Ходжаев Р.Д.

ФОРМИРОВАНИЕ ETL ЗАДАЧ ДЛЯ ПОЛНОЙ ЗАГРУЗКИ ДАННЫХ

R.D. Khodzhaev

ETL TAKS FORMATION FOR A FULL DATA LOAD

УДК: 638/57-35

В статье рассматривается формирование ETL задач для полной загрузки информации из логируемых источников в хранилища данных банков.

Ключевые слова: *хранилище данных, банк, ETL задача, полная загрузка данных, логируемый источник.*

The article discusses the formation of ETL tasks for full load of information from logged sources to data warehouses of banks.

Key words: *data warehouse, Bank, ETL task, full loading of data, логируемый source.*

Банки с большим объемом обрабатываемой информации используют хранилище данных для систем поддержки принятия решений (СППР), в отличии от типовых Автоматизированных Банковских Систем (АБС), предназначенных для транзакционной обработки данных. Структура ХД отличается от структуры базы данных АБС (операционного источника). ХД позволяет сотрудникам банка оперативно формировать отчеты для принятия бизнес-решений, несмотря на загруженность операционного источника. Для того чтобы в ХД были актуальные данные – хранилище обновляется периодически, то есть данные с помощью ETL задач загружаются из операционного источника в ХД. ETL задача представляет собой набор правил, которые ставят в соответствие структуры данных источника структурам ХД, кроме того эти правила применяются для проверки, очистки и дополнения данных. В соответствии с техникой формирования ETL задачи подразделяются на две категории: для полной и инкрементальной загрузки. Ниже предлагается формирование типовых ETL задач для полной загрузки данных в хранилища. Таблицы операционного источника, таблица ETL области, внешние таблицы, являются источником данных для полной загрузки (ПЗ). Для инкрементальной загрузки – источником данных являются представления stage области, снимки таблиц источника, таблицы ETL области, внешние таблицы. В качестве инструмента для формирования ETL задач предлагается использовать Oracle Warehouse Builder версия 11.2 (OWB). Правила трансформации и загрузки данных в целевые таблицы настраиваются при помощи операторов OWB.

Редактор общих настроек ETL задачи можно вызвать в рабочей области «Дерево метаданных» OWB (Рис.1).

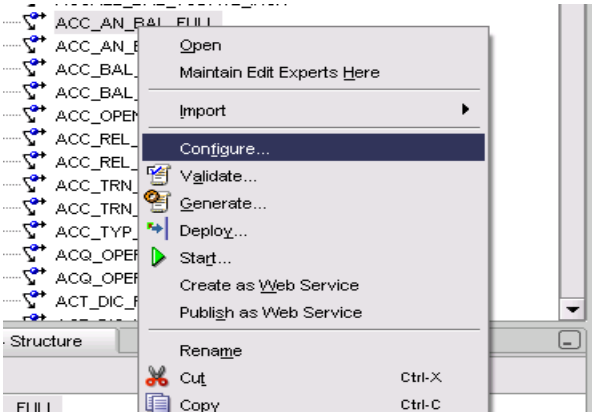

Рис. 1. Дерево метаданных OWB.

Необходимо настроить следующие параметры (основные выделены жирным шрифтом) (Рис.2). Наиболее важными являются – *Use Target Load Ordering*, *Generation Mode*, *Default Operating Mode*.

В зависимости от логики ETL задачи значения указанных параметров могут меняться, но, как правило, это:

Use Target Load Ordering = true – использование той очерёдности загрузки, которая указана в настройках ETL задачи;

Generation Mode = Set Based – генерация ETL задачи для пакетного режима;

Default Operating Mode = Set Based – выполнение загрузки в пакетном режиме.

Deployable	**true**
Generation Comments	
Language	**PL/SQL**
Referred Calendar	
⊞ Chunking options	
⊟ Code generation options	
Analyze table statements	**true**
ANSI SQL Syntax	**false**
AUTHID option	None
Bulk processing code	true
Commit Control	Automatic
Enable Parallel DML	false
Error Trigger	
Generation Mode	**Set based**
Optimized code	true
Use Target Load Ordering	**true**
⊞ Post Map Process Operators	
⊞ Pre Map Operators	
⊟ Runtime parameters	
Analyze table sample percentage	**40**
Bulk size	1000
Commit frequency	1000
Default audit level	Error Details
Default Operating Mode	**Set based**
Default purge group	WB
Maximum number of errors	**0**

Рис. 2. Настройки ETL задачи.

Рассматриваются следующие типовые полные загрузки:

- Загрузка стандартной сущности.
- Загрузка древовидной (иерархической) сущности.
- Загрузка снимков.

Ниже описывается реализация каждой типовой ПЗ с использованием следующих обозначений:

A (F_{sp}, F_s) – таблица источника, где F_{sp}– первичный ключ, F_s –поля, не входящие в первичный ключ

T_{etl} (F_{dwhp}, F_{sp}, F_a) – таблица ETL области для полного набора связей ключей источника и ХД, где F_{dwhp} – первичный ключ таблицы ХД, F_{sp} – первичный ключ таблицы источника, F_a – дополнительные поля.

T_{dwh}(F_{dwhp}, F_{dwh}) – таблица ХД, где F_{dwhp} – первичный ключ, F_{dwh} – поля, не входящие в первичный ключ, S_T_{dwh} – последовательность, реализующая суррогатный ключ F_{dwhp}.

Стандартная сущность – это таблица источника, соответствующая требованиям:

- древовидная (иерархическая)
- неизменяемый первичный ключ
- существует прямой аналог в ХД

Для реализации загрузки стандартной сущности необходимо реализовать следующие DML конструкции в редакторе OWB с помощью операторов «Панели инструментов»:

1. заполнять **T_{etl}**:

```
INSERT INTO Tetl(WH_ID, Fsp)
SELECT S_Tdwh.NEXTVAL, Fsp FROM A
```

2. заполнять **T_{dwh}**:

```
INSERT INTO Tdwh (Fdwhp, Fdwh)
SELECT Tetl.WH_ID, A.Fs
FROM Tetl, A
WHERE Tetl.Fsp = A.Fsp
```

3. устанавливать очерёдность загрузки **T_{etl} =>T_{dwh}**

Иерархическая сущность – это таблица, у которой установлен порядок подчинённости низших звеньев высшим. Структура такой сущности представляет собой дерево без циклов. Каждая запись содержит идентификатор родителя. Если родитель не указан – запись корневая. Наименование поля, содержащего идентификатор родителя – *ID_HI* и на *ID_HI* накладывается ограничение целостности foreign key:

```
ALTER TABLE A ADD (CONSTRAINT FK _ TS
__ TREE
FOREIGN KEY (ID_HI) REFERENCES A (Fsp))
```

При загрузке иерархических сущностей необходимо дополнять **T_{etl}** полем

SRC_ID_HI – идентификатор родителя в таблице источника, т.е. **T_{etl}(F_{dwhp}, F_{sp}, F_a)**, где **F_a=SRC _ID_HI**.

Для этого в редакторе OWB с помощью операторов «Панели инструментов» необходимо реализовать следующие DML конструкции:

1. заполнять **T_{etl}**:

```
INSERT INTO Tetl (WH_ID, Fsp, SRC_ID_HI)
SELECT S_Tdwh.NEXTVAL, Fsp, ID_HI FROM A
```

2. заполнять **T_{dwh}**:

```
INSERT INTO Tdwh (Fdwhp, ID_HI, Fdwh)
SELECT Tetl.Fdwhp, TREE.Fdwhp, Fs
FROM A, Tetl, Tetl TREE
WHERE Tetl.Fsp = A.Fsp
AND Tetl.SRC_ID_HI = TREE.SRC_ID (+)
```

3. устанавливать очерёдность загрузки **T_{etl} =>T_{dwh}**

Снимок – это копия данных таблиц источника.

Для загрузки снимка необходимо в редакторе OWB с помощью операторов «Панели инструментов» реализовать следующую DML-конструкцию:

1. заполнять снимок **T_{etl}**

```
INSERT INTO Tetl (Fsp, Fs)
SELECT Fsp, Fs FROM A
```

В OWB Designer указываются различные режимы генерации и выполнения ETL задач. Основные два режима – построчный (Row Based) и пакетный (Set Based). Генерация – это создание PL\SQL пакета

(package) на базе формированной ETL задачи с учётом установленных параметров. Режимом генерации управляет параметр Generation Mode. Сгенерировать ETL задачу можно как в пакетном, так и в построчном режиме, а также в обоих одновременно. Как она будет выполняться зависит от параметра Default Operating Mode.

Пакетный режим обеспечивает высокую производительность, но предоставляет минимум логирующей информации для разработчика ХД. Данный режим подходит для штатной работы загрузки данных.

Построчный режим сопровождается низкой производительностью загрузки, но при этом лог выполнения загрузки отражает детальную информацию о происходящих событиях во время работы ETL задачи. Этот режим подходит для отладки (Debug) процесса загрузки.

В пакетном режиме реализованные DML конструкции загружают данные одним блоком, например:

INSERT INTO ...
SELECT ... FROM
MERGE INTO … USING (SELECT…
DELETEFROM … WHERE

Тем самым достигается высокая производительность.

Во втором режиме реализованные DML конструкции загружают данные построчно. Логика выполнения сгенерированного пакета следующая: набор данных заполняет коллекцию (FETCH < CURSOR > BULK COLLECT INTO), а затем для каждой записи коллекции выполняется DML операция в определённые целевые таблицы (FORALL i IN ... INSERT INTO), при этом объём логирующей информации – максимальный.

Предлагаемый принцип формирования ETL задач позволит банкам в дальнейшем разработать систему полной загрузки данных в ХД из логируемых источников.

Литература:

1. Фейерштейн С., Прибыл Б. Oracle PL\SQL для профессионалов 3-е изд. СПб.: Питер, 2003. – 941 с.: ил. ISBN 5-318-00528-4
2. Griesemer B. Oracle Warehouse Builder 11g: Getting Started. Packt Publishing Ltd, 2009. – 368 p. ISBN 1847195741
3. МакДональд К., Кац Х., Кристофер Б., Кальман Дж, Нокс Д. Oracle PL/SQL для профессионалов: практические решения. СПб.: ДиаСофтЮП, 2005. – 560 с.: ISBN 5-93772-160-8 .

Рецензент: к.т.н., доцент Советбеков Б.С.