

ПРИМЕНЕНИЕ IBM INFOSPHERE DATASTAGE В СОЗДАНИИ ПРОЦЕССА ОБРАБОТКИ И ХРАНЕНИЯ КОРПОРАТИВНОЙ ИНФОРМАЦИИ

Гусаковская Евгения Геннадьевна

магистрант, Белорусский государственный университет информатики и радиоэлектроники, РБ, г. Минск

Аннотация. В данной статье изучены возможности программного продукта IBM InfoSphere DataStage, а также его применимость для создания процессов извлечения, преобразования и загрузки данных.

Введение

В настоящее время существует проблема наличия множества баз данных, содержащих неоднородную информацию. Из-за этого возникают проблемы при попытках получения целевой информационной картины о предметной области, описанной в данных источниках. И, как следствие, появляются сложности при создании отчетности. Одна из главных проблем – довольно долгое формирование отчета, из-за использования внешних источников данных. В настоящее же время руководители должны иметь возможность получить доступ к нужной им информации сразу же, как только она им понадобится. Следовательно, предприятия нуждаются в системах, способных осуществлять интеграцию и верификацию данных между различными системами. Наличие данного типа систем позволяет упростить осуществление контроля над данными и получать пользователям наиболее актуальную информацию о предметной области.

Одним из вариантов реализации данных систем является стек технологий DataWarehouse. Данный стек включает в себя средства и методы извлечения, загрузки и преобразования информации, так называемый ETL процесс. В данном докладе будет рассмотрено использование программного продукта IBM InfoSphere DataStage как основного инструмента для создания ETL процессов.

1. Понятие ETL

ETL (от англ. Extract, Transform, Load – дословно «извлечение, преобразование, загрузка») -- комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных.

Целью ETL приложения является извлечение информации из одного или нескольких источников, преобразование ее в формат, поддерживаемый системой хранения и обработки, которая является получателем данных, а затем загрузка в нее преобразованной информации [1, с. 2554]. Изначально ETL-системы использовались для переноса информации из более ранних версий различных информационных систем в более новые. В настоящее время ETL-системы все более широко применяются именно для консолидации данных с целью их дальнейшего анализа. Очевидно, что поскольку хранилища данных могут строиться на основе различных моделей данных (многомерных, реляционных, гибридных), то и процесс ETL должен разрабатываться с учетом всех особенностей используемой в модели хранилища данных. Кроме того, желательно, чтобы ETL-система была универсальной, то есть могла извлекать и переносить данные как можно большего числа типов и форматов.

Независимо от особенностей построения и функционирования ETL-системы, она должна обеспечивать выполнение трех основных этапов процесса переноса данных [2, с. 28]:

- Извлечение данных. На этом этапе данные извлекаются из одного или нескольких источников и подготавливаются к преобразованию. Следует отметить, что для корректного представления данных после их загрузки в хранилище данных из источников должны извлекаться не только сами данные, но и информация, описывающая их структуру, из которой будут сформированы метаданные для хранилища;
- Преобразование данных. Производится преобразование форматов и кодировки данных, а также их обобщение и очистка;
- Загрузка данных – запись преобразованных данных в соответствующую систему хранения.

Следует помнить, что ETL – это не только процесс переноса данных из одного приложения в другое, но и инструмент их подготовки к анализу.

2. IBM InfoSphere DataStage

IBM InfoSphere DataStage – это инструмент ETL, входящий в пакет решений IBM Information Platforms и IBM InfoSphere. Программный продукт использует графическую нотацию для построения решений по интеграции данных и доступен в различных версиях, таких как Server Edition, Enterprise Edition и MVS Edition. Он использует архитектуру клиент-сервер. Серверы могут быть развернуты как в Unix, так и в Windows. DataStage является мощным инструментом интеграции данных, часто используемый в проектах хранилищ данных для подготовки данных к генерации отчетов.

IBM DataStage состоит из следующих программных модулей [3, с. 129]:

- DataStage Server. Мощный многопоточный сервер позволяет преобразовывать большие и очень большие объемы данных. При этом данные, получаемые из источников и загружаемые в приемники, могут быть представлены в самых разнообразных форматах (текст, XML, JSON, сообщения WebSphere MQ и SeeBeyond, ADABAS и др.). Наличие нескольких ЦПУ дает возможность разбивать ETL-процедуры и их этапы (job и stage -- в терминологии IBM WebSphere DataStage) на подзадания и обрабатывать их параллельно.
- DataStage Designer. Данное средство разработки позволяет создавать ETL-процедуры и описывать преобразования данных в графическом виде. Программист создает потоки преобразования информации, перемещая элементы, называемые stage, которые ответственны за выполнение какой-то конкретной задачи: выгрузка/загрузка данных, преобразование данных, агрегирование и т.д.
- DataStage Manager. Отвечает за организацию базовых элементов любого DataStage проекта в каталогах. Данное программное средство позволяет разделять общие метаданные между несколькими серверами.
- DataStage Administrator. Средство настройки и управления сервером. Его основные функции: определение прав доступа разработчиков к различным участкам проекта, настройка производительности сервера и использования административных функций.

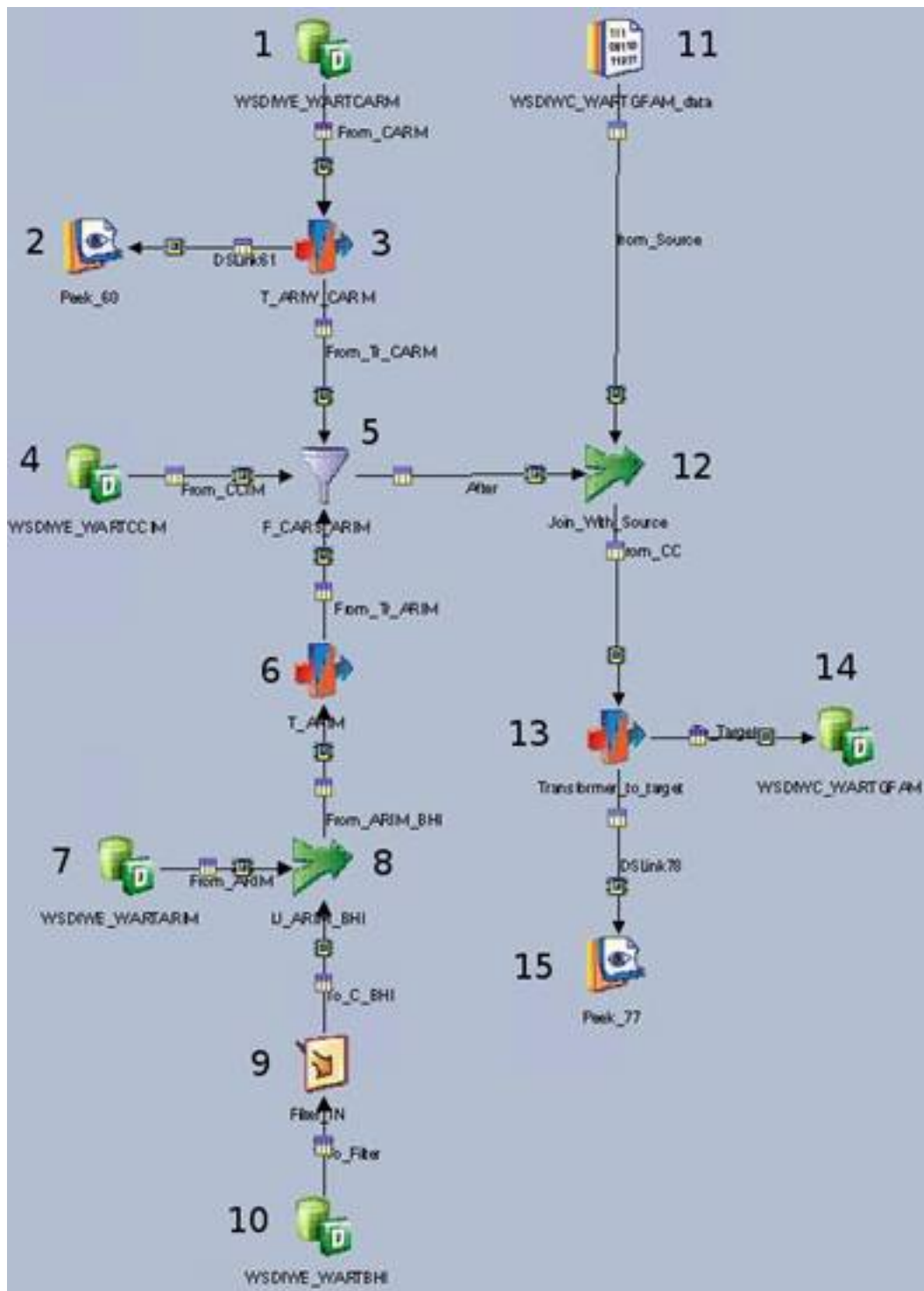


Рисунок 1. DataStage job

Рассмотрим несколько основных элементов разработки процедур [4, с. 135] в DataStage Desing на примере job-а, представленного на рисунке 1. Для удобства обозначения элементы, называемые stage, пронумерованы.

1. Db2 Connector stages – элементы, предназначенные для соединения с базой данных Dd2 для выгрузки (stages №№ 1, 4, 7, 10) или загрузки данных (stage №14).
2. Sequential file stage (элемент №11) позволяет считывать/записывать данные из файла или в файл.
3. Transformer stages – это элементы (stages №№ 3, 6, 13), служащие для направления потоков данных и преобразования самих данных.
4. Join stages (элементы №№ 8, 12), Filter stage (элемент №9) и Funnel stage (элемент №5)

являются своеобразными аналогами SQL команд join, where и union соответственно.
5. Peek stages (элементы №№ 2, 15) позволяют выполнять отладку

Заключение

Как мы можем видеть, IBM InfoSphere DataStage является мощным инструментом для создания ETL процессов. Данный программный продукт позволяет производить сложные преобразования на больших объемах данных. Графический интерфейс упрощает разработку процедур, позволяя создавать потоки преобразования информации, не прибегая непосредственно к программированию. А использование параллельных вычислений уменьшает время выполнения преобразований и загрузки данных.

Список литературы:

1. Li Jian, Xu Bihua. ETL tool research and implementation based on drilling data warehouse. // Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference, vol.6.
2. Ralph Kimball, Joe Caserta. The data warehouse ETL toolkit : practical techniques for extracting, cleaning, conforming, and delivering data. // Wiley Publishing, Inc., 2004
3. IBM Infosphere DataStage Essentials v11.5 (Course Guide KM204) // IBM 2015.
4. IBM InfoSphere DataStage Data Flow and Job Design // International Business Machines Corporation 2008.