

BARC-Product Review

Syncsort DExpress*

[Analyst: Lars Iffert, Timm Grosser, November 2011]

Anbieter

Das Unternehmen wurde 1968 als Whitlow Computer Systems gegründet und zählt daher zu den Urgesteinen in der Software-Industrie. Von Anfang an stand die Performancesteigerung in der Datenverarbeitung auf der Fahne des Unternehmens. Hier wurde insbesondere auf die Sortierung und alle Prozesse, die auf der Sortierung basieren, wie der Join und die Aggregation Wert gelegt. Seit 1981 firmiert das Unternehmen unter dem Namen Syncsort Incorporated. Bereits in den 70er Jahren erfolgte der Markteintritt in Deutschland, Frankreich und Großbritannien. Im Jahre 2009 wurde das Unternehmen umstrukturiert. Marketing und Sales wurden deutlich verstärkt und auch das Partnergeschäft ausgebaut.

Das Gros der Kunden (70%) befindet sich in Nordamerika. Hier gehören besonders die großen Unternehmen (90% der Fortune-100-Liste) zur Kundschaft.

Strategie

Wachsende Datenvolumina und immer neue Datenquellen führen zu aufwändigen Integrationsszenarien mit der Folge, dass die Datentransformationen zum kritischen Faktor hinsichtlich Performance, Implementierungszeit und Aufwand werden. Syncsort adressiert genau diesen Bedarf mit seinem „ETL 2.0“ Ansatz. Mit Konzentration auf wesentliche ETL-Funktionalität, insbesondere das „T“, bietet Syncsort eine spezialisierte Lösung an, mit dem Ziel, durch Einfachheit und ausgereifte Technologie Kosten zu sparen und kritischen Performance-Anforderungen nachzukommen. Im Vordergrund des Lösungsangebotes steht die Spezialisierung auf das „ETL“-Geschäft – der Anspruch, die Funktionalität einer breit aufgestellten Datenintegrationsplattform anzubieten, wird bewusst nicht verfolgt. Gemäß der Ambition, die DI-Performance zu steigern, bietet Syncsort Möglichkeiten an, performance-schwache Datenintegrationsteilprozesse in bestehenden DI-Werkzeugen durch eigene performance-starke Routinen zu ersetzen. Speziell für die Sortierung großer Datenbestände greift Syncsort hier auf eine lange Entwicklungshistorie und patentierte plattformsspezifische Algorithmen zurück. Heute sind diese Transformationsbausteine für IBM, Informatica, SAP, Talend und Ab Initio im Einsatz (für IBM und Informatica teilautomatisiert über Miti-Technologie). Zuletzt ist zu erwähnen, dass nach wie vor die Beschleunigung von Datentransformationen auf dem Mainframe oder in Mainframeapplikationen aus der Historie heraus zum Kerngeschäft von Syncsort zählt (Syncsort spricht hier von „Appmod Acceleration“).

Produktportfolio & Markteinordnung

Der Schwerpunkt von DExpress liegt in der Verarbeitung hochvolumiger Daten mit dem Ziel der Integrations-Beschleunigung. Zur Performance-Steigerung greift das Produkt auf patentierte Algorithmen zurück, die in Abhängigkeit der verwendeten Hard- und Software dynamisch die bestmöglichen Treiber ansteuern.

Anbieterprofil	syncsort
Anbieter	Syncsort
Vertretungen	In über 70 Ländern, Zentrale USA, GB, F, D
Mitarbeiter	350 (D: 12)
Kunden	k.A.
Neukunden	k.A.
Umsatz	k.A.

Zahlen bezogen auf 2010

* getestet wurde Syncsort DExpress 6.9

Copyright © BARC GmbH 2011.
Alle Rechte vorbehalten.

Business Application Research
Center-BARC GmbH
Steinbachtal 2b
97082 Würzburg
+49 (0)931 880651-0
www.barc.de

Das Einsatzgebiet von DMExpress ist unabhängig von Unternehmensgröße und Branche.

Durch die Produkte „Syncsort für Mainframe“, „Syncsort für VMS“, „Backup Express“ und „Fileport“ wird die Produktpalette komplettiert. Syncsort bietet Funktionen zur Sortierung und Datenmanipulation. Backup Express ist eine Lösung zum Thema Disaster Recovery und Business Continuity die aktuell besonders in virtuellen Servern eingesetzt wird. Fileport transformiert Mainframe-Dateien von Platten oder Bändern in auf offenen Systemen (UNIX, Linux, Windows) üblichen Zeichenformaten (ASCII, Unicode).

Architektur

Das ETL-Werkzeug DMExpress kann als Server auf den gängigen Betriebssystemen (Windows, UNIX, Linux) oder als Workstation auf einem Windows-Betriebssystem installiert und betrieben werden. Die grafische Entwicklungsoberfläche zur Erstellung von Daten- und Kontrollflüssen (Task-Editor für Datenflüsse, Job Editor für Kontrollflüsse) und zur Überwachung und Steuerung des DMExpress-Server sind Windowsprogramme.

Die erstellten Prozesse werden als Dateien im Dateisystem abgespeichert und können über die Entwicklungsoberfläche, den integrierten Scheduler oder durch externe Programme als Scripts zum Ablauf gebracht werden.

DMExpress enthält standardmäßig spezielle Server-Komponenten, die bei Bedarf genutzt werden können: Grid-Funktionen ermöglichen den Ablauf der ETL-Prozesse auf mehreren Rechnern. Die Advanced Data Management Funktionen dienen der Steigerung der Performance bei Aggregationen und Joins; sie enthalten dafür Erweiterung der Datentransformations-Funktionen. Impact Analysis ermöglicht Impact-Analysen sowie ihre grafische Darstellung.

Die Performance für die Befüllung von Data Warehouses steht im Vordergrund der Lösung. Zur Performance-Steigerung werden sowohl scale-up als auch scale-out Architekturen unterstützt, eigene Akzeleratoren sowie Parallelisierungsfunktionen wie Piping oder Partitionierung. Seit der Version 5.4 bietet die Software auch integrierte Re-Start- und Wiederanlauf-Verfahren für Hochverfügbarkeitsszenarien an.

Für die Verarbeitung hochvolumiger Daten unterstützt DMExpress dynamisch die Lastverteilung verschiedener oder einzelner Prozesse auf mehrere Rechner (scale-out), unabhängig von der eingesetzten Hardware oder des Betriebssystems. Dabei entscheidet das System automatisch und selbständig wie die Last am günstigsten zu verteilen ist. Das Regelwerk zur Verteilung basiert auf patentierten Algorithmen, in denen die langjährige Erfahrung in der Performance-Optimierung von Syncsort in Umgang mit unterschiedlichen Hardware-Software-Kombinationen niedergeschrieben ist.

Modelle und Metadaten

Die Metadaten werden zusammen mit den Task-Dateien in einem eigenen proprietären Format gespeichert. Innerhalb eines Jobs kann aus der Entwicklungsoberfläche hinaus in den Metadaten navigiert oder auch Verwendungsnachweise erstellt werden. Für die Suche nach Metadatenobjekten in unterschiedlichen Jobs steht eine globale Suche zur Verfügung.

Metadaten können auf Job/Task-Ebene tabellarisch in der Entwicklungsoberfläche angezeigt werden. Nachvollziehbarkeitsanalysen (Impact-/Lineage-Analysen) sind innerhalb eines Jobs möglich.

Ein zentrales Repository kann in einem "Repository-Task" aufgebaut werden. Dazu werden alle Entwicklungsobjekte (Konnektoren, Transformationen, ...) in eine Aufgabe (Task) gezogen. Bei Neuentwicklung einer Aufgabe können die benötigten Metadaten aus dem "Repository-Task" importiert werden.

Aufgrund der datei-basierten Metadatenablage kann eine Versionierung und damit auch eine Mehrbenutzerunterstützung durch externe Source-Verwaltungsprogramme realisiert werden.

Entwicklung

Für die Entwicklung von Datenflüssen steht der Task-Editor mit Templates für die gängigen Integrationsszenarien im Data-Warehouse-Umfeld bereit, die nach Bedarf angepasst und erweitert werden können. Nachdem sich der Entwickler für eine der Szenarien Aggregate, Copy, Join, Merge oder Sort entschieden hat, können die Elemente des Templates parametrisiert und eigene Transformationen über einen integrierten Expression-Editor ergänzt werden. Die erstellten Datenflüsse können dann im Job-Editor unter Berücksichtigung von Abhängigkeiten arrangiert werden. Auf diese Weise können Datenflüsse auch mehrfach in unterschiedlichen Jobs wiederverwendet werden. Für die Wiederverwendung von Metadaten eines Quell- oder Zielsystems oder einzelner Berechnungen können die Metadaten zwischen einzelnen Datenflüssen verlinkt oder importiert werden. Die Entwicklung unterstützt DMExpress mit einfachen Debugging- und Preview-Funktionen. Für die Entwicklung in Mehrbenutzerumgebungen ist der Betrieb in Zusammenhang mit einem beliebigen externen Source-Verwaltungsprogramm zu empfehlen.

Die grafische Benutzeroberfläche, die Templates und vordefinierten Transformationsfunktionen des Expression Editors vereinfachen die Entwicklung der Datenintegrationsprozesse und führen zu einem schnellen Ergebnis. Für die Entwicklung komplexer Transformationsprozesse wie den Aufbau von Historien (Slowly Changing Dimensions) oder Hierarchien (bspw. rekursive Schleifen) kann der Entwickler zusätzliche Funktionalität in Form von C- oder Java-Programmen in den Integrationsprozess einbinden.

Wie die meisten Datenintegrationswerkzeugen richtet sich DMExpress eher an Anwender, die mit Datenhaltungen vertraut sind. Für die selten zu nutzenden Add-Ons sind einfachere Programmierkenntnisse erforderlich.

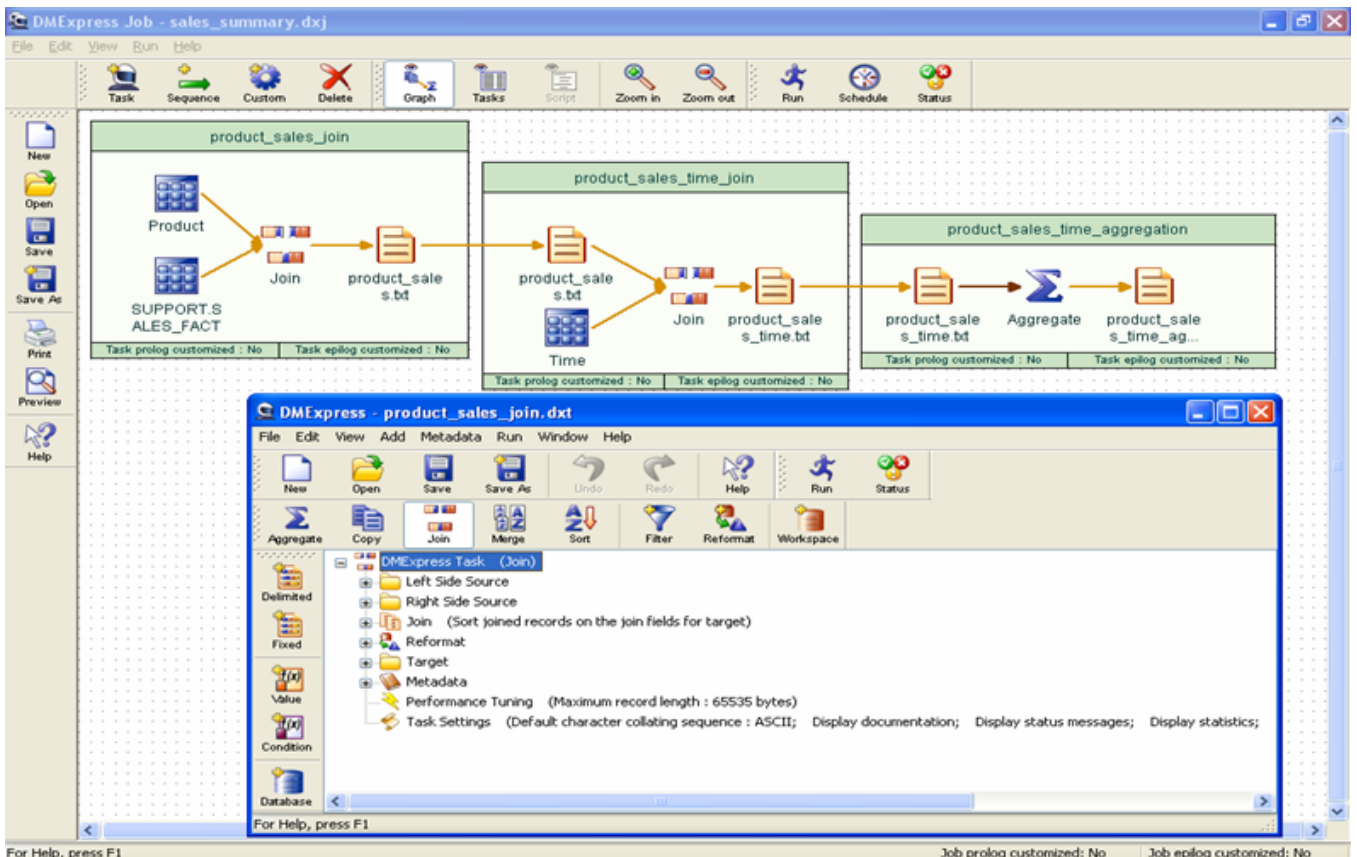


Abbildung 1: Im Hintergrund werden im Job-Editor mehrere Datenflüsse (Tasks) in einen Workflow integriert. Daten können wahlweise innerhalb des Workflows durchgängig im Hauptspeicher abgearbeitet werden. Der Task-Editor im Vordergrund zeigt das Entwicklungstemplate zur Erstellung eines Joins.

Für ausgewählte Szenarien werden vordefinierte Templates angeboten, die eine schnelle Implementierung ermöglichen und somit dem Entwickler beim Aufbau einfacher Integrationsprozesse unterstützen. Komplexere Transformationen können größtenteils durch Programmierung ergänzt werden. Für die Entwicklung selbst bietet das Werkzeug einfache Funktionen zur Testunterstützung an, zeigt im Vergleich zur Konkurrenz aber Potential zur Verbesserung. Anders als die großen Datenintegrationsplattformen bietet Syncsort vordefinierte Prozess-Templates, in denen der Entwickler anhand eines roten Ablauffadens die Komponenten konfiguriert. Es handelt sich nicht um eine grafische Entwicklungsumgebung, in der unterschiedlichste Komponenten per Drag&Drop erst zu einem Datenfluss zusammengefügt werden müssen.

Die Benutzeroberfläche ist mit dem aktuellem Release nun auch in den Sprachversionen Deutsch, Französisch, Chinesischer und Japanisch verfügbar.

Konnektivität

Über native Schnittstellen kann das Werkzeug auf nahezu jede klassische relationale Datenquelle zugreifen, wie Teradata (auch Parallel Transporter), IBM, Microsoft oder Oracle, als auch auf spalten-orientierte analytische Datenbanken wie bspw. Vertica. Zur Beschleunigung des Ladevorgangs nutzt DMExpress optimierte Techniken wie die Aufbereitung von Daten vor dem Ladeprozess. Des Weiteren werden sequentielle Dateien, Flat Files, XML-Dateien und insbesondere Mainframe Host-Dateien unterstützt. Zur Anbindung von near-time Daten über Messages oder Web-Services kann DMExpress um die Technologie des Partners Composite oder Attunity erweitert werden. Mögliche Ziele der Integrationsprozesse können relationale Datenbanken, XML-Dateien oder beliebige Text-Dateien sein. Seit Version 5.1 erweitern nun auch ein SAP BW und ERP Konnektor die Konnektivität des Integrationswerkzeuges. Zur Steigerung der Performance bietet die Lösung unterschiedliche Funktionen zur Parallelisierung wie Piping oder auch Partitionierung an. Mit dem aktuellem Release ist ein komprimiertes Lesen und Schreiben in das Hadoop HDFS möglich.

Datentransformation

Die Anwendung unterstützt grundsätzliche alle Datentransformationen (Filter / Exclude, Gruppierung, Aggregation, Berechnungen, Boolesche Logik, User Exits, Makros, Skriptsprache, Mehrstufigkeit, Schleifen, mehrdimensionale Strukturen, Sortierung). Die Datentransformation geschieht in Form von ETL und wird dem strategischem ETL 2.0 Ansatz entsprechend auch so implementiert.

Datenqualität

Syncsort positioniert sich als Spezialist für Datenintegration. Allerdings können einfache, technische Datenqualitätsoperationen wie Standardisierungsaufgaben (Datentypkonvertierung, Überprüfung auf Gültigkeitsbereiche) mit Bordmitteln umgesetzt werden. Generell gehören Datenqualitätsfunktionen nicht zum Kerngeschäft von Syncsort. Sie werden über enge Partnergeschäfte mit Trillium Software abgedeckt. Die Integration von Datenqualitätsfunktionen geschieht über den Syncsort Funktionsbaustein „Custom Task“. Die Daten werden dabei vollständig an das Datenqualitätswerkzeug übergeben und nach Verarbeitung von dort abgeholt. Die Daten können dabei über eine Datei oder direkt über den Arbeitsspeicher übermittelt werden.

Systemverwaltung

Das System ist einfach mit einem sehr kleinen Footprint zu installieren und innerhalb der bestehenden IT-Infrastruktur zu integrieren und zu betreiben. Für den Betrieb können Workflows mithilfe des Job-Editors plattformunabhängig erstellt werden. Jobs/Tasks werden plattformunabhängig gebaut und auf dem ausführenden System in der dort installierten Engine ausgeführt.

Ein Workflow besteht aus einer Abfolge von Datenflüssen und wahlweise individuellem Code. Pre- und Postprocessing-Komponenten ermöglichen den externen Programmaufruf im Workflow. Zur Ausführung wird ein Script erstellt, welches dann zum Ablauf gebracht werden kann.

Fehlerroutinen, wie bspw. Benachrichtigungen per Mail bei Fehlern, oder weitere Aktionen können über das Ablaufskript programmatisch implementiert werden.

Der Zugriff auf das Werkzeug wird über das Betriebssystem berechtigt.

Für die Protokollierung werden XML-Formate unterstützt. Eine Auditierungsfunktion wird explizit nicht angeboten. Das Scheduling ist auf Job-Ebene möglich. Weitere Möglichkeiten zur Steuerung der Prozesse sowie zur Fehlerbehandlung eröffnen Programmiermöglichkeiten im Ablaufskript. Externe Scheduler können per Commandline oder API angebunden werden.

Ein Monitoring der Prozesse ist über den Syncsort Systemmonitor und einer Analyse der Log-Dateien möglich. Externe Überwachungssysteme können über die Log-Dateien angebunden werden.

Stärken und Schwächen	
Stärken	Schwächen
<ul style="list-style-type: none"> • Gut geeignet für die performante Massendatenverarbeitung, vor allem durch optimierte Kern-Transformationen wie Sortierung, Join, Merge und Aggregation • Robustes System mit Fokus auf Beschleunigung, Performance und Skalierbarkeit • Einfache Installation und Betrieb, sowie schmaler Footprint • Bietet Optionen zur Performanceoptimierung von bestehenden IBM DataStage und Informatica Teilprozessen sowie Mainframe Applikationen • Datenintegrations-Templates für wesentliche ETL-Szenarien verfügbar • Verhältnismäßig günstig • Schnelles Beladen auch von spalten-orientierten Datenbanken 	<ul style="list-style-type: none"> • Begrenzter Funktionsumfang im Vergleich zu breit aufgestellten Datenintegrationsplattformen, da Spezialist für DI mit Fokus auf Beschleunigung von batch-orientierten Integrationsszenarien <ul style="list-style-type: none"> - Kein Fokus auf „fachliches“ Metadatenmanagement im Sinne der Bereitstellung eines Business Glossars oder Nachvollziehbarkeit der Datenherkunft für den Fachbereich - Keine direkte Unterstützung von Datenqualitätsaufgaben oder MDM - Keine direkte Unterstützung einer Einzeldatensatzverarbeitung im Sinne einer near-/real-time Datenverarbeitung • Bei Notwendigkeit einer allumfassenden Enterprise Integrationsplattform Erwerb weiterer Tools notwendig (Best-of-Breed-Ansatz) • Abbildung komplexer Transformationen über Scripting