DOI 10.37882/2223-2966.2023.6-2.15

ИССЛЕДОВАНИЕ И РАЗРАБОТКА МЕХАНИЗМА РЕАЛИЗАЦИИ ПРЕДРАСЧЕТНЫХ ТАБЛИЦ В СТРАТЕГИИ ЗАГРУЗКИ ДАННЫХ MERGE

Землин Никита Алексеевич
ФГБОУ ВО «Кубанский ГАУ имени И.Т. Трубилина»
darnik141@gmail.com
Тюнин Евгений Борисович
кандидат экономических наук, доцент,
ФГБОУ ВО «Кубанский ГАУ имени И.Т. Трубилина»,
г. Краснодар
tunin_ora@mail.ru

Аннотация: Предложено использовать таблицы предрасчета для оптимизации процесса загрузки данных и рассмотрен механизм реализации данной оптимизации при помощи инструмента Pentaho Data Integration (PDI). В заключении подведены итоги и утверждено, что использование механизма предрасчетных таблиц для загрузки данных из ODS в DM слой для стратегии Merge является экономически эффективным решением, позволяющим не увеличивать затраты на проект и продолжать работу с постоянно растущим объемом данных.

Цель — разработка механизма предрасчетных таблиц для загрузки данных из ODS в DM слой для стратегии Merge.

Метод или методология проведения работы: в статье использовались методы проектирования и создания информационной системы.

Результаты: получен наиболее продуктивный и качественный метод загрузки данных на основе стратегии Merge.

Область применения результатов: полученные результаты целесообразно применять крупным предприятиям, осуществляющими постоянную загрузку большого количества данных.

Ключевые слова: загрузка данных, ETL, ETL-технологии, информационные системы, стратегия загрузки, ODS, DM, таблица предрасчета.

RESEARCH AND DEVELOPMENT OF A MECHANISM FOR IMPLEMENTING PRE-CALCULATED TABLES IN THE MERGE DATA LOADING STRATEGY

N. Zemlin
E. Tyunin

Summary: This article describes the process of loading data from the operational storage layer (ODS) to the user's aggregate and map (DM) area, and focuses on the Merge strategy, which is the most popular and in demand. The use of pre-calculation tables to optimize the data loading process is proposed and the mechanism for implementing this optimization using the Pentaho Data Integration (PDI) tool is discussed. It concludes by stating that the use of pre-calculation tables mechanism for loading data from ODS to DM layer for Merge strategy is a cost-effective solution that allows not to increase the project cost and to continue working with the ever-growing amount of data.

Purpose: To develop a mechanism for pre-calculated tables to load data from ODS to DM layer for Merge strategy.

Work method or methodology: the paper used methods of designing and creating an information system.

Results: the most productive and qualitative method of data loading based on Merge strategy was obtained.

Scope of the results: it is advisable to apply the obtained results to large enterprises, which carry out the constant loading of large amounts of data.

Keywords: data loading, ETL, ETL-technologies, information systems, loading strategy, ODS, DM, pre-calculation table.

В настоящее время информационные технологии играют все более важную роль в постоянно развивающемся мире. Они широко используются в научных исследованиях, предпринимательской деятельности и медицине. Использование информационных технологий и систем является необходимым и востребованным практически во всех областях деятельности. Автоматизирование рабочих процессов с помощью информационных технологий позволяет повысить качество предоставляемых услуг и труда, а также увеличить прибыль. Благодаря использованию информационных технологий становится возможным также сбор и анализ большого объема данных, что позволяет принимать более обоснованные решения и выявлять тенденции в различных процессах. Большое количество данных, собранных с помощью информационных технологий, можно быстро и грамотно преобразовывать, выявляя необходимую информацию и десятков террабайт данных.

Ведя речь конкретно про стратегии, при помощи которых данные поступают из оперативного слоя хранения данных (ODS) в область пользовательских агрегатов и витрин (DM) можно выделить следующие виды: merge, increment, replace — данные три типа наиболее популярные, но также ещё существуют такие виды, как scd, scd2, scd2audit.

Стратегиями загрузки данных из ODS слоя в DM в основном являются три первых, хотя в очень редких ситуациях они могут применятся и для загрузки данных с источников (SRC) в ODS.

Наиболее востребованной и имеющей популярность среди прочих стратегий загрузки является Merge. При этом типе стратегии, данные из источника-таблицы ODS поступают в DM при помощи основного запроса, фильтр может подвергаться кастомизации в плане

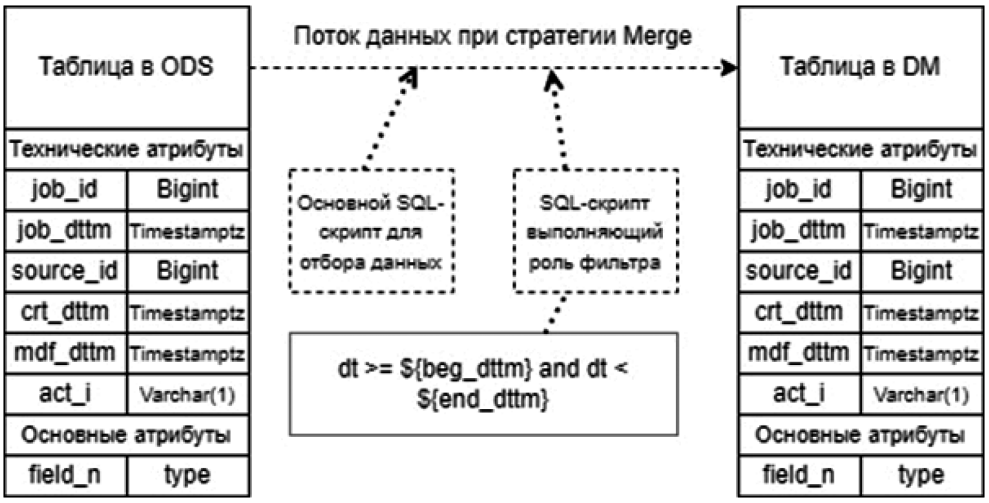

Рис. 1. Схема потока данных стратегии загрузки Merge

указания ключевого поля, по которому будет производиться фильтрация. Чаще всего используется атрибут mdf_dttm — дата изменения записи, но в редких индивидуальных случаях может применяться и crt_dttm. Схема потока данных стратегии Merge представлена на рисунке 1.

Помимо вставки данных в таблицу, может происходить как их обновление, так и удаление, метка действия кроется в атрибуте act_i.

Данная стратегия загрузки подходит для таблиц средних размеров, но чаще всего применяется для крупных источников данных, где зачастую пользователи сталкиваются с проблемой, что обработать такой объем данных не получается из-за нехватки вычислительных мощностей сервера. Бесконечно увеличивать ресурсы не получится, поэтому в данной статье будет продемонстрирован альтернативный путь решения, а именно использования таблиц предрасчета, которые позволяют в рамках одной загрузки последовательно рассчитывать блоки данных отдельно друг от друг, всякий раз освобождая ресурсы системы.

Для реализации механизма было выбрано средство Pentaho Data Integration (PDI) — это инструмент для интеграции данных с открытым исходным кодом, предоставляемый компанией Pentaho. Он позволяет быстро и легко извлекать, преобразовывать и загружать данные из различных источников, таких как базы данных, файлы, веб-сервисы и другие приложения.

Данное исследование направлено на улучшение уже имеющейся стратегии загрузки данных, разработанной в ETL-средстве PDI. Все начинается с условия Filter rows, которое имеет Boolean значение и если оно равно True, то происходит обычная загрузка данных, без использования предрасчетных таблиц. Однако, если же значение равно False, то идет инициализация первой предрасчетной таблицы fr_temp с последующей инициализацией джоба exsql_temp_pre_tbl, который содержит в себе SQL-скрипт, инициализирующий новое табличное пространство в виде временной таблицы. После инициализируется джоб логирования wlog_temp, который удобным обозначением передает сообщение о состоянии в систему логирования PDI (рис. 2).

После выполнения двух заданий в виде filter rows и rf_temp происходит проверка на необходимость создания второй предрасчетной таблицы, и если вновь условие равно True, то описанные выше действия повторяются. Далее идет выполнение джоба exsqls_temp_tbl, содержимое которого представлено на рис. 3.

Стратегия Merge позволяет производить удаление, обновление и вставку записей в зависимости от содержимого атрибута act_i в обрабатываемой строке. Где параметр I — вставка, D — удаление и U — обновление. В связи с этим данная стратегия и обладает повышенной популярностью, так как зачастую в крупных базах данных большинство ключевых таблиц является фактовыми для заполнения которых и подходит данная стратегия. Временная таблица загружается по такому же сценарию, как обычная. Идет наложение фильтра, основная выборка, джоин из таблицы, проверка идентификатора действия.

Разработанный механизм предрасчетной таблицы для загрузки данных их ODS в DM слой для стратегии Merge является экономически эффективным решением. Внедряя данный механизм в стандартную стратегию загрузки данных, открывается возможность не увеличивать затраты на проект, а именно на вычислительные ресурсы, параллельно продолжая подстраиваться под постоянно растущий объем данных со стороны источника и без перебойно забирать их в целую базу данных.

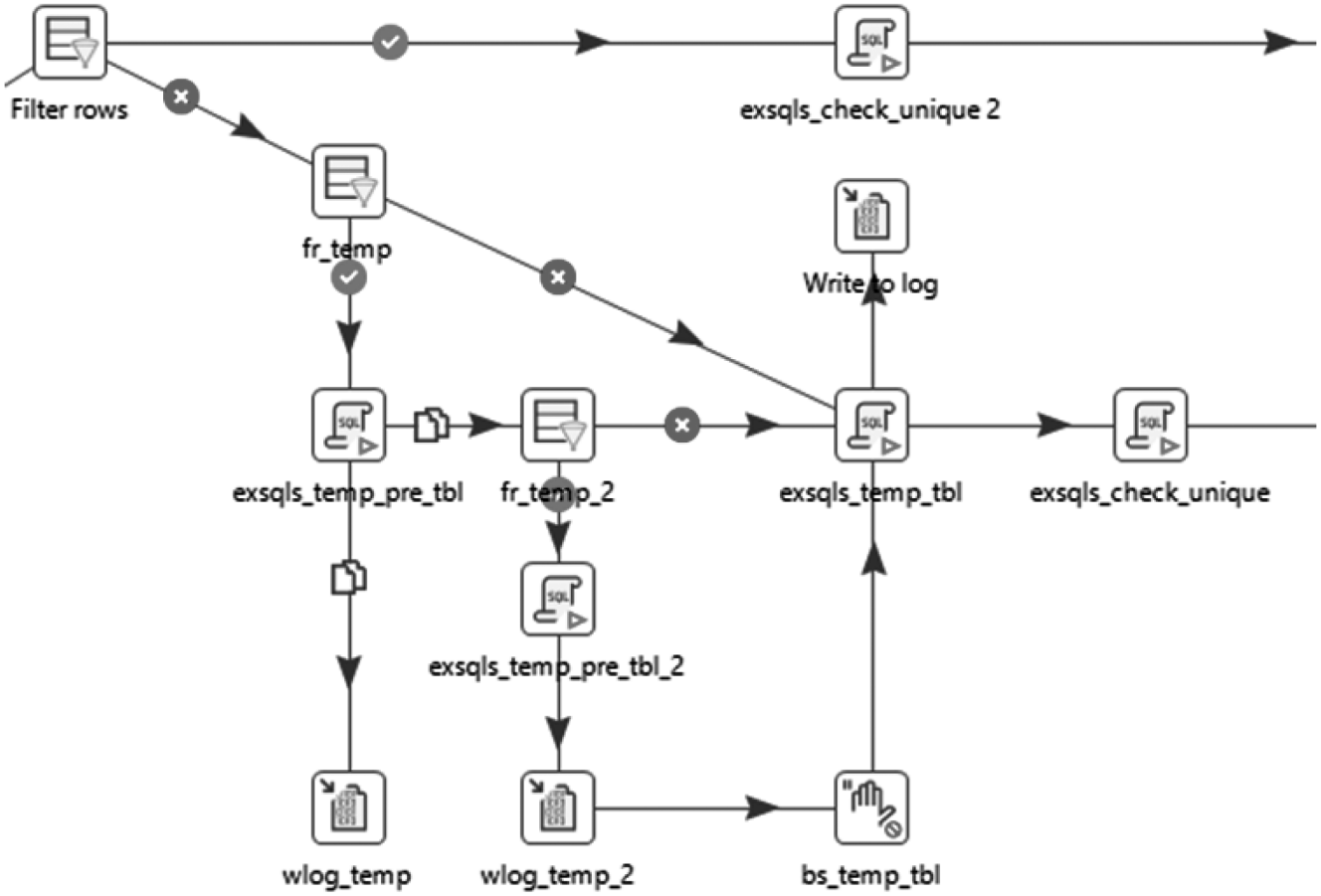

Рис. 2. Новые джобы стратегии

SQL-скрипт для выполнения. (утверждения, разделенные;) Вопросительные знаки будут заменены аргументами.

```
--Создаем временную таблицу на основе запроса из yaml-файла,
create temporary table etl${p_trg_table_name}
on commit drop
TABLESPACE ?
as (
with sel as
?,
     fltr as
         (
          select *
          from sel
          where
             ${p_src_filter_expression}
         ),
     jn as
         (
           select ?,src.act_i,?
           from fltr src
           left join
           ${%%p_con_n%%.SCHEMA}.${p_trg_tb_n} trg
              on ?
           where
              (
               ${p_cond_str} ?
               or src.act_i = 'D'
              )
         )
select * from jn
)
?
;
```

Рис. 3. Содержимое джоба exsqls_temp_tbl

ЛИТЕРАТУРА

1. Комиссаренко Н. Не только AirFlow: Apache Luigi и еще 3 ETL-оркестратора для Big Data Pipeline'ов [Электронный ресурс] URL: https://medium.com/@bigdataschool/не-только-airflow-apache-luigi-и-еще-3-etl-оркестратора-для-big-data-pipelineов-bcbe227cfee3 (дата обращения: 19.04.2023).
2. Арьков В.Ю., Бизнес-аналитика. Извлечение, преобразование и загрузка данных / В.Ю. Арьков — Москва, изд. Рейдеро, 2020. — 128 с.
3. Калачанов В.Д. Экономическая эффективность внедрения информационных технологий: учеб, пособие / В.Д. Калачанов., Л.И. Кобко. — М.: Изд-во МАИ, 2019.
4. Лукьянов Г.В. Информационная модель в проектировании информационных систем: учебное пособие / Г.В. Лукьянов. — Москва: Московский гуманитарный университет, 2016. — 29 с.
5. Как мы оркестрируем процессы обработки данных с помощью Apache Airflow [Электронный ресурс] URL: https://habr.com/ru/company/lamoda/blog/518620/ (дата обращения: 20.04.2023)
6. Лазицкас Е.А. Базы данных и системы управления базами данных: учебное пособие / Е.А. Лазицкас, И.Н. Загуменникова, П.Г. Гилевский. — 2-е изд. — Минск: Республиканский институт профессионального образования (РИПО), 2018. — 268 с.
7. Круценюк К.Ю., CASE-технологии структурного анализа. Моделирование бизнес-процессов в BPWin. Часть I: Учебное пособие / Круценюк К.Ю.// Норильский государственный индустриальный институт, 2011 г. — 124 с.
8. Благодаров А.В., Клиент-серверные приложения баз данных: Учебное пособие / Благодаров А.В., Гринченко Н.Н., Громов А.Ю.// Рязанский государственный радиотехнический университет, 2017 г. — 72 с.
9. Силен Д., Основы Data Science и Big Data. Python и наука о данных / Д. Силен, А. Мейсам., А. Мохамед. –СПб.: Питер, 2017. — 336 с.
10. Андреас В. BIG DATA. Вся технология в одной книге / В. Андреас. Изд: Эксмо, серия: Top Business Awards, — 2021. — 384 с.
11. Билл Ф., Революция в аналитике. Как в эпоху Big Data улучшить ваш бизнес с помощью операционной аналитики / Ф. Билл. Изд: Альпина Паблишер, — 2020. — 316 с.
12. Rafik A., AlievJanusz KacprzykWitold PedryczMo, 11th international conference on theory and application of soft computing, computing with words and perceptions and artificial intelligence / A. Rafik AlievJanusz KacprzykWitold PedryczMo. Изд: Springer Nature Switzerland (Zug), — 2022. — 758 с.
13. Arthur Gibadullin, Digital and information technologies in economics and management / Gibadullin Arthur, Изд: Springer International Publishing, — 2022. — 280 с.
14. Beskopylnyi A.N, Conference «INTERAGROMASH 2021». precision agriculture and agricultural machinery industry, volume 2 / A.N. Beskopylnyi, M.M. Shamtsyan // Изд: Don State Technical Universyty, — 2022. — 1058 с.
15. Computational science and its applications / 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part I. Изд: Springer-Verlag GmbH (Гейдельберг), — 2022. — 1090 с.
16. 20TH International multidisciplinary scientific geoconference sgem 2020 / Том. 2.1. Informatics, geoinformatics / Изд: Общество с ограниченной ответственностью СТЕФ92 Технолоджи (София), Albena, — 2020. — 612 с.

Журнал «Современная наука: актуальные проблемы теории и практики»