*Ходжаев Р.Д.*

# АЛГОРИТМЫ РЕАЛИЗАЦИИ МЕТОДОВ ПРЕДОТВРАЩЕНИЯ ОШИБОК ИНКРЕМЕНТАЛЬНОЙ ЗАГРУЗКИ В ХРАНИЛИЩЕ ДАННЫХ

*R.D. Khodzhaev*

# ALGORITHMS OF IMPLEMENTATION OF METHODS FOR PREVENTING ERRORS IN INCREMENTAL LOADING OF DATA WAREHOUSES

*УДК:627.315 (043)*

*В статье рассматривается разработка алгоритмов реализации методов предотвращения ошибок в хранилище при инкрементальной загрузки для трех типов операционных источников банков Республики Таджикистан.*

***Ключевые слова:*** *банк, хранилище данных, алгоритмы, предотвращения ошибок.*

*The article describes the development of algorithms for the implementation of methods for preventing errors in the warehouses of the incremental load for three types of operational sources of banks of the Republic of Tajikistan.*

***Key words:*** *bank, store data, algorithms, error prevention.*

При проектировании инкрементальной загрузки данных следует учитывать следующие основные факторы:

1. тип базы данных(БД) источника;
2. механизм захвата измененных данных;
3. структура и назначения stage-области;
4. ETL - задачи, технологии и инструменты их реализации;
5. структура БД Хранилища Данных (ХД).

В данной статье во всех рассматриваемых случаях принимается, что БД ХД имеет многомерную структуру - схему star или snowfiake. Учет структуры БД ХД является сложной задачей и выходит за рамки данной статьи. Ниже рассматриваются задачи и назначение основных механизмов инкрементальной загрузки данных в ХД: CDC (Change Data Capture), stage области, ETL (Extract, Transform, Load).

**Задачи и назначение механизма захвата измененных данных**

Механизм захвата измененных данных (CDC - Change Data Capture) должен обеспечивать выполнения следующих процессов:

Доступ к источнику: непосредственно к данным (таблицам, представлениям и т.п.) или опосредовано через активные или архивные лог файлы.

Выделение необходимых для загрузки в ХД данных.

Пересылка данных в stage область.

Запись данных об изменениях в таблицы stage области.

**Задачи и назначение stage области**

Основным назначением stage области является накапливание изменений и предоставление для инкрементальных загрузок только новых изменений. Измененные данные помещаются в таблицы stage области. Типовая структура таблиц stage области следующая:

$A(F_S)$ - таблица - источник, где $F_s$ - поля A. Тогда наименование и структура представления будет следующая:

$A_{stg}(F_s, F_t)$, где $F_s$ - поля, совпадающие по наименованию и типу с полями таблицы источника. Их может быть меньше чем в соответствующей таблице источника A, так как не все типы поддерживаются в процессе передачи данных между источником и stage областью. $F_1$ - набор технических полей.

**Задачи и назначение ETL**

Основные функции ETL - это извлечение данных из stage области и источника, их преобразование и дальнейшая загрузка обработанных данных в таблицы ХД. В примере ниже рассматривается задача ETL.

INSERT INTO $A_{dwh}$ ($F_{dwhi}$, $F_{dwh2}$), …… $F_{dwhn}$)

SELECT TRUNC($A.F_{al}$) as $F_l$, ($B.F_{b1}$+$B.F_{b2}$) as $F_2$ …, Fn

FROM A, B, $A_{stg}$, $B_{stg}$

WHERE...,

где $A_{dwh}$ - таблица в ХД, A и B - таблицы источника, $A_{stg}$ и $B_{stg}$ - таблицы stage области.

В данном примере:

Функция Extract (извлечение данных) ETL задачи выполняется в шаге запроса SELECT FROM A,B,$A_{stg}$,$B_{stg}$ Функция Transform (преобразование данных) в секции: TRUNC($A.F_{al}$), ($B.F_{b1}$+$B.F_{b2}$). Функция Load - (операция загрузки) с помощью INSERT INTO в таблицу $A_{dwh}$.

После рассмотрения основных компонентов инкрементальной загрузки данных - CDC, stage области и ETL задач, предлагается реализация алгоритмов методов по предотвращению причин возникновения ошибок для промежуточного, логируемого источников, а также для источников с метками времени.

**Алгоритм для промежуточного источника**

Алгоритм реализации метода предотвращения ошибок в ХД при инкрементальной загрузке данных из промежуточного источника состоит их следующих этапов:

1. Создание stage области.
2. Создание механизма формирования новых\ старых снимков.
3. Формирование ETL задач на основании новых\ старых снимков

Во время реализации алгоритма необходимо учитывать тип механизма CDC, используемый промежуточным источником. Для промежуточного источника предлагается механизм CDC, который должен будет захватывать измененные данные следующим образом:

Промежуточному источнику необходимо иметь снимки источника - новые и старые, чтобы получить набор измененных данных. Эти снимки будут находиться в stage области.

Например, в источнике есть таблица A ($F_1$, $F_2$); контейнер (таблица) для старого снимка и для нового в stage области: $A_{stg(old)}$ ($F_1$, $F_2$) и $A_{stg(new)}$ ($F_1$, $F_2$). Для данного метода CDC должен обеспечивать:

1. Помечать снимок $A_{stg(new)}$ ($F_1$, $F_2$)как старый и передавать данные из него в Astg(0id) (Fj, F2). Данные $A_{stg(oid)}$ ($F_n$ $F_2$) уже не актуальны и поэтому можно удалить содержимое перед обновлением. Данное действие возможно реализовать через SQL операторы:

TRUNCATE $A_{stg(old)}$;
/
INSERT INTO $A_{stg(old)}$
SELECT FROM $A_{stg(ntw)}$;
/
TRUNCATE $A_{stg(new)}$;
/

2. Извлекать данные из таблицы источника A в $A_{stg(new)}$ stage области.

Данное действие можно реализовать с помощью операторов SQL:

INSERT INTO $A_{stg(new)}$
SELECT FROM A@"DBLINK_TO_SOURCE"; где DBLINK_TO_SOURCE - связь между БД stage области и БД источника.

Далее вычисляется дифференциал через ETL задачи, работающие следующим образом: Данные для удаления:

SELECT FROM $A_{stg(old)}$ MINUS
SELECT FROM $A_{stg(new)}$ Данные для добавления: SELECT FROM $A_{stg(new)}$ MINUS
SELECT FROM $A_{stg(old)}$

Для промежуточного источника алгоритм можно также реализовать при помощи объектов файловой системы:

В stage области необходимо иметь таблицу Astg(diff) и файлы в файловой системе операционной системы - $A_{stg(new)}$.txt и $A_{stg(old)}$.txt. То есть, основная задача CDC - это заполнение снимка Astg(new).txt в stage области. На операционном источнике данные таблицы A экспортируются в объект файловой системы (например A.txt), затем этот файл пересылается в stage область. Данные полученного файла A.txt копируются в Astg(new).txt, но перед этим данные из As(g(new)-txt перемещаются в Astg(old).txt. После этого вычисляется дифференциал средствами ETL задач, и этот дифференциал загружается в таблицу Astg(diff) для последующей загрузки в ХД.

**Алгоритм для логируемого источника**

Алгоритм реализации метода предотвращения ошибок в ХД при инкрементальной загрузке данных из логируемого источника состоит их следующих этапов:

1. Создание stage области.
2. Создание механизма фиксации точки отсчёта.
3. Формирование ETL задач на основании таблиц stage области.

При реализации алгоритма необходимо учитывать тип механизма CDC, используемый логируемым источником. Для логируемого источника CDC должен работать следующим образом:

Необходимо передавать архивные лог-файлы СУБД (Система управления базами данных) из источника в stage область. Далее происходит чтение лог-файлов и извлечение необходимых данных, которые вставляются в таблицы stage области. Описанные действия CDC можно реализовать с помощью продукта Oracle Streams [1].

В следующем примере рассматривается структура stage области, которая необходима для реализации алгоритма

Пусть A(F1, F2) - таблица источника, где F1 и F2 - поля этой таблицы. В этом случае необходимо в stage области иметь таблицу «приемник» со следующей структурой Astg (F1, F2, Ft), где Fb F2 _ такие же поля как в таблице A источника, Ft - набор технических полей, таких как физическое время транзакции, номер транзакции, тип выполняемой DML операции (insert 1 update 1 delete).

Данные в таблице $A_{stg}$ будут храниться в следующем виде:

| | | | | | |
|---|---|---|---|---|---|
| F11 | F21, | 'INSERT' | 03.04.2011 | 01:01:01 | 123456 |
| F12 | F22 | 'UPDATE' | 03.04.2011 | 01:01:02 | 123457 |
| F13 | F23 | 'INSERT' | 03.04.2011 | 01:01:06 | 123458 |
| F14 | F24 | 'DELETE' | 03.04.2011 | 01:01:07 | 123459 |

При реализации данного алгоритма перед запуском очередной инкрементальной загрузки необходимо блокировать прием архивных лог файлов на время загрузки. Таким способом фиксируется точка отсчета и тем самым устраняется причина возникновения ошибок во время инкрементальной загрузки.

Причина несоответствия между состоянием базовых таблиц и данными в логах изменений для легируемого источника устраняется при помощи stage области. Для ETL задач источником данных являются данные stage области, которые заполняет CDC Oracle Steams.

**Алгоритм для источника с метками времени**

Алгоритм реализации метода предотвращения ошибок в ХД при инкрементальной загрузке данных из источника с метками времени состоит их следующих этапов:

Создание stage области.

Создание механизма фиксации точки отсчёта.

Формирование ETL задач на основании таблиц stage области.

При реализации алгоритма необходимо учитывать тип механизма CDC, используемый источником с метками времени. Для источника с метками времени механизм CDC должен работать следующим образом:

В каждой таблице в источнике создается поле аудита, например change Jimestamp, которое отражает время последней модификации записи любого из полей. Рассматриваемая в качестве примера таблица A(FjF2) меняет структуру на A(FbF2, change timestamp). Основная суть CDC в данном алгоритме - это выявление более новых записей. Для этого

CDC должен зафиксировать время начала последней загрузки. То есть в stage области должна быть таблица журнала загрузок с временем последней инкрементальной загрузки, например

**LOAD_JRN(start Jimestamp, finish Jimestamp, status) ,где**

*start Jimestamp* - время начала загрузки,

*finish Jimestamp* - время окончания загрузки,

*status* - *состояние* (успешно \ не успешно закончилась загрузка).

После этого в начале инкрементальной загрузки CDC фиксирует дату начала новой загрузки, ищет дату начала последней загрузки, а затем в таблицы stage должен загрузить новые записи следующим образом:

INSERT INTO As,g(F], F2, change Jimestamp 1,2)
SELECT Fb F2, change Jimestamp
FROM A@"DBLINK_SOURCE"
WHERE change Jimestamp > НАЧАЛО ПРЕДЫДУЩЕГО
AND change Jimestamp <= НАЧАЛО – ТЕКУЩЕГО

Таким образом, в $A_{stg}$ попадают новые и изменённые данные. Далее вычисляется дифференциал через ETL задачи, работающие следующим образом: Пусть $A_{dwh}$ - таблица ХД, соответствующая таблице источника A.

$F_{pk}$ - первичный ключ таблицы источника Тогда
Данные для изменения: SELECT FROM $A_{stg}$
WHERE Fpk in (SELECT $F_{pk}$ FROM $A^{dwh}$)
Данные для добавления:
SELECT FROM $A_{sfg}$
WHERE FpkNOTIN (SELECT Fpk FROM $A_{dwh}$)

Причина несоответствия измененных данных устраняется внедрением stage области.

Предлагаемые выше алгоритмы реализации методов предотвращения ошибок позволят разработать инкрементальную загрузку данных в ХД в соответствии с потребностями банков Республики Таджикистан.

**Литература:**


1. Madhu Tumma. Oracle Streams, High Speed Replication and Data Sharing. Rampant, 2005 - 289p.
2. Барсегян А.А., Куприянов М.С., Степанеыко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. - СПб.: БХВ-Петербург, 2004. - 336 е.: ил. ISBN 5-94157-522-X
3. Фейерштейн С., Прибыл Б. Oracle PL\SQL для профессионалов 3-е изд. СПб.: Питер, 2003. - 941 е.: ил. ISBN 5-318-00528-4
4. Л. Хоббс, С. Хилсон, Ш. Лоуенд Oracle 9iR2: Разработка и эксплуатация хранилищ баз данных, М.: Кудиз-Образ, 2004. - 585 С.


**Рецензент: к.т.н., доцент Глазунов Д.В.**