УДК 004.4

Извлечение данных – важная часть процесса ETL

Кузьмина Юлия Васильевна
Брянский государственный университет имени академика И.Г. Петровского
магистрант

Кубанских Олеся Владимировна
Брянский государственный университет имени академика И.Г. Петровского
кандидат физико-математических наук, ст. преподаватель кафедры информатики и прикладной математики

Аннотация
В статье описывается процесс извлечения данных из различных источников в единое хранилище данных.
Ключевые слова: извлечение, ETL, хранилище данных, логические карты данных, диаграмма сущность-связь.

Extraction is important part in ETL process

Kuzmina Yuliya Vasilevna
Bryansk State University named after academician I.G. Petrovsky
undergraduate

Kubanskikh Olesya Vladimirovna
Bryansk State University named after academician I.G. Petrovsky
candidate of physical and mathematical sciences, high teacher of department of informatics and applied mathematics

Abstract
Process of extraction of data from various sources in warehouse is described in article.
Keywords: extraction, ETL process, data map, SCD, E-R diagram.

ETL состоит из трех фаз: извлечение, преобразование, загрузка. Полная схема процесса ETL показана на рис. 1.

Рисунок 1. Полная схема процесса ETL

Первая фаза – извлечение – является важной составляющей, фундаментом для других фаз цикла ETL. В статье процесс извлечения описан подробнее.

Первый шаг при проектировании системы извлечения – создание логических карт данных (англ. data map). Карта данных обычно представляет собой электронные таблицы, которые содержат сопоставление форматов данных первоначального источника и конечного пункта нахождения данных. Иногда карты данных содержат также отчет о происхождении данных.

Линейный отчет может содержать следующие данные:

1) Компоненты целевой таблицы: имя таблицы, имя столбца, тип таблицы (таблицы фактов или измерений), тип данных.

2) Компоненты исходной таблицы, такие как исходная база данных, имя таблицы, имя столбца, тип данных.

3) Тип преобразования.

1) Четыре компонента целевой таблицы (имя таблицы, имя столбца, тип таблицы, тип данных) обязательны. Пятый, дополнительный компонент это - SCD, измерения, которые изменяются со временем. SCD – это Slowly Changing Dimension, что можно перевести как медленно изменяющиеся измерения.

Типы SCD:

• Тип 1 – измененный атрибут просто обновляется (перезаписывается), чтобы отразить последнее значение, при этом история не сохраняется.

• Тип 2 – при изменении каких-то данных добавляется новая строка, старая не удаляется. Обе строки (и старая и новая) содержат в качестве атрибута флаг актуальности.

• Тип 3 – при изменении данных к существующей строке добавляется новый атрибут.

2) Исходный компонент (Source Component): исходная база данных, откуда данные извлечены, определяется как начальный пункт цикла ETL.

3) Преобразование: преобразование данных, приведение к необходимому формату, может быть выражено в SQL или псевдокоде.

Как построить логические карты данных?

Прежде, чем начать проектировать логические карты данных, мы должны знать, как выглядят итоговые данные.

• Определение источника данных.

• Сбор и документирование исходной системы.

• Создание системы, которая отслеживает отчеты, показывающие информацию о том, кто ответственный за каждый источник. Может содержать такие особенности источника, как имя интерфейса, транзакции за день, приоритеты, количество ежедневного использования, использование отделов, деловое использование, платформы и некоторые комментарии.

• Создание и анализ E-R диаграммы (отношения предприятия «сущность» - «связь»). Она показывает, как различные объекты предприятия связаны друг с другом.

E-R диаграмма содержит уникальный идентификатор: status_id, status_code, status_description.

• Типы данных: отношения между типами это - очень важная особенность отношений и играет большую роль в последующем извлечении данных.

• Дискретные отношения: необходимы для отображения многих измерений. Часто это единственная центральная таблица (или представление), в которой хранятся все объекты, таким образом доступ к данным становится более легким.

• Количество элементов отношений и колонок. Есть три типа количества элементов:

- один к одному - отношение первичного ключа между таблицами;

- один ко многим – отношения внешнего ключа в таблицах;

- многие ко многим – например, три таблицы связаны с двумя таблицами отношениями типа «один ко многим».

После подготовки логической карты данных необходимо:

• Проанализировать содержание данных, которое ставит нас перед выбором данных для процедуры извлечения.

• Бизнес-правила должны быть полными для процесса ETL.

• Объединение источников данных к единственному источнику данных может привести к эффективной работе. Но если измерения совершенно отличаются, тогда объединение будет неэффективным.

• Логическая карта данных - это только спецификация для физического плана работы. Она показывает основные моменты для внимания во входной и выходной базе данных. Это иногда дает эффективный результат. Может также использоваться в качестве источника информации и представляться конечным пользователям.

Для того чтобы сделать цикл ETL полным и извлекать данные из различных баз данных, нужно приложить много усилий. Создание логической карты данных играет важную роль для процесса извлечения. Поскольку процесс извлечения происходит для многих баз данных с различными форматами, может потребоваться намного больше концентрации на процессе извлечения, таким образом, логическое отображение данных эффективно от источника до места назначения. Точное извлечение данных из базы данных - основное, необходимое условие для остальных фаз в цикле ETL.

Библиографический список

1. Patel S., Pandia M. How is Extraction important in ETL process? / International Journal of Advance Innovations, Thoughts & Ideas / IJAITI 2012
2. Kimball R., Caserta J. The Data Warehouse ETL Toolkit. Wiley Publishing, inc, 2004
3. Ross M. Slowly Changing Dimensions Are Not Always as Easy as 1, 2, 3; / Kimball Group (March 10, 2005). Intelligent Enterprise.