DOI 10.37882/2223-2966.2024.6-2.19

ПРОБЛЕМЫ И НЕДОСТАТКИ ОБРАБОТКИ И ЗАГРУЗКИ ДАННЫХ СРЕДСТВАМИ IPS INFORMATICA В ХРАНИЛИЩА ДАННЫХ

PROBLEMS AND DISADVANTAGES AND DOWNLOAD OF DATA PROCESSING USING IPS INFORMATICA IN DATA WAREHOUSE

E. Emelyanov

Summary. In IT systems data stored in many source systems. As rule databases, storages, systems on statistics and etc not related to each other. In data contains important information for business management, but to do this it is necessary to obtain data from diverse sources and present it in a form convenient for managers and analysts in the shortest possible time.

The purpose of the study is to identify the obvious problems and shortcomings of ETL data processing and processing.

For the provision of services, the main disadvantages of ETL tools are used when including data in a data warehouse based on IPS Informatica.

This study should clearly highlight the problems and shortcomings of processing results and loading data using ETL tools.

Keywords: data warehouse, data base, source system, mapping, structured query language.

Емельянов Егор Гурьевич
аспирант, Технологический университет имени дважды героя Советского Союза, летчика-космонавта А.А. Леонова
did-qvi@mail.ru

Аннотация. В информационных системах данные хранятся в разных источниках. Часто это не связанные между собой БД, хранилища событий, системы статистики и т.п. В этих данных есть важная информация для управления бизнесом, но для того необходимо достать данных из разнородных источников и представить их в удобном для менеджеров и аналитиков виде за минимально возможное время.

Целью исследования является обозначить явные проблемы и недостатки обработки и загрузки данных ETL средствами.

Задачей является рассмотреть основные недостатки ETL средств при загрузке данных в хранилище данных на примере ПО IPS Informatica.

Результатом данного исследования должно являться явное обозначение проблем и недостатков обработки и загрузки данных с помощью ETL средств.

Ключевые слова: хранилище данных, база данных, система источник, маппинг, язык структурированных запросов.

Системы хранения данных: это системы, которые хранят большие объемы данных из различных источников и предлагают единое представление данных для целей анализа. Они служат централизованным хранилищем, позволяя организациям получать доступ к данным из нескольких источников и анализировать их в упрощенном порядке. Такие системы используются с целью поддержки принятия решений по управлению и привлечению потенциальных клиентов, анализа эффективности той или иной политики компании и для составления различной необходимой отчетности. В подобных системах, как правило, объем данных исчисляется в терабайтах и петабайтах и по этой причине очень важно обеспечить максимальное короткое время отклика получения ответа на запрос которое должно исчисляться секундами или минутами.

Архитектура таких систем стоит из следующих элементов:

1. Хранилище данных. Хранилище данных собирает информацию из разных источников, включая приложения, файлы и базы данных. Оно обрабатывает информацию с помощью различных инструментов, чтобы данные были готовы к аналитическим целям. Например, хранилище данных может собирать информацию из реляционной базы данных, в которой данные хранятся в таблицах строк и столбцов, и совмещать с данными из не реляционных БД или различными API сервисами.
2. Инструменты ETL. Инструменты извлечения, преобразования и загрузки (ETL) — это процессы базы данных, которые автоматически извлекают, изменяют и подготавливают данные к формату, подходящему для аналитических целей. Хранилища данных используют ETL для преобразования и стандартизации информации из различных источников перед тем, как сделать ее доступной для инструментов OLAP.
3. Сервер OLAP. Сервер OLAP — это базовая машина, которая питает систему OLAP. Он использует инструменты ETL для преобразования информации в реляционных базах данных и подготовки их к операциям OLAP.
4. База данных OLAP. База данных OLAP — это отдельная база данных, которая подключается к хранилищу данных. Инженеры по обработке данных иногда используют базу данных OLAP, чтобы предотвратить нагрузку на хранилище данных

анализом OLAP. Они также используют базу данных OLAP для упрощения создания моделей данных OLAP.

5. Аналитические инструменты OLAP. Бизнес-аналитики используют инструменты OLAP для взаимодействия с кубом OLAP. Они выполняют такие операции, как нарезка, нарезка в виде кубов и поворотные операции для получения более глубокого понимания конкретной информации в OLAP-кубе.

Системы OLAP используются для аналитической обработки данных, объединенных из различных источников. В таких системах осуществляются сложные запросы к большим объемам исторических данных для проведения комплексной аналитики, которая в дальнейшем может помочь в принятии бизнес-решений по управлению и развитию компании или предприятия в целом.

Можно выделить следующие характеристики OLAP систем:

1. Сосредоточенность на сложных запросах данных: Базы данных OLAP предназначены для обработки сложных запросов к данным, включающих множество измерений и иерархий. Это позволяет проводить расширенный анализ данных и выявлять закономерности и тенденции.
2. Многомерный анализ: Базы данных OLAP оптимизированы для многомерного анализа. Это предполагает анализ данных по нескольким осям или измерениям и позволяет пользователям исследовать взаимосвязи и корреляции между различными наборами данных.
3. Использование в аналитических системах: Системы OLAP обычно используются в аналитических системах, таких как инструменты бизнес-аналитики (BI), хранилища данных и системы поддержки принятия решений. Эти системы требуют сложных возможностей анализа и отчетности для поддержки принятия бизнес-решений.
4. Низкий объем крупных транзакций: Базы данных OLAP обрабатывают небольшой объем крупных транзакций, эффективно обрабатывая обновления или вставки данных. Основное внимание уделяется анализу данных, а не манипулированию ими.
5. Денормализованная структура данных: Базы данных OLAP имеют денормализованную структуру данных. Это означает, что данные хранятся таким образом, что снижается необходимость в сложных соединениях при запросе данных. Это приводит к сокращению времени ответа на запросы и повышению производительности.
6. Оптимизирован для операций чтения: Системы OLAP оптимизированы для операций чтения. Это позволяет им обрабатывать большое количество запросов и запросов на получение данных. Это критически важно для приложений, требующих быстрого и эффективного анализа данных.
7. Высокая задержка данных: Системы OLAP имеют высокую задержку данных. Эта задержка возникает потому, что системе необходимо обработать и агрегировать данные, прежде чем сделать их доступными для анализа, что создает разрыв между временем обновления данных и их доступностью для анализа.

Инструменты ETL, часто используемые в сочетании с SQL, являются основополагающими элементами инженерии данных, предназначенными для решения сложных задач управления данными. Эти инструменты могут извлекать данные из многих источников, будь то традиционные реляционные базы данных, системы NoSQL или облачные платформы. Однако их настоящее мастерство проявляется на этапе трансформации. Здесь данные подвергаются тщательной очистке для устранения аномалий, обогащению для повышения их ценности и структурированию, чтобы сделать их пригодными для аналитических целей. Помимо этих основных функций, современные инструменты ETL решают проблемы, связанные с большими данными и аналитикой в реальном времени. Теперь они предлагают возможности потоковой обработки, позволяющие предприятиям обрабатывать данные в режиме реального времени, а также интеграцию машинного обучения для прогнозирования тенденций и аномалий. Кроме того, с развитием облачных вычислений многие инструменты ETL теперь являются облачными, что обеспечивает масштабируемость, гибкость и экономическую эффективность. Их интеграция с современными решениями для хранения данных обеспечивает предприятиям бесперебойный конвейер данных от извлечения данных до генерации аналитической информации. В более широком контексте хранилищ данных и аналитики инструменты ETL являются не просто помощниками; они являются движущей силой, предоставляющей предприятиям возможность использовать истинный потенциал своих данных.

На техническом уровне процессы ETL и ELT включают в себя несколько этапов:

1. Извлечение данных: это первый шаг в извлечении данных из различных систем. Исходные данные могут быть в разных форматах, а процесс извлечения гарантирует, что они будут переданы в инструмент ETL для дальнейшей обработки.
2. Преобразование данных: после извлечения данные преобразуются. Это включает в себя очистку данных для устранения несоответствий, их обогащение для повышения их ценности и структурирование для обеспечения пригодности для анализа. Python и SQL часто используются для манипулирования и обработки данных на этом этапе.

3. Загрузка данных. Последний шаг включает загрузку обработанных данных в целевую базу данных или хранилище данных. В зависимости от требований это может быть полная загрузка, при которой загружаются все данные, или инкрементная загрузка, при которой загружаются только новые или измененные данные.

Инструменты ETL предлагают несколько преимуществ, которые делают их незаменимыми в современной инженерии данных:

1. Качество данных: инструменты ETL обеспечивают качество данных, устраняя несоответствия и аномалии. Такие функции, как очистка и проверка данных, играют в этом решающую роль.
2. Процесс интеграции данных. Поскольку необработанные данные поступают из разных источников, интеграция их в единое целое является сложной задачей. Инструменты ETL упрощают эту интеграцию, плавно объединяя данные из различных источников. Экономия времени: ручные процессы ETL отнимают много времени и подвержены ошибкам. Инструменты ETL автоматизируют эти процессы, поэтому рабочие процессы выполняются быстро и точно.
3. Масштабируемость. Современные инструменты ETL, особенно облачные решения, обеспечивают масштабируемость. Это означает, что они могут обрабатывать большие объемы данных, масштабируясь вверх или вниз в зависимости от требований. Экономическая эффективность: автоматизируя процессы ETL, предприятия могут сэкономить на затратах, связанных с ручной обработкой данных. Кроме того, облачные инструменты ETL предлагают модели ценообразования с оплатой по мере использования, гарантируя, что предприятия платят только за то, что они используют.

Хранилище данных определяется как централизованное хранилище, в котором компания хранит все ценные активы данных, интегрированные из разных каналов, таких как базы данных, неструктурированные файлы, приложения и т.д. Хранилище данных обычно создается и используется в первую очередь для целей отчетности и анализа данных. Благодаря способности хранилищ данных собирать все данные в одном месте, они служат ценным инструментом бизнес-аналитики (BI), помогая компаниям получать аналитическую информацию о бизнесе и намечать будущие стратегии.

У хранилищ данных есть несколько определяющих особенностей:

1. Предметно-ориентированность. Данный показатель означает, что информация данных в хранилище вращается вокруг некоторого предмета по сравнению с озером данных. Это означает, что на складе хранятся не все данные компании, а только интересующие их темы. В качестве иллюстрации можно построить конкретный склад для отслеживания только информации о продажах.
2. Интеграция означает, что хранилище данных имеет общие стандарты качества хранимых данных. Например, любая организация может иметь несколько бизнес-систем, отслеживающих одну и ту же информацию. Хранилище данных действует как единый источник достоверной информации, предоставляя самую свежую и подходящую информацию. Временной вариант относится к согласованности хранилища данных в течение определенного периода, когда данные переносятся в хранилище и остаются неизменными. Например, компании могут работать с историческими данными, чтобы узнать, какими были продажи 5 или 10 лет назад в отличие от текущих продаж.
3. Энергонезависимость подразумевает, что как только данные попадают в хранилище, они остаются там и не удаляются с новыми записями данных. Таким образом, при необходимости можно восстановить старые архивные данные. В кратком изложении затрагивается тот факт, что данные используются для анализа данных. Часто они агрегируются или сегментируются в витринах данных, что облегчает анализ и составление отчетов, поскольку пользователи могут получать информацию по подразделениям, разделам, отделам и т.д.

Современные хранилища данных имеют три стандартных подхода к построению уровней архитектуры: одноуровневую, двухуровневую и трехуровневую архитектуру. Наиболее распространенной является трехуровневая модель, состоящая из нижнего, среднего и верхнего уровней.

- Нижний уровень представлен системами отчетов, обычно системами реляционных баз данных. Разнообразные серверные инструменты позволяют извлекать, очищать, преобразовывать и загружать данные на этот уровень. Существует два разных подхода к загрузке данных в хранилище данных: ETL и ELT. Оба процесса включают функции «Извлечение», «Загрузка», «Преобразование», но в разной последовательности.
- Средний уровень служит посредником между базой данных и конечным пользователем. Это дом для сервера OLAP (онлайн-аналитическая обработка), который преобразует данные в форму, более подходящую для анализа и запросов.
- Верхний уровень называется интерфейсным или клиентским уровнем. Он содержит API (интерфейс прикладного программирования) и инструменты, предназначенные для анализа данных, составления отчетов и интеллектуального анализа данных (процесс обнаружения закономерностей в больших наборах данных для прогнозирования результатов).

Загрузка данных в хранилища осуществляются с помощью ETL средств, в которых задаются правила обработки и загрузки данных из не реляционных и реляционных баз данных. Одним из таких средств является IPS Informatika которая, принимая на вход правила осуществляет загрузку данных из различных баз данных в некоторый CASH (далее кэш), осуществляет трансформации над такими данными, согласно переданным правилам, и после обработки помещает их в конечную таблицу приемник в базе данных.

ETL средства в ПО Informatika обладают следующими преимуществами:

1. Консолидированное представление данных. ETL обеспечивает консолидированное представление данных для углубленного анализа и отчетности. Управление многочисленными наборами данных требует времени и координации и может привести к неэффективности и задержкам. ETL объединяет базы данных и различные формы данных в единое, унифицированное представление. Процесс интеграции данных улучшает качество данных и экономит время, необходимое для перемещения, категоризации или стандартизации данных. Это облегчает анализ, визуализацию и осмысление больших массивов данных.
2. Точный анализ данных. ETL обеспечивает более точный анализ данных для соответствия нормативным и регулятивным стандартам. Вы можете интегрировать инструменты ETL с инструментами обеспечения качества данных для профилирования, аудита и очистки данных, обеспечивая их достоверность.
3. Автоматизация задач. ETL автоматизирует повторяющиеся задачи обработки данных для эффективного анализа. Инструменты ETL автоматизируют процесс миграции данных, и вы можете настроить их на периодическую интеграцию изменений данных или даже во время выполнения. В результате инженеры по обработке данных могут больше времени уделять инновациям и меньше — решению таких утомительных задач, как перемещение и форматирование данных.

ETL средства в ПО IPS Informatica обладают следующими недостатками:

1. Масштабирование. Как было указано выше, перед транспортировкой данных их необходимо обработать. Для этого данные предварительно материализуются и индексируются в кэше. Фактически кэш представляет из себя файл, элементы которого разбиты на массивы значения которого содержат первичный ключ, указатели на дочерний связанный объект и номер главного элемента массива. Такой подход удобен с точки горизонтального расширения т.к. в этом случае будут увеличиваться элементов внутри самого массива, а само кол-во массивов будет статично на последующих этапах обработки данных. Обратная ситуация с вертикальным расширением, при котором увеличивается как кол-во входящих массивов, так и операций по загрузке и выгрузке данных из кэша тем самым значительно увеличивая общий объем памяти выделяемого для кэша, для обработки данных, и приводя к резкому снижению времени раскрытия и обработки элементов внутри каждого массива.
2. Переполнение кэша. Кэш является промежуточной областью хранения данных элементы которого защищаются в случае, если в дальнейшей обработке их участие исключается. Однако это правило не работает в случае, например, если связи 1 к 100000000 т.е. 1 элемент массива связан с 100000000 элементами других массивов. Как правило данная проблема решается расширением общего объема кэша путем дополнительных вычислительных ресурсов сервера. В случае если ответ от сервера не был получен или же превышено время ожидания ответа, то трансформации, выполняемые в рамках текущей транзакции, будут отменены с соответствующей ошибкой о переполнении кэша. На практике данную проблему решают путем создания несколько кэшей доступ к которым осуществляется последовательно. Однако такой подход увеличивает время трансформации во много раз т.к. остальные процессы простаивают по причине ожидания завершения основного процесса и выделении ресурса для их работы

Выводы

В рамках данной статьи рассмотрены основные недостатки ETL средств ПО IPS Informatica при загрузке данных в хранилище данных. По результатам проведенного анализа можно сделать вывод о том, что при увеличении кол-ва обработки данных возникает необходимость в выделении дополнительных вычислительных ресурсов сервера, а в случае их отсутствия, или недостатка, существенное снижение производительности и времени обработки данных.

ЛИТЕРАТУРА

1. Сбор, объединение и преобразование данных с помощью Power Query. 2022. [с. 128–256].
2. Справочная документация по эксплуатации программного обеспечения informatica [Электронный ресурс]. Режим доступа: https://docs.informatica.com .
3. Клеппман М.А. Высоконагруженные приложения. Питер, 2022. [с. 350–526].

© Емельянов Егор Гурьевич (did-qvi@mail.ru)
Журнал «Современная наука: актуальные проблемы теории и практики»