XXXII Неделя науки СПбГПУ. Материалы межвузовской научно-технической конференции. Ч.V : С.29-30

УДК 004.658.6

А.В.Иванов (4 курс, каф ИУС), А.А.Зотов (асп., каф ИУС),
С.А.Пархоменко (главный системный аналитик, ООО “Деловые консультации, СПб”)

РАЗРАБОТКА НАБОРА COM-ОБЪЕКТОВ ДЛЯ MS DTS, РЕАЛИЗУЮЩЕГО ЗАГРУЗКУ ИНФОРМАЦИИ ИЗ ТЕКСТОВЫХ ФАЙЛОВ В ТАБЛИЦЫ МЕТАДАННЫХ

При разработке аналитической системы одной из наиболее важных задач является обеспечение взаимодействия с внешними источниками. Внешние источники могут иметь разную природу и характер. Ими могут быть базы данных различных форматов, документы, глобальные сети и др. Особый интерес представляют внешние источники в виде текстовых файлов. Чаще всего загрузка данных от источников такого типа носит слабо формализуемый и, следовательно, неавтоматизированный характер.

Эта особенность текстовых файлов ведет к тому, что все действия по извлечению информации из файла приходится делать вручную, что ведет к большим затратам времени на однообразную работу, т.е. созданию таблиц и перенос данных вручную из файла отчета в таблицу. Поэтому возникает необходимость разработки способа автоматизации переноса данных из отчета в таблицы баз данных. Эта задача не является тривиальной, так как формат представления информации в текстовых файлах может иметь сложную структуру, а различные файлы часто имеют существенные отличия в структурах организации данных.

Сложность данной задачи можно увидеть на примере Microsoft® SQL Server™ 2000, в котором предусмотрена возможность разбора текстовых файлов. С его помощью можно работать только с простыми текстовыми файлами, а файлы с чуть более сложной структурой организации данных разобрать правильно практически невозможно. Это обусловлено тем, что для разбора сложных по структуре текстовых файлов необходимы средства описания форматов, которые в описанном средстве отсутствуют.

Под форматом текстового файла понимается некоторое описание структуры организации данных этого файла, которое описывает расположение тех данных, которые нужно загрузить из этого файла в таблицу базы данных.

Фирмам, работающим с базами данных, чаще всего приходится иметь дело не просто с текстовыми файлами, а с файлами, которые содержат отчеты. Поэтому при разработке описания формата файла учитывались особенности отчетов (например: наличие headera и footera).

После рассмотрения нескольких видов отчетов было выдвинуто предположение что, имея текст и некоторое описание формата отчета, можно автоматически получать данные в виде таблицы с фиксированным количеством столбцов. Описание формата отчета должно быть формализовано, поскольку выборка данных из отчета осуществляется автоматически. Следовательно, требуется разработать формат описаний, который должен позволять описание структуры сложных отчетов, и, в то же время, быть достаточно простым, поскольку предполагается, что описание отчетов будет создавать администратор системы, а не разработчик формата. В описание следует включить набор правил выделения данных и признак типа отчета, чтобы исключить интерпретацию неверного типа документа.

Для создания описаний формата отчета был выбран язык разметки XML. Выбор обусловлен тем что язык XML имеет жесткую структуру, что облегчает разработку и последующую работу с программой. На этом языке был разработан формат описаний, который позволяет описать практически любой отчет.

Используя этот формат описаний была разработана программа в виде задачи (task) для Microsoft® SQL Server™ 2000 Data Transformation Services.

Задача является основной функциональной единицей DTS Package (пакет), позволяя как трансформировать данные, так и выполнять многие другие операции. Стандартный пакет задач Microsoft® SQL Server™ 2000 Data Transformation Services уже включает некоторые типовые задачи. Для расширения возможностей DTS предусмотрена возможность написания и подключения собственных задач, реализующих то или иное действие. Задача, подключаемая к DTS, должна быть оформлена в виде COM-объекта. Она должна быть поддерживать два уникальных для DTS интерфейса: CustomTask интерфейс и CustomTaskUI интерфейс, если пользователь предполагает включить в задачу какой-то свой пользовательский интерфейс.

В интерфейсной части задаются исходный файл отчета, XML файл описания, имя сервера, базы данных, название таблицы, куда будут помещены данные из отчета. По комаде execute пакет, в котором содержится задача, запускается на выполнение. По данным из интерфейсной части создается соединение с сервером и в заданной базе данных создается пустая таблица, после чего производится загрузка информации из отчета.

После того, как данная задача была реализована, время загрузки информации из текстовых файлов содержащих отчеты уменьшилось многократно, т.к. время написания XML описания 2–10 минут, а время разбора файла вручную напрямую зависит от сложности его структуры и размера, а так как обычно размеры отчетов достаточно велики, то и время их разбора также велико. Еще одним плюсом реализованной системы, позволяющим сэкономить время на загрузку данных, является то, что при поступлении нового отчета с форматом, для которого уже когда-то было написано XML-описание, его не нужно писать заново, а достаточно загрузить файл, в котором оно содержится.