Syncsort DMExpress – высокопроизводительная интеграция данных

В конце января 2009 г. компания Syncsort анонсировала доступность 5-й версии своего продукта DMExpress для интеграции данных, выпускаемого с 2004 г. На текущий момент это одно из наиболее производительных и экономичных решений интеграции данных на рынке.

Введение

Мировой рынок продуктов Data management and integration (DM&I) активно развивается уже в течение многих лет. По прогнозам Gartner[1], к 2013 г. он составит порядка $33,2 млрд. И хотя доля продуктов Data Integration Tools (DIT) к 2013 г. в общем объеме будет значительно меньше – $2,7 млрд в сравнении с продуктами Database Management System ($30,5 млрд), их ежегодный прирост составит 12,3% против 6,7% для DMS.

Рынок DIT в последнее время получает новый импульс также за счет таких новых инициатив и развития секторов рынка как Master Data Management (MDM), Business Intelligence (BI), Service-Oriented Architectures (SOAs).

По данным уже упомянутого исследования Gartner[1], к 2013 г. 75% ресурсов, выделяемых на управление информацией, будет затрачиваться на интеграцию и анализ различных типов неструктурированных данных, соответственно, на анализ структурированных – будет уходить только 25%.

Такой интерес к DTI и BI-решениям связан, во-первых, с резко возросшим интересом к бизнес-анализу самого широкого круга компаний – от средних и небольших до самых крупных – в связи с пониманием необходимости более реактивного ведения бизнеса в современных условиях, а, во-вторых, резко возрастающими объемами хранилищ данных – до петабайт и более (рис. 1). Так, например, согласно WinterCorp TopTen программе, самые большие в мире базы данных утраиваются в размере каждые два года, начиная с 1999 г. В связи с этим все решения, позволяющие расширить границы по производительности, в зависимости от объема данных стали приобретать особую остроту.

Один из нишевых игроков, по данным отчета[2] "Gartner Magic Quadrant for Data Integration Tools" – компания Syncsort с продуктом DMExpress. Решение DMExpress является логическим развитием программного пакета SyncSort. Он предоставляет все возможности SyncSort и добавляет такой существенный функционал как оптимизированный доступ к БД, управление метаданными, поддержка всех типов и преобразований данных, функции, условные значения, арифметические операции и управление заданиями. Сорокалетний опыт высокопроизводительной обработки данных, непрерывная доходность и большая лояльная база заказчиков являются основой Syncsort, на которой растет его присутствие на рынке. На текущий момент решениями компании Syncsort пользуются 96 компаний из списка Fortune 100, свыше 80% компаний – из списка Dow Jones и почти все государственные учреждения США, среди которых такие гиганты, как Microsoft, Cisco, Siemens, American Express, BP, Toyota и др.

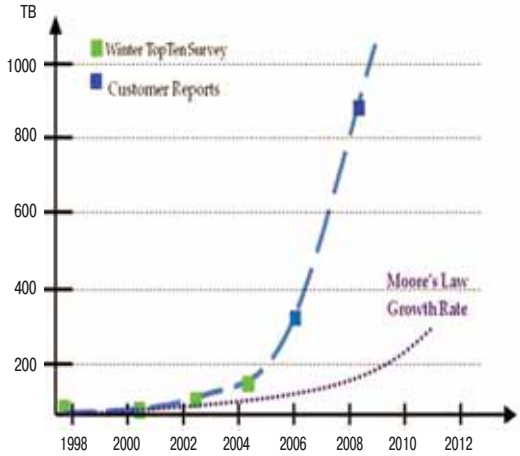

Рис. 1. Экспоненциальный рост объемов баз данных для бизнес-анализа заставляет искать пути повышения производительности решений для всех этапов, связанных с поддержанием хранилищ данных.

Среди основных вендоров, предлагающих продукты DTI, можно отметить следующие: Microsoft, Sybase, Oracle, Syncsort, Pitney Bowes (software division), Informatica, iWay, Sun Microsystems, Tibco Software, ETI, Pervasive Software, Open Text, IBM, SAS, Business Objects (an SAP company).

Данная публикация – первая из серии, посвященной особенностям DTI-продуктам.

Функциональные особенности Syncsort DMExpress

DMExpress впервые был анонсирован в 2004 г. В начале 2009 г. была выпущена уже пятая версия этого продукта.

Высокоскоростная обработка данных DMExpress – результат патентованных алгоритмов по производительности, современной параллельной технологии обработки и многолетних исследований.

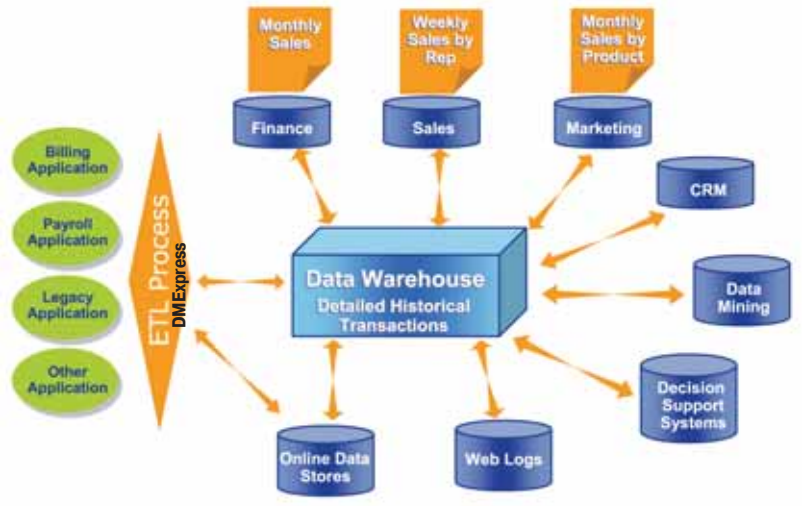

Рис. 2. Использование DMExpress для ускорения ETL-процессов может стать ключом для своевременного принятия бизнес-решений и эффективного управления.

1) "The Gartner Data Management and Integration Vendor Guide, 2009", 24 April 2009, Regina Casonato, Mark A. Beyer, Ted Friedman, Gartner RAS Core Research Note G00167064.

2) "Magic Quadrant for Data Integration Tools", 22 September 2008, Ted Friedman, Mark A. Beyer, Andreas Bitterer, Gartner RAS Core Research Note G00160825.

DMExpress может быть развернут на UNIX-, Windows- и Linux-платформах.

DMExpress позволяет интегрировать большой объем данных из множественных гетерогенных источников, преобразовывая и объединяя данные с высокой скоростью, создавая объединенные обозрения для отчетов, анализа или других приложений.

DMExpress в большей степени ориентирован на пакетную обработку, чем на онлайновую. Однако при этом имеет ряд преимуществ. Так, при обработке больших томов данных DMExpress за счет внутренних механизмов оптимизации позволяет в значительной степени ускорить или заменить долготекущие процессы на более быстрые. В дополнение к этому, DMExpress меньше загружает центральный процессор и меньше использует ресурсов памяти. Если достаточно часто приходится выполнять крупномасштабные пакетные задания, DMExpress может значительно снизить потребность в аппаратных инвестициях, сокращая время обработки заданий на часы и даже дни.

Медленные операции ETL (Extract, Transform, Load) влияют на все процессы, которые производятся с данными далее. Эту проблему чувствуют на самых высоких уровнях, где требуется деловая аналитика, на ежедневном бизнес-уровне, где актуальная информация необходима для сделок, продаж, управления и отчетности.

Использование DMExpress для ускорения ETL-процессов может стать ключом для своевременного принятия бизнес-решений и эффективного управления, в частности, для онлайновых и ряда других BI-приложений:

- Changed Data Capture (Delta Processing);
- Horizontal Data Pivot;
- Multi-level Hierarchical Aggregation;
- Join/lookupWeb Log Processing;
- Data Standardization, Cleansing, and Validation;
- Data Warehousing;
- Database and Batch Loads;
- Data Mining;
- Customer Relationship Management;
- BPM;
- EAI;
- OLTP Systems.

Для ускорения ETL-операций DMExpress интегрирует набор следующих технологий:

- коммерческие алгоритмы;
- патентованные методы управления данными;
- динамическую оптимизацию;
- быстрый ввод/вывод;
- современную параллельную обработку.

Разработанный для повышения скорости, DMExpress может сокращать время для процессов ETL до 90%, ускоряя задачи от простых загрузок базы данных до формирования крупномасштабных корпоративных информационных хранилищ. Безотносительно к операциям ETL, DMExpress позволяет ускорить:

- чтение баз данных и простых файлов непосредственно, используя расширенные методы ввода/вывода;
- загрузку непосредственно в базу данных или простых файлов, устраняя отдельные шаги загрузки. Использование DMExpress для предварительной сортировки вне базы данных драматично уменьшает время, требуемое для загрузки и индексирования реляционных баз данных, таких как: Oracle, SQL Server и DB2;
- агрегирование данных с целью ускорения обработки запроса;
- любые преобразования на уровне записей и отдельных полей.

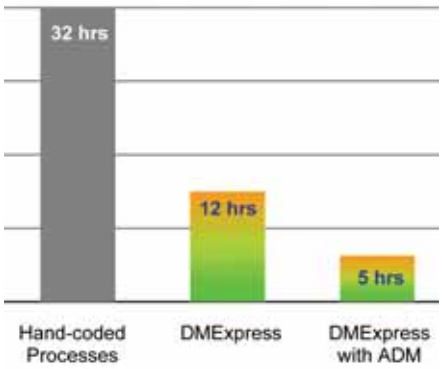

Рис. 3. Использование дополнительных компонент DMexpress, в частности, Advanced Data Management может существенно улучшить время обработки запросов – в 2 и более раз.

Решение легко развертывать для построения, обслуживания и внедрения приложений, при этом персонала требуется значительно меньше, чем для альтернативных решений. ETL-приложения легко разрабатываются любым сотрудником с минимальными навыками программирования. Операционная гибкость позволяет быстро адаптироваться под изменяющуюся информационную либо бизнес-среду.

DMExpress имеет ряд дополнительных компонент, позволяющих расширять возможности базового продукта, в частности, это:

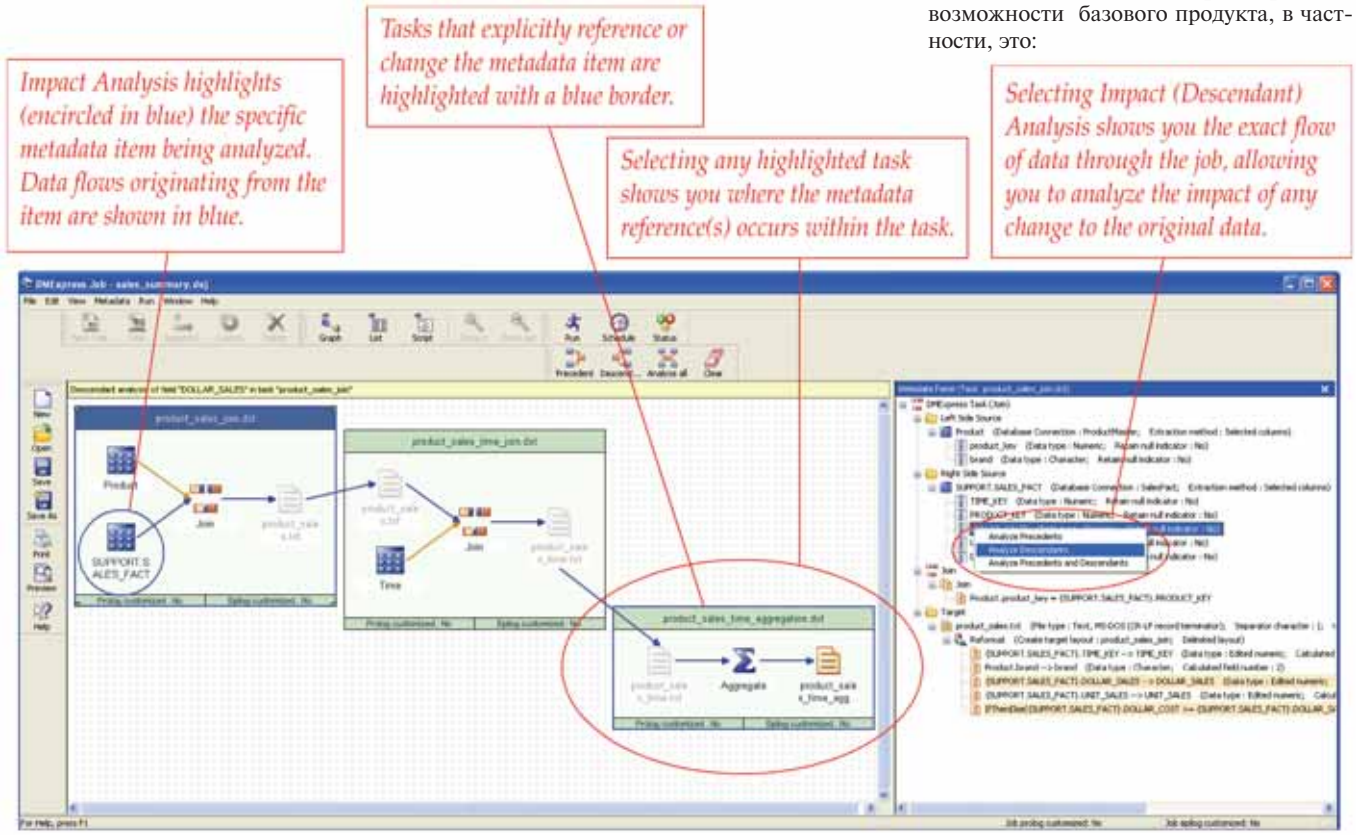

Рис. 4. Компонета Impact Analysis оценивает эффект изменения определенной переменной и отражает этот эффект графически. Можно проследить все места, где на определенный элемент данных ссылаются, анализируя поток метаданных в пределах Job Editor.

– Advanced Data Management (ADM);
– Data Source and Data Target;
– Impact Analysis;
– Grid Computing;
– ES/MTO поддержку.

Advanced Data Management

ADM – отдельно лицензируемая компонента Syncsort's DMExpress. ADM включает следующий функционал:

– высокопроизводительную агрегацию данных;
– высокопроизводительное соединение данных;
– расширенные возможности по преобразованию данных, включая: UNICODE-поддержку, целевое отображение размещения, внешние функции, арифметические функции, функции даты и времени, строковые функции, определяемые пользователем значения, разделение данных и расширенные функции агрегации данных.

При этом ускорение операций (по данным разработчика) может составить 2 и более раз (рис. 3).

Data Source and Data Target

Базовая версия DMExpress обеспечивает поддержку гетерогенным источникам и адресатам, включая простые файлы, структурированные файлы, каналы и данные в памяти. Отдельно лицензируемые Data Source and Data Target компоненты обеспечивают прямой доступ к таблицам базы данных и другим данным.

DMExpress Data Source and Data Target включает поддержку:

– XML источника и адресата;
– основных источников и адресатов для реляционных баз данных:
  – Oracle;
  – SQL Server;
  – DB2;
  – Teradata;
  – Sybase;
  – Red Brick;
  – Vertica;
  – Netezza;
– как источника и адресата для ODBC, которые обеспечивают прямой доступ к MySQL, Access и Excel;
– как источника для мэйнфреймов;
– как источника для SAP.

Данная компонента позволяет существенно сократить ввод/вывод, устранить извлечение и загрузку утилит, упростить развитие приложений, а также увеличить производительность приложений.

Impact Analysis

Отдельно лицензируемая компонента – Impact Analysis – позволяет графически отражать эффект изменения глобальной переменной; идентифицирует, какие входные переменные влияют на значение определенной выходной переменной, а также обеспечивает глобальный поиск переменных среди всех заданий.

Impact Analysis включает три особенности, которые работают как hand-in-hand и дающие возможность разработчикам приложений быстро понять элементы потока задачи:

– Impact Analysis;
– Lineage Analysis;
– Global Find.

Impact Analysis оценивает эффект изменения определенной переменной и отражает этот эффект графически (рис. 4). Можно проследить все места, где на определенный элемент данных ссылаются, анализируя поток метаданных в пределах Job Editor (Редактор задания).

Lineage Analysis идентифицирует, какие входные переменные влияют на значение определенной переменной вывода. Можно проследить все места, где на определенный элемент данных ссылаются, анализируя поток метаданных в пределах Редактора задания.

Global Find позволяет идентифицировать, какие задания и задачи затрагиваются, когда необходимо изменить какие-либо приложения. Это может быть полезным первым шагом в анализе воздействия или анализе происхождения, определяя местонахождение всех заданий, где интересующий элемент происходит.

Grid Computing

Отдельно лицензируемая компонента обеспечивает высокопроизводительную архитектуру для CPU-интенсивных приложений, давая возможность пользователям использовать коллективную мощность обработки множества компьютеров.

ES/MTO Support

Эта компонента обеспечивает поддержку MicroFocus Enterprise Server с Mainframe Transaction Option (ES/MTO), давая следующие преимущества:

– бесшовную интеграцию с ES/MTO языком управления заданиями сортировки шагов (PGM=SORT);
– способность установить ES/MTO переменные среды MF_ALIAS и MFJEXTSM с целью использования DMExpress технологии сортировки;
– существенное сокращение общего времени, а также использования системных ресурсов (центральный процессор, память, дисковый ввод/вывод);
– снижение требований к производительности, сопоставимое с DMExpress MicroFocus COBOL Sort verb accelerator.

Оценка производительности Syncsort DMExpress

Тестирование состояло в том, что DMExpress v4.8 извлекал, преобразовывал, чистил и загружал 5,4 Тбайт "сырых" TPC-H данных в Vertica Analytic Database за 57 минут 21,51 секунды. Тестирование проводилось на HP BladeSystem c7000 x86 серверах, работающих под управлением RedHat Linux

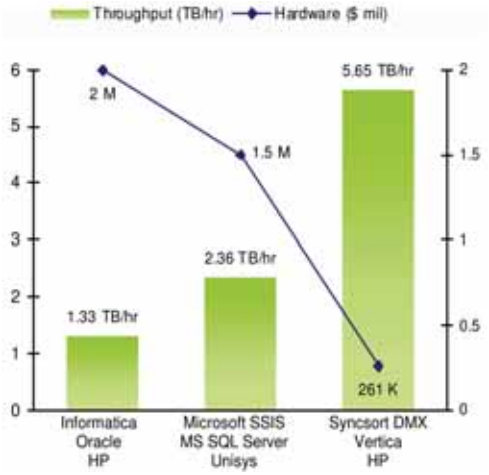

Рис. 5. Результаты тестирования DMExpress v4.8 по извлечению, преобразовыванию, чистке и загруке 5,4 Тбайт "сырых" TPC-H данных в Vertica Analytic Database на базе HP BladeSystem c7000 x86 серверах, работающих под управлением RedHat Linux ОС.

операционной системы. Результаты тестирования представлены на рис. 5.

Основные результаты тестирования:

– 5,4 Тбайт загружены за 57 минут 21.51 секунды – Syncsort, Vertica, и HP;
– предыдущий рекорд – загрузка 1 Тбайт данных за 25 минут 20 секунд – Microsoft и Unisys;
– независимый аудит компанией DSS Labs;
– характеристика процесса:
  · DMExpress читает индустриально стандартные TPC-H данные с дискового источника в память;
  · DMExpress выполняет чистку данных на источнике, чтобы гарантировать, что они соответствуют бизнес-правилам, и создает reject-файл;
  · DMExpress посылает данные в форме ASCII с разделителем полей через каналы;
  · копия Vertica получает данные через каналы;
  · копия Vertica загружает данные в базу данных Vertica;
  · процесс повторен для всех таблиц.

Заключение

Возрастающая необходимость к анализу бизнес-информации со стороны компаний всех уровней с одновременно экспоненциально возрастающими объемами данных, требующими анализа при ограниченности IT-бюджетов, во многом будет способствовать оптимизации всех процессов, связанных с поддержанием DW и, в частности, ускорению ETL-процессов, которые по важности занимают первые позиции в опросах. Так, по результаты исследования, проведенного среди участников IDC CEMA BI Roadshow 2007, на вопрос: "какие функции вы планируете добавить к вашим BI-инструментам в течение следующих 12 месяцев?", 54% опрашиваемых ответили: "в хранение данных" – самый высокий приоритет, за ним идут: "сбор данных" – 45%, "ETL" – 32%.

Ирина Сайфуллина,
компания "Москит"