

Microsoft®
SQL Server™ 2005

SQL Server 2005 통합 서비스에 대한 소개

요약

본 백서는 오늘날과 같은 기업 환경에서 데이터 통합 기술을 활용하여 의미 있고 안정적인 정보를 제공함으로써 경쟁 우위를 유지하려는 기업들이 당면하고 있는 과제에 대해 설명합니다. 특히 SQL Server 2005 Integration Services(SSIS)를 통해 정보 기술 부서가 회사의 데이터 통합 요건에 부합할 수 있는지를 중점적으로 논의하며 실제 시나리오를 예를 들어 설명합니다.

본 문서는 예비 문서이며 여기에서 설명된 소프트웨어의 최종 상용 버전이 출시되기 전에 상당 부분 변경될 수 있습니다. 이 문서에 포함된 정보는 문서 발행 시에 논의된 문제들에 대한 Microsoft Corporation의 당시 관점을 나타냅니다. Microsoft는 변화하는 시장 상황에 부응해야 하므로 이를 Microsoft 측의 공약으로 해석해서는 안되며 발행일 이후 소개된 어떠한 정보에 대해서도 Microsoft는 그 정확성을 보증하지 않습니다.

이 문서는 오직 정보를 제공하기 위한 것입니다. Microsoft는 이 설명서에서 어떠한 명시적이거나 묵시적인 보증도 하지 않습니다. 해당 저작권법을 준수하는 것은 사용자의 책임입니다. 저작권에서의 권리와는 별도로, 이 설명서의 어떠한 부분도 Microsoft의 명시적인 서면 승인 없이 어떠한 형식이나 수단(전기적, 기계적, 복사기에 의한 복사, 디스크 복사 또는 다른 방법) 또는 목적으로도 복제되거나, 검색 시스템에 저장 또는 도입되거나, 전송될 수 없습니다.

Microsoft가 이 설명서 본안에 관련된 특허권, 상표권, 저작권 또는 기타 지적 재산권 등을 보유할 수도 있습니다. 서면 사용권 계약에 따라 Microsoft로부터 귀하에게 명시적으로 제공된 권리 이외에, 이 설명서의 제공은 귀하에게 이러한 특허권, 상표권, 저작권 또는 기타 지적 재산권 등에 대한 어떠한 사용권도 허여하지 않습니다.

© 2005 Microsoft Corporation. 전원 보유.

Microsoft, Visual Basic, Visual C# 및 Visual Studio는 미국 및 기타 국가에서 마이크로소프트의 상표이거나 등록 상표입니다. 여기에 인용된 실제 회사와 제품 이름은 해당 소유자의 상표일 수 있습니다.

Contents

들어가는 글	2
실제 시나리오	2
데이터 원본	2
데이터 사용	2
데이터 통합 요건	2
데이터 통합의 과제	3
기술적 문제	3
조직적 문제	5
경제적 문제	5
SQL Server 2005 Integration Services	6
SSIS 아키텍처	6
작업 흐름과 데이터 흐름 엔진	6
파이프라인 아키텍처	6
통합 시나리오	7
기존 DW 로딩을 위한 SSIS	7
SSIS 및 데이터 품질	9
기존 ETL을 능가하는 SSIS 응용 프로그램	10
통합 플랫폼인 SSIS	12
통합된 개발 플랫폼	12
프로그래밍 용이성	14
스크립팅	14
데이터 통합의 실현	16

들어가는 글

기업 데이터를 의미 있는 유용한 정보로 탈바꿈시킬 수 있는 능력은 오늘날과 같은 기업 환경에서 경쟁 우위를 확보하는 데 필수적인 요소입니다. 전 세계 기업들의 IT 부서들에게 있어 급격히 증가하는 데이터를 효과적으로 활용하여 과거를 더욱 완벽하게 이해하고 향후 나아갈 방향을 확립한다는 목표는 가장 어려운 모험이 되어 왔습니다. 현재 데이터 통합과 관련한 문제들은 다음과 같이 크게 3가지로 구분할 수 있습니다.

- 기술적 문제
- 조직적 문제
- 경제적 문제

본 백서에서는 이러한 문제를 자세히 짚어 보고 Microsoft® SSIS(SQL Server™ 2005 Integration Services)를 통해 이들 문제를 해결할 수 있는 방법에 대해 논의할 것입니다. 먼저 이러한 문제들을 실제 시나리오에서 검토해 보겠습니다.

실제 시나리오

한 주요 다국적 운송 회사에서는 자체 데이터웨어하우스를 사용하여 운영 실적을 분석하고 예약된 배송의 변수를 예측하고 있습니다.

데이터 원본

이 회사의 주요 데이터 원본은 DB2 기반 주문 입력 시스템의 주문 데이터, SQL Server 기반 CRM(Customer Relationship Management) 시스템의 고객 데이터 그리고 Oracle 기반 ERP 시스템의 벤더 데이터입니다. 주요 시스템의 데이터 외에도 선적 감독관이 수기로 입력하는 "특수" 이벤트를 추적하는 스프레드시트의 데이터가 데이터웨어하우스에 통합됩니다. 현재 기상 정보, 교통 상태 및 벤더 상세 정보(하청업체의 배송 등)와 같은 외부 데이터도 다양한 소스에서 입수되는 텍스트 파일로 약간의 시차를 두고 데이터웨어하우스로 포함됩니다.

데이터 사용

이들 데이터가 여러 다양한 이기종 환경에서 작성되었을 뿐만 아니라 그 요구 사항과 지리적 위치도 매우 다양합니다. 이러한 다양성이 로컬 시스템의 확산을 주도했다고 해도 과언이 아닙니다. IT 부서에 대한 주요 영향 중 하나는 적어도 고객 데이터에 대해서만은 "1개 버전의 사실 데이터"를 구축할 수 있다는 것입니다.

데이터 통합 요건

이처럼 데이터의 다양성, 비즈니스 요구 사항 그리고 사용자 요구 사항 등의 측면에서 IT 부서는 다음과 같은 데이터 통합 요구 사항을 정의했습니다.

- 다양한 내, 외부의 소스에서 통합된 이전 및 최신 데이터를 안정적이며 일관되게 제공해야 한다.
- 보다 신속하게 데이터를 입수할 수 있도록, 제공업체와 벤더의 데이터를 웹 서비스나 FTP 등과 같은 여타 직접적인 매커니즘을 통해 이용할 수 있도록 해야 한다.
- 중복 데이터를 정리 및 삭제해야 하며 여타 다른 방법으로 데이터 품질을 보장해야 한다.
- 전 세계적으로 규제 요구가 증가하는 추세에 따라 기업은 명확한 감사 추적 기록을 유지해야 한다. 안정적인 데이터를 유지하는 것만으로 충분치 않으며 데이터를 추적 및 인증해야 한다.



데이터 통합의 과제

외견상, 실제 시나리오에서 데이터 통합 문제는 매우 단순합니다. 여러 소스에서 데이터를 입수하여 데이터를 정리 및 전환한 다음, 적절한 데이터 스토어로 로딩하여 분석 및 보고용으로 사용할 수 있습니다. 불행하게도 일반적인 데이터웨어하우스나 비즈니스 인텔리전스 프로젝트의 경우 기업은 데이터 통합 단계에서 가용 자원의 60-80%를 소모하고 있습니다. 왜 이러한 어려움이 발생하는 것일까요?

기술적 문제

기술적 문제들은 주로 소스 시스템에서 시작됩니다. 오늘날 트랜잭션(소비자가 입수, 구매 또는 확보) 데이터 수집에서 트랜잭션 이전(웹 클릭이나 RFID와 같은 메커니즘을 통해 고객 의도 추적) 데이터 수집으로 옮겨가는 추세가 확대되고 있습니다. 데이터는 현재 데이터베이스 및 텍스트 파일 같은 기존 소스와 형식으로 입수될 뿐 아니라 점차 다양한 형식(전용 파일에서 Microsoft Office 문서, XML 기반 파일 등을 망라)으로 이용되고 있으며, 웹 서비스와 RSS(Really Simple Syndication) 스트림 같은 인터넷 기반 소스에서 제공되기도 합니다. 가장 대표적인 문제는 다음과 같습니다.

- 다양한 형식의 여러 소스
- 구조적, 반구조적 및 비구조적 데이터
- 서로 다른 시간에 도착하는 소스 시스템에서 데이터 공급
- 대용량 데이터

이론적으로는 어떤 방법에 의해서건 필요한 모든 데이터를 한 곳에 보관하게 되는 경우에도 다음과 같은 새로운 문제가 제기됩니다.

- 데이터 품질
- 다양한 데이터 형식 인식
- 데이터를 비즈니스 분석에 의미 있는 형식으로 전환

다행스럽게도 필요한 모든 데이터를 입수하여 정리하고 유용한 형식으로 변환 및 대응할 수 있다고 가정해 봅시다. 여전히 기존 데이터 이동 및 통합과는 다른 변화가 이루어지게 됩니다. 즉, 고정된 장기 배치 프로세스에서 유동적인 단기 주문형 프로세스로 전환되는 것입니다. 배치 지향 프로세스는 주로 사용자가 시스템에 과중한 작업을 요청하지 않는 "시스템 정지 시간" 중에 수행됩니다. 이는 주로 미리 정의된 야간의 6-8 시간 동안 이루어지는 배치 윈도우이며 이 시간에는 아무도 근무하지 않는 것으로 가정합니다. 모든 규모와 유형의 기업들이 다국적으로 범위를 확장하고 있기 때문에 이는 더 이상 현실적이지 않습니다. 지금은 전혀 시스템이 중지되지 않거나 아주 잠시 동안의 시스템 중지 시간만 있을 뿐이며 세계 어딘가에서 누군가는 야간에도 업무를 처리하고 있습니다. 다국적 기업들의 경우, 해가 지지 않습니다. 따라서

- 가능한 빨리 데이터를 로딩해야 하는 부담이 늘어나고,
- 동시에 여러 대상을 로딩해야 하며
- 대상이 더욱 광범위하게 확대됩니다.

결국 이 모든 것을 반드시 해야 하는 것은 물론, 그것도 가능한 빨리 수행해야 합니다. 온라인 비즈니스와 같이 극단적인 경우에는 데이터를 지속적으로 통합해야 합니다. 따라서 진정한 배치 윈도우는 존재하지 않으며 대기 시간도 수 분을 초과해선 안됩니다. 이러한 시나리오의 대다수에서 의사 결정 프로세스는 지속적으로 실행되는 소프트웨어를 통해 자동화됩니다.

확장성과 성능은 결코 시스템 중지를 허용할 수 없는 비즈니스 요구에 직면했을 때 그 중요성이 더욱 높아집니다.

올바른 기술을 선택하지 않은 경우 시스템은 데이터웨어하우징 및 통합 프로세스의 거의 모든 단계를 스테이징해야 합니다. 이기중(특히 비표준) 데이터 원본이 ETL(Extract, Transform, and Load) 프로세스에 포함되어야 하고 데이터에 더욱 복잡한 작업(데이터 및 텍스트 마이닝과 같은)이 수행되어야 하는 경우 데이터 스테이징에 대한 요구는 더욱 증가합니다. 그림 1에 제시된 바와 같이 스테이징 요구가 증가하면서 "루프를 종료"하는 데(예: 분석하고 새 데이터에 조치를 취하는 데) 소요되는 시간도 증가합니다. 이러한 기존 ETL 아키텍처(로딩에 앞서 발생하는 부가 가치 ETL 프로세스와 정반대)는 시스템이 새로운 비즈니스 요구에 대처할 수 있는 능력을 심각하게 저하시키게 됩니다.

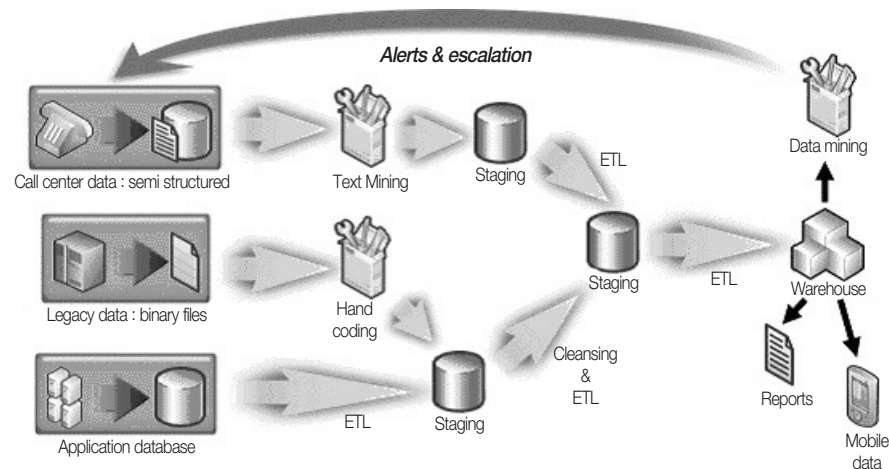


그림 1

마지막으로 데이터 통합이 전체 조직의 통합 아키텍처에 어떻게 연결되는지의 문제는 전사적인 비즈니스 문제를 해결하기 위해 응용 프로그램 통합을 위한 실시간 트랜잭션 기술과 배치 지향적인 대용량 데이터 통합 기술이 모두 요구되는 경우에 그 중요성이 더욱 높아집니다.



조직적 문제

대규모 조직에서는 데이터 통합과 관련하여 크게 2가지 문제에 직면하고 있습니다. 그것은 바로 "파워"와 "안전지대(comfort zone)"입니다.

파워 문제 : 데이터는 파워이며 일반적으로 모든 사람들이 데이터를 회사의 매우 귀중한 공유 자산으로 인식하게 하는 것은 상당히 어렵습니다. 전사적인 차원의 데이터 통합을 성공적으로 수행하기 위해서 다양한 데이터 원본 소유자들이 모두 프로젝트의 목적과 방향을 진심으로 이해하고 받아들여야 합니다. 관련 당사자들의 협력 없이는 데이터 통합 프로젝트를 성공적으로 진행할 수 없습니다. 임원진의 적극적인 지원, 합의 도출 및 여러 이해 당사자들과의 강력한 데이터 통합 팀 구축이 이들 문제를 해결할 수 있는 중요한 성공 요인이라 할 것입니다.

안전지대(Comfort Zone)의 문제 : 데이터 통합의 문제는 각각의 개별적인 요구의 맥락에서 분석하는 경우, 다양한 방법으로 해결될 수 있습니다. 데이터 통합의 약 60%는 수작업 코딩을 통해 해결됩니다. 이와 유사한 문제를 해결하는 데 사용되는 기술은 복제, ETL, SQL, EAI 등 매우 다양합니다. 사람들은 친숙한 기술에 더 비중을 두게 마련입니다. 이러한 접근법에는 중복된 기능들이 존재하며 개별 사례에서 해당 작업을 처리할 수 있다고 하더라도 이러한 기술은 서로 다른 유형의 문제들을 해결하도록 최적화되어 있습니다. 전사적 데이터 통합 문제를 해결하려고 하는 경우, 올바른 기술과 함께 탄탄한 아키텍처가 없다면 성공을 거두지 못할 것입니다.

경제적 문제

앞서 개략적으로 설명한 조직 및 기술 관련 문제들로 인해 데이터 통합은 데이터웨어하우스/비즈니스 인텔리전스 프로젝트에서 가장 높은 비용 부담을 발생시키고 있습니다. 데이터 통합의 비용을 높이는 주요 요인은 다음과 같습니다.

- 데이터 통합에 필요한 형식으로 데이터를 추출하는 것은 조직의 파워 게임 때문에 느리고 번거로운 프로세스가 될 것입니다.
- 여러 소스의 데이터를 정리하여 하나의 의미 있는 일관된 형식에 대응한다는 것은 특히 어려운 일입니다.
- 종종 표준 데이터 통합 도구는 프로젝트의 데이터 변환 요건을 충분히 충족할 수 있는 기능이나 확장성을 제공하지 않습니다. 따라서 해당 작업을 완료하기 위해 특별한 ETL 코드를 개발하는 데 막대한 컨설팅 비용을 지출하기에 이릅니다.
- 기업 내 여러 부서들이 별도로 관리하는 데이터를 통합하는 문제에 주안점을 두고 있습니다. 이들을 모두 통합해야 한다면 이러한 작업을 전사적 데이터 통합 아키텍처로 통합하는 데 따른 추가 비용이 발생합니다.

기업의 데이터웨어하우징 및 비즈니스 인텔리전스 요구가 점차 증가함에 따라 문제가 있는 데이터 통합 아키텍처의 경우, 유지 관리의 어려움이 더욱 심화되며 총 소유비용도 크게 증가하게 됩니다.

SQL Server 2005 Integration Services

표준 데이터 원본을 ETL 데이터로 통합하는 방식이 계속해서 대부분 데이터웨어하우스의 주류를 이루고 있습니다. 그러나 더욱 다양한 데이터 원본, 규제 요건 그리고 전 세계적인 운영과 온라인 운영을 포함해야 한다는 요구가 기존 데이터 통합 요건을 빠르게 변화시키고 있습니다. 이처럼 빠르게 성장하고 변화하는 상황에서 데이터의 가치를 최대한 도출하여 활용할 수 있도록 해야 한다는 요구가 그 어느 때보다 높아지고 있습니다. 효과적인 데이터 통합은 효과적인 의사 결정을 위한 기본 토대입니다. SQL Server Integration Services는 오늘날 기업 환경에서 효과적으로 데이터를 통합할 수 있도록 빠르고 유연하며 확장성이 뛰어난 아키텍처를 제공합니다.

본 백서에서는 SQL Server Integration Services(SSIS)가 어떻게 ETL 운영에 대한 기존 요구는 물론, 범용 데이터 통합의 변화하는 요구를 효과적으로 처리하는 도구로서 활용될 수 있는지에 대해 검토합니다. 또한 SSIS가 여타 주요 ETL 공급업체들이 제공하는 도구 및 솔루션과 어떤 측면에서 근본적인 차이가 있으며 이러한 차별성이 전 세계 모든 규모 기업들의 변화하는 요구를 어떻게 충족하는지에 대해서도 논의합니다.

SSIS 아키텍처

작업 흐름과 데이터 흐름 엔진

SSIS는 작업 중심적인 작업 흐름 엔진과 신속하고 확장 가능한 데이터 흐름 엔진으로 구성됩니다. 데이터 흐름은 전반적인 작업 흐름의 맥락에서 존재하며 작업 흐름 엔진은 데이터 흐름 엔진에 런타임 자원과 운영 지원을 제공합니다. 이러한 작업 흐름과 데이터 흐름의 결합으로 SSIS는 기존 ETL이나 데이터웨어하우스 시나리오에서는 물론, 데이터 센터 운영과 같은 많은 다른 확장 시나리오에서도 효과적입니다. 본 백서에서는 주로 데이터 흐름 관련 시나리오에 초점을 두게 될 것입니다. 데이터 센터 중심적인 워크플로에 SSIS를 사용하는 것은 별도의 주제로 다루어져야 합니다.

파이프라인 아키텍처

SSIS의 코어에는 데이터 변환 파이프라인이 존재합니다. 이 파이프라인은 데이터가 메모리로 로딩되면 해당 데이터의 행 세트를 초고속으로 처리하는 버퍼 중심의 아키텍처를 보유하고 있습니다. 이 접근 방법은 특정 변환이나 작업상의 요구 사항 또는 실제 하드웨어가 장애 요인이 되는 경우에도 데이터 스테이징 없이 단일 작업으로 ETL 프로세스의 모든 데이터 변환 단계를 수행하는 것입니다. 그럼에도 불구하고 최대 성능을 실현하기 위해 이 아키텍처에서는 스테이징 작업을 피하고 있으며 메모리에서 데이터를 복사하는 작업도 가능한 하지 않습니다. 이는 데이터웨어하우스와 통합 프로세스의 거의 모든 단계에서 스테이징 작업을 거쳤던 기존 ETL 도구와 대조적입니다. 스테이징 없이 데이터를 조작할 수 있는 능력은 기존 관계형 파일과 플랫폼 파일 데이터를 넘어 이전 ETL 변환 기능 이상으로 확장됩니다. SSIS를 통해 모든 유형의 데이터(구조적, 비구조적, XML 등)는 버퍼로 로딩되기 전에 테이블(열과 행) 구조로 변환됩니다. 테이블 데이터에 적용할 수 있는 모든 데이터 작업은 데이터 흐름 파이프라인의 모든 단계에서 데이터에 적용할 수 있습니다. 이는 단일 데이터 흐름 파이프라인이 다양한 데이터 원본을 통합하고 데이터 스테이징 과정 없이 이러한 데이터에 복잡한 작업을 수행할 수 있다는 것을 의미합니다.

또한 업무 또는 운영상의 이유로 스테이징 과정이 필요한 경우에도 SSIS는 이를 훌륭히 지원합니다.

이러한 아키텍처 덕분에 SSIS는 기존 데이터웨어하우스 중심적인 ETL에서 새로운 정보 통합 기술에 이르기까지 다양한 데이터 통합 시나리오에서 사용될 수 있습니다.

통합 시나리오

기존 DW 로딩을 위한 SSIS

기본적으로 SSIS는 완벽한 기능을 갖춘 포괄적인 ETL 도구입니다. 그 기능, 확장성 및 성능은 시장의 하이엔드 경쟁업체들이 제공하는 최고의 제품들에 필적할 수 있습니다. 또한 데이터 통합 파이프라인 아키텍처를 통해 다수의 동시 소스에서 데이터를 가져와 여러 복잡한 변환 작업을 수행한 다음, 동시에 여러 대상으로 데이터를 전송할 수 있습니다. 이 아키텍처를 통해 SSIS는 대형 데이터베이스뿐 아니라 복잡한 데이터 흐름에도 유용하게 사용될 수 있습니다. 데이터가 소스에서 대상으로 전송될 때 데이터 흐름을 분리하거나 병합하고 다른 데이터 흐름과 결합하는 것은 물론, 조작할 수 있습니다. 그림 2에서는 이러한 흐름의 예를 보여 줍니다.

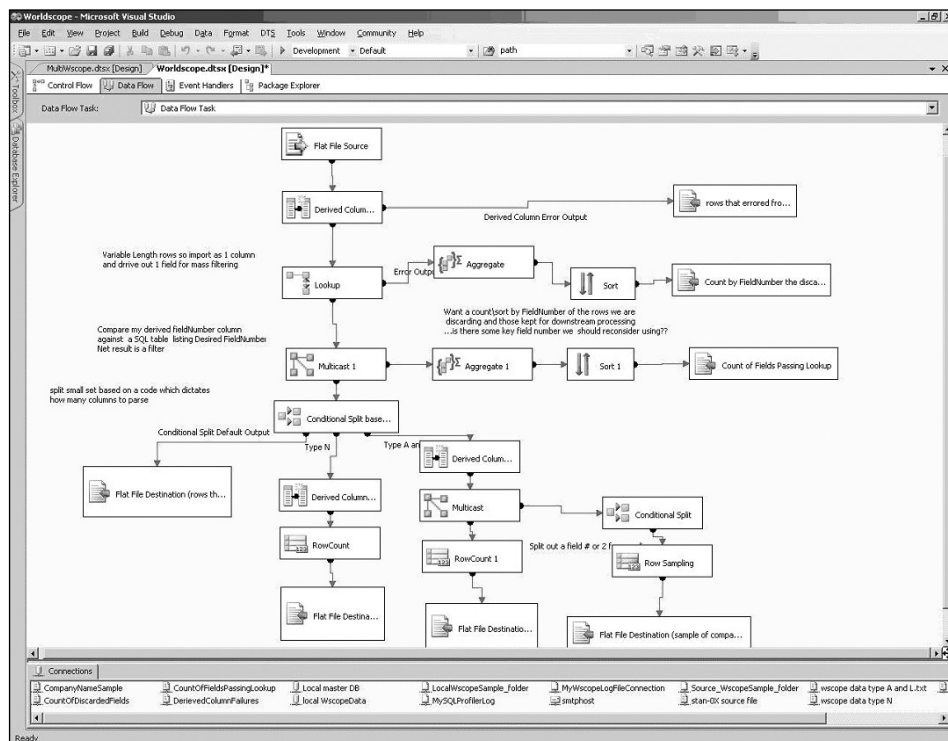


그림 2

SSIS는 어댑터라는 특수 구성 요소 세트를 사용하여 OLE DB, 관리되는 (ADO.NET), ODBC, 플랫폼 파일, Excel 및 XML을 비롯한 다양한 소스 간에 데이터를 주고 받을 수 있습니다. SSIS는 맞춤형 데이터 어댑터(자체적으로 또는 타사에서 개발한)에서 데이터를 사용할 수도 있습니다. 따라서 SSIS 데이터 흐름에서 원활하게 사용될 수 있도록 데이터 원본으로 레거시 데이터 로딩 로직을 래핑할 수 있습니다. SSIS에는 데이터웨어하우스 구축에 필수적인 데이터 조작을 가능하게 하는 강력한 데이터 변환 구성요소 세트가 포함되어 있습니다. 이러한 변환 구성 요소는 다음과 같습니다.

- Aggregate 한 번의 전달로 여러 차례의 집계를 수행합니다.
- Sort 흐름 중의 데이터를 정렬합니다.
- Lookup 참조 데이터 세트에 유연한 캐싱된 조회 작업을 수행합니다.
- Pivot 및 UnPivot 명칭이 나타내는 바와 같이 2개의 별도 변환 작업을 수행합니다.
- Merge, Merge Join 및 Union All 조인과 결합 작업을 수행할 수 있습니다.
- Derived Column 문자열, 숫자, 일/시 등의 작업 및 코드 페이지 변환과 같은 열 수준의 작업을 수행합니다. 실제로 여타 공급업체들이 별개의 여러 변환 기능으로 분리해 놓은 모든 기능을 이 한 가지 구성 요소가 모두 포함하고 있습니다.
- Data Conversion 다양한 형식(숫자, 문자열 등) 간에 데이터를 전환합니다.
- Audit 선형 메타데이터가 있는 열과 다른 작업 감사 데이터를 추가합니다.

이러한 코어 데이터웨어하우징 변환뿐 아니라 SSIS는 SCD(Slowly Changing Dimensions)와 같은 고급 데이터웨어하우징에 대한 지원을 포함하고 있습니다. SSIS의 SCD 마법사는 사용자가 입력한 내용을 기초로 SCD를 관리할 수 있도록 요구 사항을 지정하는 과정을 안내하고 여러 변환 과정을 통해 전체 데이터 흐름을 생성하여 SCD 로딩을 실행합니다. 2개의 새로운 SCD 형식과 함께 표준 Type 1과 2 SCD에 대한 지원(고정 속성 및 유추된 구성원)이 제공됩니다. 그림 3은 SCD 마법사의 한 페이지를 보여 줍니다.

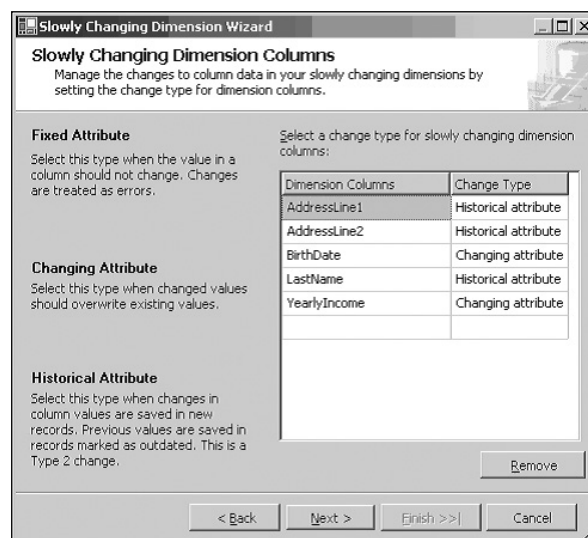


그림 3

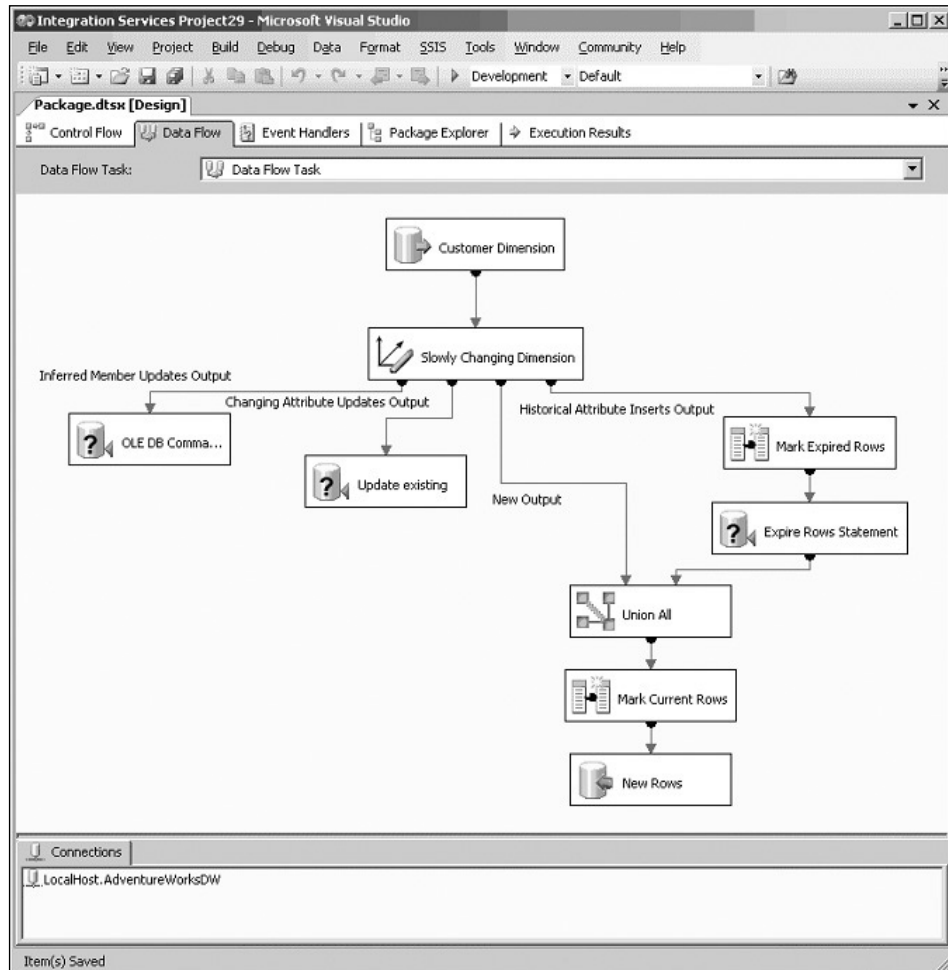


그림4 SCD 마법사로 생성된 데이터 플로우

또한 SSIS는 Analysis Services MOLAP(multidimensional OLAP) 캐시를 데이터 흐름 파이프라인에서 직접 로딩하는 데 사용할 수 있습니다. 이는 SSIS가 관계형 데이터웨어하우스를 생성하는 것은 물론, 분석 응용 프로그램의 다차원 큐브를 로딩하는 데에도 사용된다는 것을 의미합니다.

SSIS 및 데이터 품질

SSIS의 주요 특징은 데이터를 통합할 수 있는 것은 물론, 데이터를 조작하기 위한 다른 기술들을 통합할 수 있다는 것입니다. 이에 따라 SSIS는 첨단 "퍼지 로직(fuzzy logic)" 기반 데이터 정리 구성 요소를 포함할 수 있습니다. Microsoft Research 연구소에서 개발된 이들 구성 요소들은 이 분야에 대한 최신 연구 성과를 보여주고 있습니다. 그 접근 방법은 도메인 독립적이며 주소/zip 참조 데이터와 같은 특정 도메인 데이터를 전혀 활용하지 않습니다. 따라서 이들 변환은 단순히 데이터를 처리하는 것 뿐만 아니라 대다수 형식의 데이터를 정리하는 데에도 사용할 수 있습니다.

SSIS는 Analysis Services의 데이터 마이닝 기능과 긴밀히 통합되어 있습니다. 데이터 마이닝은 데이터 세트에서 패턴을 추출하여 마이닝 모델로 캡슐화합니다. 무엇보다도 중요한 것은 이 마이닝 모델은 어떤 데이터가 데이터세트에 속해 있는지, 어떤 데이터가 비정상인지를 예측하는 데 사용할 수 있기 때문에 데이터 품질 구현을 위한 수단으로 이용할 수 있습니다.

SSIS에서는 복잡한 데이터 라우팅을 지원하기 때문에 비정상 데이터를 확인하여 자동으로 수정하고 더 나은 값으로 대체합니다. 이에 따라 "페루프" 정리 시나리오를 수행할 수 있습니다. 그림 5는 이러한 페루프 정리 데이터 흐름의 예를 보여 줍니다.

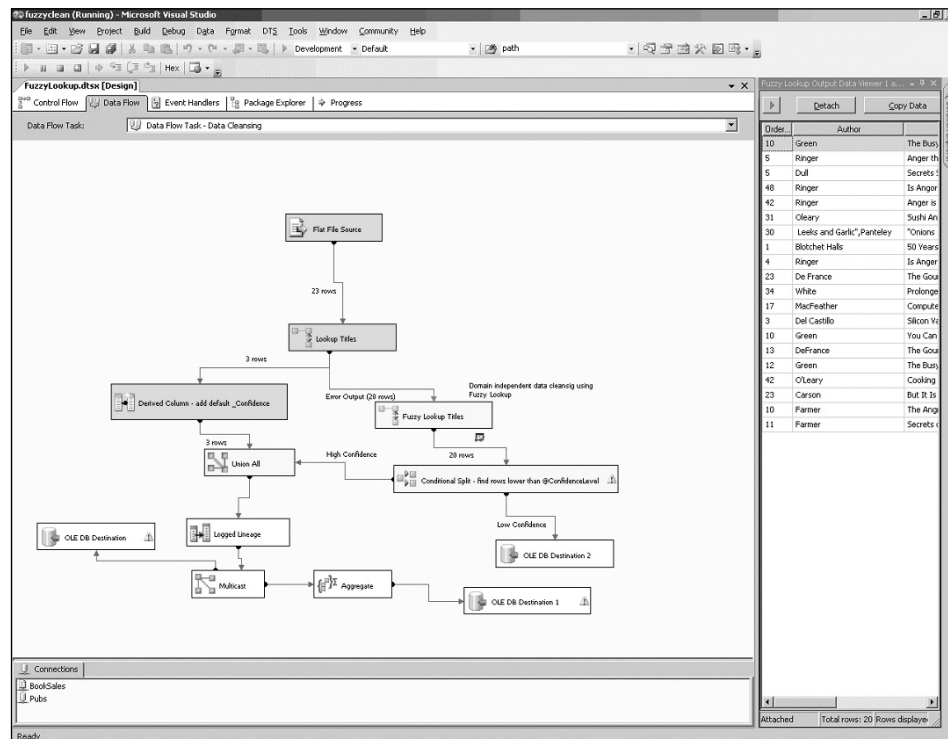


그림 5

SSIS는 기본 제공되는 데이터 품질 기능 이외에도 타사 데이터 정리 솔루션과 긴밀히 연동되도록 확장할 수 있습니다.

기존 ETL을 능가하는 SSIS 응용 프로그램

거의 모든 종류의 데이터를 처리할 수 있는 데이터 흐름 파이프라인, Analysis Services와의 긴밀한 통합, 다양한 데이터 작업 기술과의 연동 지원 및 풍부한 작업 흐름 엔진 포함 등을 강점으로 갖춘 SSIS는 ETL과 같이 이전 방식을 고수하는 많은 시나리오에서 사용될 수 있습니다.

서비스 지향 아키텍처

SSIS는 데이터 흐름 파이프라인에서 디스크상의 파일은 물론 HTTP상의 URL에서 입수되는 데이터를 모두 포함하는 XML 데이터 소싱을 지원합니다. XML 데이터는 테이블 형식의 데이터로 "분할"되어 데이터 흐름에서 쉽게 조작할 수 있습니다. 이러한 XML에 대한 지원은 웹 서비스에 대한 지원과 연동할 수 있으며 SSIS는 제어 흐름에서 웹 서비스와 상호 작용하여 XML 데이터를 포착할 수 있습니다.

또한 XML은 파일에서, MSMQ(Microsoft Message Queuing)에서 그리고 HTTP를 통해 웹상에서 포착될 수 있습니다. SSIS는 XSLT, XPATH, diff/merge 등을 통해 XML을 조작할 수 있으며 XML을 데이터 흐름으로 스트리밍할 수도 있습니다.

이러한 지원을 통해 SSIS는 유연한 SOA(Service Oriented Architectures)에 포함될 수 있습니다.

데이터 및 텍스트 마이닝

SSIS는 Analysis Services의 데이터 마이닝 기능과 긴밀히 통합되어 있을 뿐 아니라 텍스트 마이닝 구성 요소도 포함하고 있습니다. 텍스트 마이닝(텍스트 분류라고도 함)은 비즈니스 범주와 텍스트 데이터(단어와 구) 간의 관계를 식별하는 것입니다. 텍스트 마이닝을 통해 텍스트 데이터의 주요 용어들을 발견할 수 있으며 이를 기초로 "관심" 텍스트를 자동으로 식별할 수 있습니다. 그 다음 "페루프" 작업을 실행해 고객 만족도 향상 및 제품 및 서비스 품질 향상과 같은 비즈니스 목표를 달성할 수 있습니다.

주문형 데이터 원본

SSIS의 가장 독보적인 특징 중 하나는 DataReader 대상으로서, ADO.NET DataReader로 데이터를 전달하는 것입니다. 이 구성 요소가 데이터 흐름 파이프라인에 포함되는 경우, DataReader 대상을 포함한 패키지는 데이터 원본으로 사용되고 ADO.NET DataReader 그 자체로 제시됩니다. 이에 따라 SSIS는 데이터웨어하우스를 로딩하는 기존 ETL로 사용될 수 있을 뿐 아니라 필요 시 여러 소스에서 통합, 조정 및 정리된 데이터를 제공할 수 있는 데이터 원본의 역할을 수행합니다. 예를 들어 이는 SSIS 패키지를 자체 데이터 원본으로 사용하는 여러 데이터 원본에서 가져온 데이터를 Reporting Services가 사용하도록 할 수 있습니다.

이 모든 것이 통합된 가능한 시나리오는 정규 보고서의 일부인 RSS 피드에서 관심 있는 자료를 식별하여 제공하는 데서 찾을 수 있습니다. 그림 6은 인터넷을 통해 RSS 피드에서 데이터를 가져오고 웹 서비스의 데이터와 통합하며 RSS 피드에서 관심 있는 자료를 찾기 위해 텍스트 마이닝을 수행한 다음, 관심 자료를 DataReader 대상으로 보내 Reporting Services 보고서에서 사용할 수 있도록 하는 SSIS 패키지를 보여 줍니다.

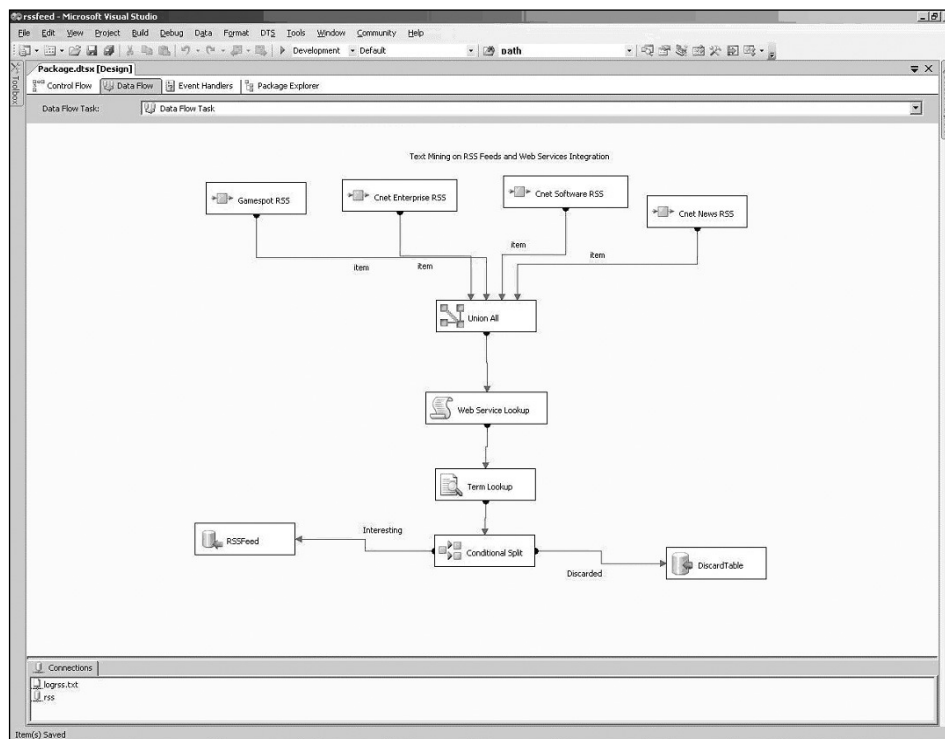


그림6

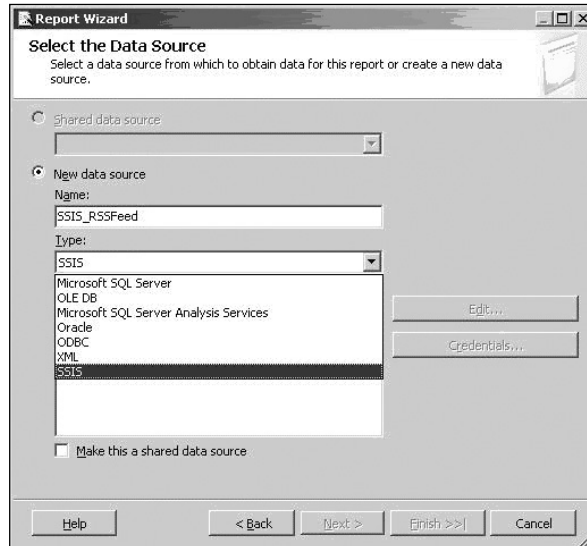


그림 7 Report 마법사의 데이터 원본으로서 SSIS 패키지를 사용하는 것을 보여 주고 있습니다.

ETL 도구 측면에서 이 시나리오는 실제 데이터 추출이나 변환 또는 로딩이 발생하지 않기 때문에 매우 특별한 경우라고 할 수 있습니다.

통합 플랫폼인 SSIS

SSIS는 과거와는 다른 새로운 시나리오를 지원하는 것은 물론, 데이터 통합을 위한 진정한 플랫폼이라는 점에서 ETL 도구를 능가합니다. SSIS는 완벽한 BI 응용 프로그램을 개발할 수 있도록 지원하는 SQL Server BI(Business Intelligence) 플랫폼의 일부입니다.

통합된 개발 플랫폼

SQL Server Integration Services, Analysis Services 및 Reporting Services는 모두 SQL Server BI(Business Intelligence) Development Studio라는 Visual Studio® 기반의 공통 개발 환경을 사용합니다. BI Development Studio는 BI 응용 프로그램 개발에 IDE(Integrated Development Environment)를 제공합니다. 이 공유 인프라는 다양한 개발 프로젝트(통합, 분석 및 보고) 간에 메타데이터 수준에서 통합이 이루어질 수 있도록 합니다. 이러한 공유 구조의 예로는 데이터 원본의 오프라인 스키마/뷰 정의인 DSV(Data Source View)를 들 수 있으며 이는 세 가지 BI 프로젝트 유형 모두에서 사용될 수 있습니다.

이 IDE는 "체크 인/체크 아웃"과 같은 팀 기반 기능에 대한 지원과 더불어 버전 제어 소프트웨어(예: VSS)와의 통합과 같은 기능을 제공하기 때문에 비즈니스 인텔리전스 응용 프로그램을 위한 전사적 차원으로 구성된 팀 개발 환경의 요구를 충족할 수 있습니다. 그림 8은 Integration, Analysis 및 Reporting 프로젝트로 구성된 BI Development Studio 솔루션을 보여 줍니다.

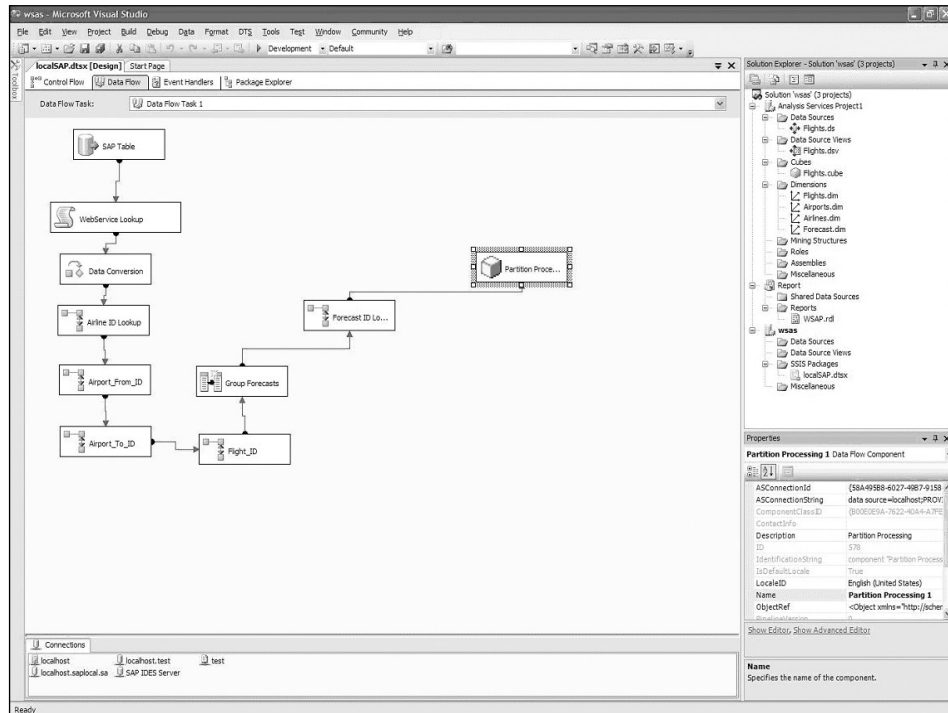


그림8

이는 BI 응용 프로그램 개발을 위한 단일 공간을 제공하는 것은 물론, 다른 Visual Studio 프로젝트(Visual C#®, Visual Basic®, .NET 등 사용)를 개발하는 데에도 사용할 수 있다는 점에서 개발자들에게 진정한 의미의 완벽하고 포괄적인 개발 환경을 제공할 수 있습니다.

BI Development Studio는 통합 BI 개발 환경 외에도 SSIS 패키지의 진정한 런타임 디버깅 기능을 갖추고 있습니다. 대표적으로 분기점 설정 및 watching 변수와 같은 표준 개발 구조 지원 등을 들 수 있습니다. 가장 독보적인 기능은 데이터 흐름 파이프라인에서 처리될 때 데이터의 행을 볼 수 있도록 하는 Data Viewer입니다. 이와 같은 데이터 시각화는 표준 텍스트 그리드, 분산 플롯 또는 막대 그래프 같은 그래픽 형태로 이루어질 것입니다. 동시에 여러 형식으로 데이터를 표시할 수 있도록 여러 뷰어들을 연결할 수도 있습니다. 그림 9는 분산 플롯과 텍스트 그리드를 사용하여 시각화된 그래픽 데이터의 예제를 보여 줍니다.

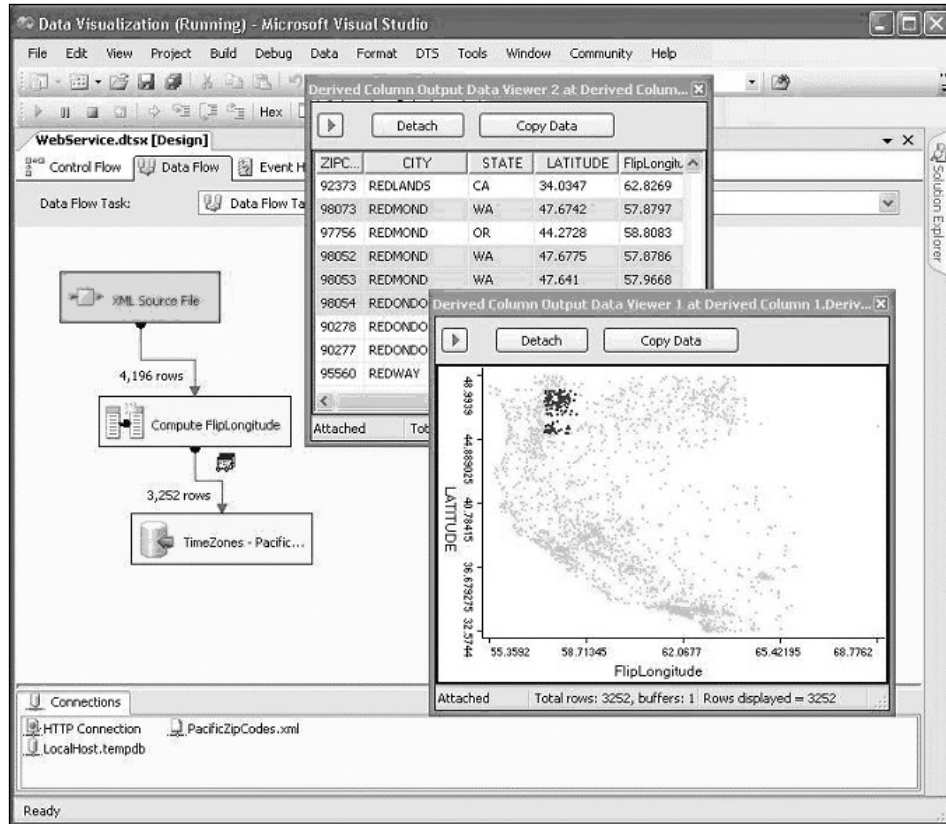


그림9

프로그래밍 용이성

SSIS는 전문 개발 환경을 제공할 뿐 아니라 풍부한 API 세트를 통해 모든 기능을 표시합니다. 이러한 API들은 관리되는 (.NET Framework)과 native (Win32) 모두이며, 이를 통해 개발자들은 .NET Framework에 의해 지원되는 모든 언어 (Visual C#, Visual Basic, .NET 등)와 C++로 맞춤형 구성 요소를 개발함으로써 SSIS의 기능을 확장할 수 있습니다. 이러한 맞춤형 구성 요소들은 작업 흐름 작업과 데이터 흐름 변환(소스 및 대상 어댑터 포함)이 될 수 있습니다. 이를 통해 레거시 데이터 및 기능을 SSIS 통합 프로세스로 쉽게 통합할 수 있기 때문에 레거시 기술에 대한 과거의 투자를 효과적으로 활용할 수 있습니다. 또한 타사 구성 요소도 쉽게 포함할 수 있습니다.

스크립팅

앞서 언급한 확장성에는 재사용 가능한 맞춤형 구성 요소뿐만 아니라 스크립트 기반 확장성까지 포함됩니다. SSIS는 작업 흐름 및 데이터 흐름 모두를 위한 스크립트 구성 요소를 보유하고 있습니다. 이를 통해 사용자들은 Visual Basic, .NET으로 스크립트를 작성하여 임시 기능 (소스 및 대상 데이터 포함)을 추가하고 .NET Framework 어셈블리로 패키징된 모든 기존 기능을 재사용할 수 있습니다.

그림 10은 데이터 흐름 내에서 데이터 행을 조작하는 스크립트의 예제를 보여 줍니다.

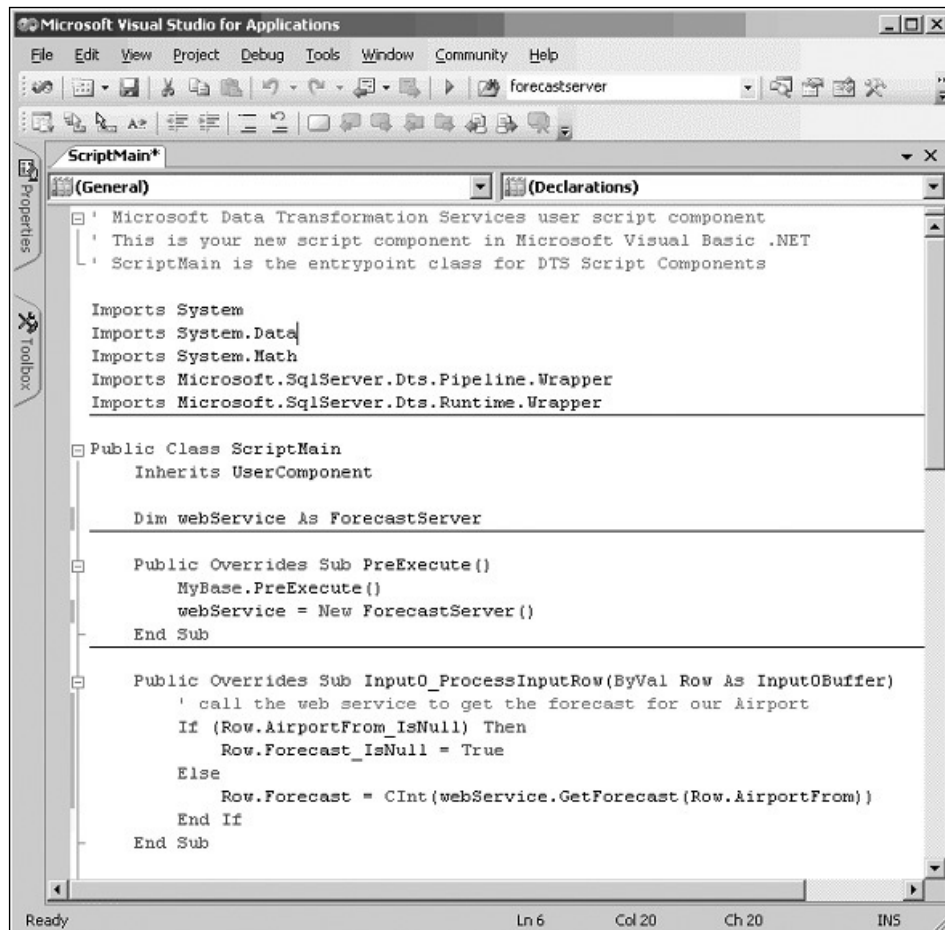


그림10

이 확장성 모델에 따라 SSIS는 데이터 통합 도구는 물론, 데이터 마이닝, 텍스트 마이닝 및 UDM 같은 기술이 쉽게 플러그인될 수 있는 통합 버스가 될 수 있습니다. 이를 통해 많은 임의의 데이터 조작 및 구조가 포함된 복잡한 통합 시나리오를 지원하게 됩니다.

데이터 통합의 실현

유연하고 확장 가능한 아키텍처를 통해 SSIS는 본 백서의 전반부에서 설명한 데이터 통합의 기술적 문제 중 대부분을 해결할 수 있습니다. 또한 그림 11에 제시된 바와 같이 SSIS는 불필요한 스테이징 작업을 없애거나 적어도 최소화할 수 있습니다. 단일 파이프라인 작업으로 복잡한 데이터 조작을 수행할 수 있기 때문에 이제 루프 종료와 적절한 조치를 수행하는 데 있어 의미 있는 시간 내에 매우 신속하게 데이터의 변경과 패턴에 대응할 수 있게 되었습니다. 이는 데이터 스테이징 작업이 필수적이며 루프 종료 및 데이터에 중요한 작업을 수행하는 데 실용적이지 못했던 기존 아키텍처와는 상반되는 것입니다.

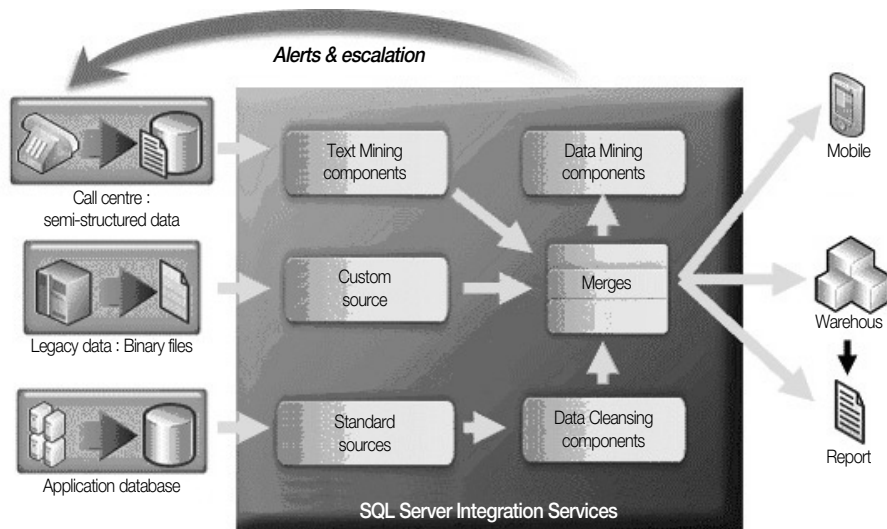


그림 11

SSIS의 뛰어난 확장성을 바탕으로 기업들은 재사용 가능한 SSIS 익스텐션에 맞춤형 코드를 포함시켜 로깅, 디버깅, BI 통합 등의 기능을 최대한 활용함으로써 데이터 통합을 위한 맞춤형 코드에 대한 기존 투자를 활용할 수 있습니다. 이는 본 백서의 전반부에 소개된 일부 조직적 문제들을 해결할 수 있습니다.

SQL Server 제품에 SSIS가 포함됨에 따라 여타 하이엔드 데이터 통합 도구와 비교할 때 비용 측면에서 정당성을 확고하게 입증할 수 있게 되었습니다. 즉, 초기 비용 구매가 낮아졌을 뿐 아니라 Visual Studio 및 나머지 SQL Server BI 도구와 긴밀하게 통합되어 응용 프로그램 개발 및 유지 관리 비용도 유사한 다른 도구에 비해 크게 절감할 수 있습니다. SSIS의 매우 합리적인 총소유 비용(TCO)(그리고 SQL Server의 나머지 요소)을 바탕으로 전사적 데이터 통합을 모든 시장 부문에서 현실화할 수 있게 되었으며, 최대 규모이자 가장 고가의 솔루션을 공급하는 기업들이 독점하는 영역에서 벗어날 수 있게 되었습니다. 이와 동시에 SSIS의 아키텍처는 최첨단 하드웨어를 활용하고 최고 수준의 고객 요건에 맞춰 성능과 규모를 제공할 수 있도록 향상되었습니다. SSIS는 최상위 대기업에서부터 중견, 중소기업에 이르는 모든 규모의 고객들이 탁월한 확장성을 바탕으로 방대한 데이터를 통합할 수 있도록 지원합니다. Microsoft 고객 지원 인프라인 SQL Server의 나머지 기능(광범위한 장기 베타 테스트에서부터 풍부한 온라인 통신, 프리미어 지원 계약에 이르기까지)과 여타 Microsoft 제품과의 일관성 및 통합에 힘입어 SSIS는 데이터 통합의 새로운 지평을 열 단연 독보적인 도구라고 할 수 있습니다.

Microsoft®

한국마이크로소프트(유)

서울특별시 강남구 대치동 892번지 포스코센터 | 전화 : 080-985-2000 | 인터넷 : www.microsoft.com/korea