между источником данных и процессором, как правило, находиться программный компонент (адаптер), скрывающий детали реализации при доступе к источнику. Одним из примеров данного подхода является технология интеграции корпоративной информации (Enterprise information integration, EII) [5].

Важным элементом федеративной системы являются метаданные, которые используются процессором федерализации данных для доступа к источникам информации. В некоторых случаях эти метаданные могут состоять исключительно из определений виртуальной картины, которые ставятся в соответствие 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.

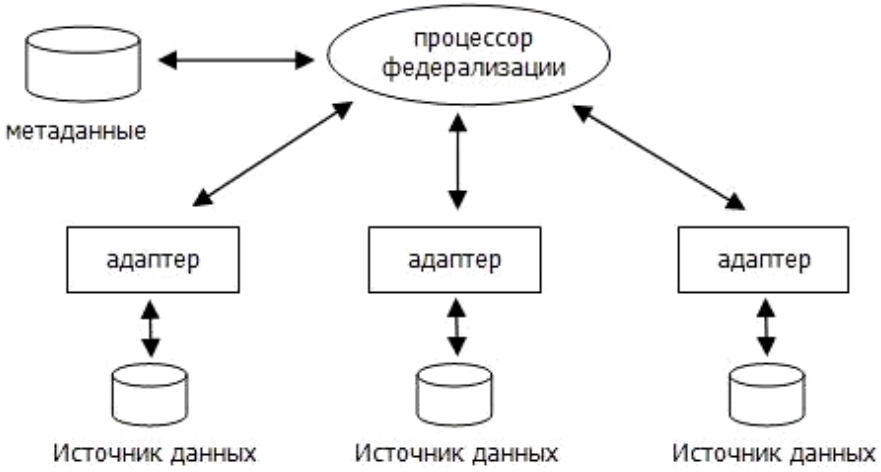

Рис. 2. Схема интеграции данных на основе федерализации

Основным преимуществом федеративного подхода является то, что доступ происходит к текущим данным, без задержек в обновлении, в отличие от подхода, основанного на создании единого хранилища данных. Этот *подход избавляет также* от необходимости копировать данные из источника в хранилище с использованием сложных систем согласования и загрузки.

Среди *недостатков* подхода можно выделить нелинейно возрастающую сложность реализации при увеличении числа источников (наличие существенных различий в модели данных может значительно усугубить ситуацию), высокие требования к качеству связи. При выполнении запросов могут возникать задержки, связанные с необходимостью обмена данными между источниками и процессором федерализации.

Распространение данных. Приложения распространения данных осуществляют передачу данных от источника к получателю. Изменения в источнике данных фиксируются и передаются получателю (подписчику) в оперативном режиме. Подписчик в свою очередь, получив сообщение, фиксирует из-

менения в своей базе данных. Таким образом достигается высокая оперативность в обновлении данных, фактически в режиме реального времени. Отличительным признаком такого подхода является гарантированная доставка данных от источника к получателю. Даже если в данный момент получатель будет недоступен по причине обрыва связи или сбое в электропитании, то, как только получатель выйдет на связь, данные будут переданы ему повторно. Примерами технологий, поддерживающих распространение данных являются интеграция корпоративных приложений (Enterprise application integration — EAI) и тиражирование корпоративных данных (Enterprise data replication — EDR) [5].

На ранних этапах развития данного подхода связь между источниками и получателями строилась по принципу «каждый с каждым». Но с появлением веб-сервисов и концепции сервисно-ориентированной архитектуры приложений (SOA — Service Oriented Architecture) появилась новая схема распространения данных. Теперь между источником и получателем существует посредник — как правило, это сервисная шина предприятия (Enterprise Service Bus — ESB), которая берет на себя функции гарантированной доставки сообщений (рис.3). Источники данных (распространители) и получатели (подписчики) разрабатываются как сервисы, которые регистрируются в реестре сервисной шины. При фиксации изменений в системе-источнике, распространитель отправляет сообщение сервисной шине, которая, получив сообщение, обращается к реестру для получения списка подписчиков на соответствующие данные. Затем сообщение будет отправлено каждому из подписчиков. Концепция сервисного подхода к управлению данными, включая задачу их интеграции, в ряде источников получила название *Information as a Service* [2].

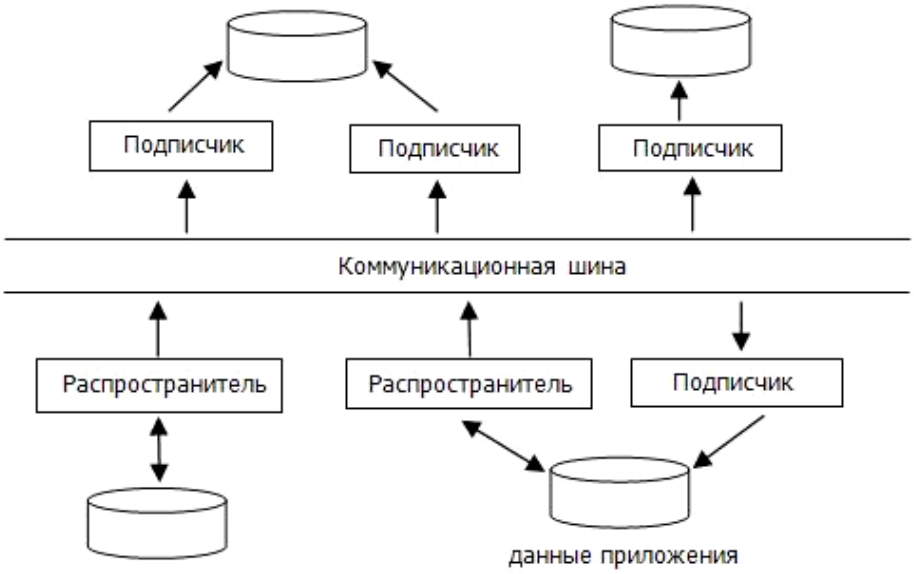

Рис. 3. Схема интеграции на основе распространения данных

Преимуществами такого подхода является высокая надежность и масштабируемость решений, простота в реализации, оперативность обновления

данных (фактически в режиме реального времени). Подключение новых приложений к схеме информационного взаимодействия, как правило, не требует перенастройки уже интегрированных приложений. *К недостаткам* можно отнести высокую сложность интеграции с внешними приложениями, не предоставляющими сервисов для доступа к данным. Для эффективного информационного взаимодействия все ресурсы, входящие в состав схемы, должны быть высокоструктурированными. Некоторые реализации накладывают существенные ограничения на объемы передаваемых данных.

2. Принципы интеграции данных в недропользовании

Интеграция разнородных данных была и остается сложной задачей, непрерывный рост объемов данных и постоянно изменяющиеся требования бизнеса не способствует ее облегчению. Тема очередного исследования института TDWI (The Data Warehousing Institute — институт хранилищ данных) была посвящена проблемам интеграции данных. Более 69 % участников исследования TDWI заявили, что проблемы интеграции данных представляют «очень серьезное» или «серьезное» препятствие для внедрения новых приложений [5]. Согласно мнению респондентов, основными проблемами в области интеграции данных являются: качество данных и вопросы безопасности; недостаток экономических моделей и неадекватное финансирование; неразвитая инфраструктура интеграции данных.

Построению интеграционных решений в основном препятствуют большое разнообразие источников данных, отсутствие открытых стандартов, постоянно изменяющиеся требования. Как правило, для решения конкретных задач предприятие выбирает программные продукты лучшие в своем классе, но со временем, когда появляется необходимость в создании корпоративной информационной среды, выясняется, что эти продукты не совместимы или плохо совместимы между собой. Тем временем уже накоплены большие объемы исторических данных и переход на комплексные решения крупных производителей, таких как IBM, Oracle, гарантирующих совместимость на уровне данных, затруднителен.

Рассмотрим проблемы интеграции на примере комплексной информационно-аналитической системы (ИАС) в сфере недропользования. Целью создания подобных систем является обеспечение информационной поддержки при принятии управленческих решений на этапах поиска, разведки и добычи, направленных на повышение качества использования недр и охране недр от вредоносного воздействия работ, связанного с их использованием. Информация, накапливаемая в такой системе, как правило, выходит за рамки деятельности одной организации и охватывает деятельность нескольких крупных нефтедобывающих компаний, каждая из которых имеет свою корпоративную информационную систему. Естественно, что модели хранения данных в этих системах могут существенно различаться, могут использоваться различные справочники, способы именования скважин, пластов, коды состояний и т.п. Интеграцией данных в таком контексте будет являться не только физическое сосредоточение всей информации в одной базе данных или простое объединение схем хранения. Здесь необходим более сложный подход, включающий в себя очистку и согласование данных из различных источников, в том числе используемых имен, кодов состояний, справочников и др.

Традиционно применяемые информационные системы в сфере недропользования, такие как Tigress, PetroVision, Finder успешно решают задачи вертикальной интеграции, т.е. интеграции данных в пределах одной компании

или группы компаний. В этом случае процесс согласования данных выполняется на этапе ввода, поскольку внутри компании используются единые справочники, классификаторы, регламенты и стандарты. Процесс согласования существенно упрощается и не требует разработки специализированных решений, поэтому их наличие в таких системах не предусматривается. В случае горизонтальной интеграции (интеграции данных различных компаний) применение такого способа согласования невозможно, так как невозможно обязать всех недропользователей использовать одни и те же справочники, а единые утвержденные стандарты учета данных в этой области отсутствуют. Кроме того, этап ручного ввода отсутствует и заменяется пакетной загрузкой данных из различных источников. Поэтому применение перечисленных систем в нашем случае нецелесообразно, поскольку такого рода системы не решают задач согласования при горизонтальной интеграции.

Существующие сегодня решения по интеграции данных в сфере недропользования основываются преимущественно на полном переносе данных «как есть» из учетных систем отдельным блоком. Такой подход исключает процесс согласования и удобен при анализе данных за относительно короткий временной период или анализе только текущих данных. В случае рассматриваемой ИАС временной период может быть значительным, за это время лицензионный участок может неоднократно передаваться различным недропользователям. Это осложняет ситуацию, поскольку на данном лицензионном участке одни и те же объекты разными недропользователями могут быть описаны по-разному. Поэтому в нашем случае данные не могут быть перенесены «как есть» и процесс согласования просто необходим для обеспечения качества и непротиворечивости полученных данных.

Наиболее эффективной платформой для бизнес-аналитики, являются данные, консолидированные в едином хранилище. Рассмотрим преимущества такого решения. Данные поступают в хранилище из корпоративных систем нефтедобывающих компаний, проходят согласование и очистку единожды. Сложные аналитические запросы выполняются на уже подготовленных данных внутри одной СУБД, что исключает передачу больших объемов данных по каналам связи и какие либо дополнительные операции согласования, таким образом, обеспечивается минимальное время выполнения запроса. Информационно-аналитическая система при выполнении запроса остается полностью автономной. Другим важным преимуществом является то, что формат обмена с поставщиками данных может быть любым, включая плоские файлы. Это не требует разработки специализированных компонентов обмена данными от компаний-поставщиков данных (от реализации которых они, естественно, откажутся).

Использование описанного решения предъявляет большие требования к ETL-системе хранилища, которая берет на себя все функции извлечения, преобразования, согласования и загрузки данных. Для рассмотренной ИАС основную сложность представляет именно процесс согласования данных, поскольку охватывает деятельность многих компаний, каждая из которых имеет свою систему учета данных. Например, одно и то же месторождение в разных системах может именоваться как «комсомольское (новое)», «новокомсомольское», «н.комсомольское», вплоть до того, что название может содержать орфографические ошибки. При загрузке поступающих данных необходимо будет учесть, что это разные наименования одного объекта и связать их с уже загруженными данными. Такого рода данные характеризуется высокой степенью взаимосвязанности, и при загрузке в хранилище эти связи нужно максимально

сохранить. Только так мы получим целостную и правильную картину добычи на данном месторождении.

Таким образом, решение проблемы использования разнородных данных, циркулирующих в недропользовании, видится в разработке алгоритмического и программного обеспечения системы интеграции данных, основной задачей которого является *автоматизация процессов согласования и загрузки данных*. Среди концептуальных принципов системы можно выделить следующие.

1. *Обучаемость.* Оператор, контролирующий процесс загрузки данных, получит возможность определять правила поведения и реакцию системы на специфические ситуации, возникающие в процессе загрузки. Со временем система будет способна сама принимать решения в подобных ситуациях.

2. *Гибкость*. Кроме возможности создавать схемы загрузки данных эксперт получит возможность дополнять систему новыми функциями и сценариями обработки данных без внесения изменений в ядро системы или разработки дополнительных модулей. Данную функциональность целесообразно разработать и реализовать на основе скриптового движка Rhino, поддерживающего runtime-компиляцию.

3. *Горизонтальная интеграция*. Система позволит осуществлять загрузку и согласование данных из нескольких внешних систем и сохранять связи между объектами на основе «интеграционных таблиц». Если объект изменится во внешней системе, то после передачи данных изменения зафиксируются и в объекте ИАС.

4. Для обеспечения широких возможностей по согласованию данных, поступающих из внешних систем, требуется *формальный язык согласования данных*.

ЛИТЕРАТУРА

1. *Асадуллаев С.* Архитектуры хранилищ данных. Цикл статей // IBM-2009. [Электрон. ресурс]. Режим доступа: http://www.ibm.com/developerworks/ru/library/sabir/axd_1/index.html.

2. *Дубова Н.* Краткий курс интеграции данных // Открытые системы-2007 [Электрон. ресурс]. Режим доступа: http://www.osp.ru/os/2007/09/4567212/.

3. *Сысоев Т. М.* Интеграция и поиск распределенных данных на основе Semantic Web технологий: Автореф. дис. … канд. техн. наук. 2007 [Электрон. ресурс]. Режим доступа: http://aspirantura.mipt.ru/zastchita/avtoreferats/fupm/f_2xqr2t.

4. *Черняк Л. С.* Интеграция данных: синтаксис и семантика // Открытые системы-2010 [Электрон. ресурс]. Режим доступа: www.osp.ru/os/2010/06/11170978/.

5. *White C.* Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise // DMReview. 2005. №11. P. 25–43.

A.A. Kresov, V.V. Uvarov

PRINCIPLES OF DATA INTEGRATION IN SUBSOIL MANAGEMENT

The article analyzes three basic approaches to integration of heterogeneous information resources: consolidation, federalization, dissemination, suggesting principles underlying development of data integration system in subsoil management.

Subsoil management, data, integration, consolidation, federalization, dissemination, storage, data adjustment, ETL, EII, SOA.