한국SAS의

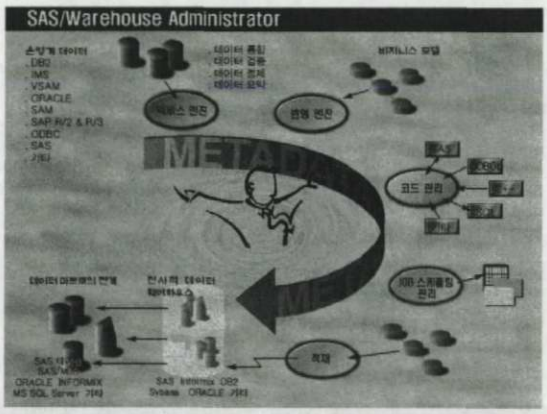
'End-to-End' 데이터 웨어하우징 솔루션

SAS는 데이터 웨어하우스 추출에서 저장, OLAP, DSS, 마이닝에 이르기까지의 전 과정에 대한 솔루션을 제공한다. 지난 20여년간 의사결정지원 분야에서 탁월한 기능으로 세계적으로 명성을 얻고 있는 SAS는 데이터 웨어하우징 분야에서도 Back-End와 Front-End에 있어서의 뛰어난 기능을 바탕으로 의사결정 지원을 위한 최적의 데이터 웨어하우징 솔루션을 제공하고 있다.

데이터 웨어하우스 추출 도구 'SAS/Warehouse Administrator'

데이터 웨어하우스 구축 과정에서 운영계 원천 데이터를 추출/변환 및 적재시키는 작업의 중요성이 새롭게 대두되고 있다. 그 이유는 이 기종에 존재하는 다양한 데이터 소스로부터 데이터를 추출하여 최적의 수행으로 데이터 웨어하우스에 축적시키기 위한 프로그램이 매우 복잡하며, 사용자 프로그램 방식의 경우 변경 데이터 관리가 까다롭다는 점, 그리고 추출/변환/축적 과정에 대한 관리, 곧 추출 과정에서의 메타 데이터 관리가 이루어질 수 없다는 점 등의 제반의 해결하기 어려운 문제점들 때문이다.

SAS/Warehouse Administrator와 SAS/CDC (Changed Data Capture)는 SAS의 데이터 웨어하우스 구축 전략인 SAS 'End-to-End' 데이터 웨어하우스 솔루션 제품들 중에서 추출/적재, 그리고 변경 데이터 관리를 담당하고 있는 제품이다. SAS/Warehouse Administrator(이하 SAS/WA)의 기능을 데이터 웨어하우스 구축 단계별로 살펴보면 다음과 같다.



〈그림 1〉 SAS/웨어하우스 추출도구

DW(Data Warehouse) Population

SAS/WA는 운영계 원천 데이터의 추출 작업을 GUI로 제공하는 제품으로서 내부적으로는 BASE, SAS/Connect, SAS/Access 제품군 등의 기능을 이용하는 소스 코드 생성기이다.

BASE와 SAS/Access 제품군을 내부적으로 이용 다양한 운영계 원천 데이터-DB2, IMS, VSAM, TAPE-들을 일관성있게 관리할 수 있다. 그리고 SAS/Connect 제품의 기능을 이용하여 운영계 시스템으로부터 데이터 웨어하우스 서버로 데이터를 고속 전송하여 데이터 전달의 신뢰성을 높였다(많은 변수가 있지만 평균적으로 TCP/IP의 경우 400k/sec, APPC의 경우 100k/sec를 지원한다).

또한 데이터 변환, OLAP/DSS를 위한 key code

의 재구성, 그리고 파생 데이터 생성 작업시 GUI에서 point&click 방식으로 작업이 가능하며 복잡한 애플리케이션 로직이 추가되는 경우 SAS 4GL(COBOL 프로그램의 자동 변환 기능 내장) 방식으로 로직의 추가가 가능하고 이러한 작업들의 기록은 모두 메타 데이터로 생성, 일목요연하게 관리됨으로써 데이터 웨어하우스의 유지, 보수가 손쉬워진다.

특히 APPC 프로그램을 할 때 개발자가 부담하여야 하는 통신 문제, 코드 변환(EBCIDIC 〈-〉 ASCII) 문제, 그리고 한글 데이터 처리 문제들을 BASE와 SAS/Connect 제품을 이용하여 해결해 줌으로써 DW 데이터 추출 작업시 개발의 편의성을 보장해준다.

〈표 1〉은 사용자들이 DW 추출 작업시 SAS/WA 제품을 활용하여 적용할 수 있는 항목들을 요약한 것이다.

변경 데이터 관리

데이터 웨어하우스 구축 과정 중 변경 데이터 관리는 매우 중요하다. 따라서 변경된 부분에 대한 효과적이고 빠른 검색을 위한 다양한 기법들이 있다. 크게 두 가지 기법이 존재하는데, 첫번째는 전체를 변경하는 방법이고, 두번째는 변경된 부분만을 단계적으로 반영하는 방법이다.

〈표 1〉 사용자들이 DW 추출 작업시 SAS/WA 제품을 활용하여 적용할 수 있는 항목들

과 정	기 능
Extraction	DB2/MVS Source 지원 IMS/MVS Source 지원 VSAM Source 지원 TAPE 지원 DBMS log를 이용한 변경 데이터 추출 Replication Engine(프로그램 불필요)
Transformation	conversion(값 전환) separation(값 분리) concatenation(값 합성) selection(열 선택) subset(행 선택) column 추가 Application logic 추가
Apply	다양한 Target DBMS 지원 Load Append Destructive Merge Constructive Merge Scheduling Job Stream 지원 기능 Error시 Retry 기능 시작 시간 설정 기능
기 타	즉시 적용 가능 여부

전체를 변경하는 방법은 데이터 웨어하우스를 기업의 새로운 비즈니스 사이클에 맞추어 새롭게 구축할 때 주로 쓰이며, 이 기법의 장점은 손쉽게 많은 양의 데이터를 관리할 수 있다는 것이다. 그러나 비용이 많이 들고 이력 데이터의 관리가 운영계 시스템의 운영 시간대로 한정될 수밖에 없는 단점을 가지고 있다.

변경된 부분을 단계적으로 반영하는 두번째 방법이 바로 SAS/CDC에서 제공하는 기법이다. SAS/CDC는 비동기식으로 운영되며, DB2,IMS의 log에 접근하여 insert, delete, update등의 정보를 SAS 파일 형태로 가공한 다음, SAS/WA에서 생성한 메타 데이터를 이용하여 복잡한 소스 테이블과 타겟 테이블간의 대응관계를 자동으로 인식한 후 타겟 테이블에 변경된 부분만을 반영시킨다. SAS/CDC를 이용하면 비교적 저렴하게 DB2와 IMS의 아카이벌 로그에 접근하여 변경 데이터을 관리할 수 있다.

메타 데이터 관리

데이터 웨어하우스는 사용자 중심의 의사결정 지원 및 다차원 분석(OLAP)을 위한 데이터 모델이다. 이러한 데이터 모델은 사용자의 요구에 따라 끊임없이 변화되며, 이 변화된 환경은 메타 데이터로 일관되게 관리할 수 있어야 한다.

SAS/WA가 자동으로 생성하는 메타 데이터는 데이터 소스, 타겟 테이블과 적재 주기, 타겟 테이블의 필드 속성, 데이터 소스와 타겟 테이블간의 대응 관계, 데이터 정제 작업시 기준 및 범위, 데이터 변형 작업에 대한 정보, 파생 데이터 생성시 계산식, 요약 테이블 생성에 관한 정보등 데이터 웨어하우스에 관한 제반 정보를 사용자에게 제공함으로써 효율적으로 데이터 웨어하우스를 구축/보수/유지할 수 있는 편이성을 가져다 준다.

SAS/WA에서 생성된 메터 데이터는 그 즉시 반

출이 가능하여 SAS/EIS 사용자들은 SAS/WA에서 생성된 메타 데이터를 그대로 이용하여 다차원 분석 작업에 적용시킬 수 있다.

DM(Data Mart) Population

데이터 마트는 데이터 웨어하우스의 데이터를 이용, OLAP Tool이나 DSS 프로그램들을 이용하여 데이터를 질의(query)하기 위한 저장소이다. 여타의 추출 도구들의 경우 DW, DM Population 기능이 통합되어 있지 않지만 SAS/WA 제품은 DW와 DM Population 기능이 제품 하나에 통합되어 있다. 따라서 추가 구매의 부담이 없으며 DW와 DM 구축의 전 과정을 일관성있게 관리할 수 있는 장점을 가지고 있다. 또한 제품 구성이 단순하기 때문에 간단하게 추출/변환 작업의 성능을 최대한으로 조정할 수 있다.

〈표 2〉는 사용자들이 DM 추출 작업시 SAS/WA를 적용할 수 있는 항목들을 요약한 것이다.

데이터 웨어하우스 구축에 있어서 흔히 간과하기 쉬운 요소이지만, 운영계 시스템상의 데이터를 데이터 웨어하우스로 정확하고, 시기 적절하게, 통합적으로 추출/적재시키는 작업은 기술적으로 매우 복잡한 과정이다.

SAS/WA는 이러한 복잡하고 정교한 작업을 GUI에서 포인트 앤 클릭 방식으로 구현하는 추출도구이며, 국내에서 유일하게 한글 데이터 처리 부분을 검증 받은 도구이다.

〈표 2〉 사용자들이 DM 추출 작업시 SAS/WA를 적용할 수 있는 항목들을 요약한 것

과 정	기 능
Extraction	다양한 DBMS Source 지원
	동기식/비동기식
	Replication Engine(프로그램 불필요)
Transformation	conversion(값 전환)
	separation(값 분리)
	concatenation(값 합성)
	selection(열 선택)
	subset(행 선택)
	column 추가
	Application logic 추가
Apply	다양한 Target DBMS 지원
	Load
	Append
	Destructive Merge
	Constructive Merge
Scheduling	Job Stream 지원 기능
	Error시 Retry 기능
	시작 시간 설정 기능
	기 타
	즉시 적용 가능 여부

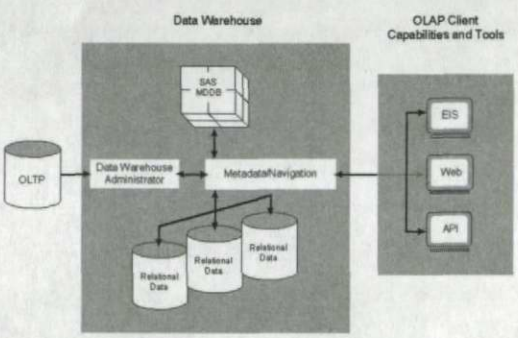
OLAP

OLAP의 정의

OLAP(On-Line Analytical Processing)이란 다양한 비지니스 관점에서 쉽고 빠르게 다차원적인 데이터에 접근하여 의사결정에 활용할 수 있는 정보를 얻을 수 있게 해주는 기술이다. 다차원(multidimensionality)이란 사람들이 일반적으로 생각하는 방식을 표현하는 것으로서 시간, 지역, 제품, 실적 등 비즈니스의 중요 관심사들이 각각의 차원을 이룬다.

Slicing & Dicing

다양한 차원으로 데이터를 검색하며, 간단한 조작으로 차원간 이동이 가능하여 사용자가 보고자하는 관점에서 손쉽게 정보를 살펴볼 수 있게 해주는 기능이다. SAS 시스템은 다차원 데이터 모델에 포함될 수 있는 차원의 수가 무제한적이다.



〈그림 2〉 SAS OLAP 솔루션 아키텍처

OLAP에 대한 두가지 접근방법

다차원적인 데이터 분석을 위한 OLAP의 접근방법은 크게 두 가지로 구분되는데, 미리 큐브 모양의 구조로 요약된 다차원 데이터베이스를 이용하여 다

차원분석을 수행하는 MOLAP(Multidimensional OLAP)과 관계형 데이터베이스를 이용하는 ROLAP(Relational OLAP)이 그것이다.

MOLAP은 거의 모든 데이터가 숫자로 되어있고 고도로 요약되어 있는 재무 데이터와 같은 경우에 더 적절하며, ROLAP은 데이터 양이 방대하고 변수간의 관련성이 그리 높지 않은 인사 데이터와 같은 경우에 주로 쓰인다.

SAS는 SAS MDDB 서버를 근간으로하는 MOLAP과 SAS의 관계형 테이블을 근간으로하는 ROLAP 양자를 모두 지원한다.

1) MDDB Server

MDDB Server는 통상 RDB 형태로 저장되어있는 데이터 웨어하우스의 데이터를 다차원 데이터 구조로 변환, 적재하여 OLAP 클라이언트의 요청에 따라 데이터를 전달하는 역할을 하는 다차원 데이터 레퍼지토리로서 다음과 같은 특징을 가지고있다.

- -GUI 방식으로 손쉽게 다차원 데이터베이스를 구축, 관리할 수 있다.
- -희박행렬은 저장되지 않기 때문에 저장 공간을 절약할 수 있다.
- -무제한적인 차원간 조합과 무제한적인 분석변수의 수를 지원한다.
- -증가식(Incremental) 갱신 방법을 지원 한다.
- -자주 사용되는 계산값은 미리 다차원 데이터베이스 내에 저장하여 사용할 수도 있으며 실행과 동시에 계산하여 가져올 수도 있다. MDDB 생성시 8개까지의 통계치를 미리 지정할 수 있으며, 실행시 13개의 통계치가 계산가능하다.
- -시스템 차원과 애플리케이션 차원 모두에서 보안 기능을 제공한다.

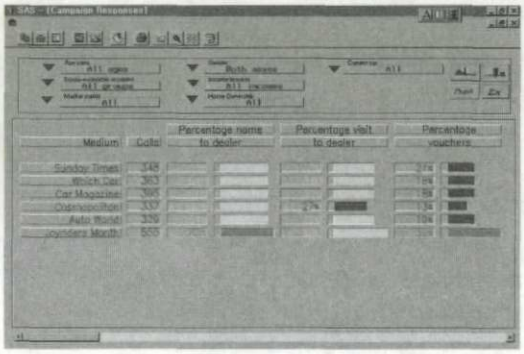
2) 개방형 클라이언트/서버 아키텍처

- -SAS OLAP 솔루션이 제공하는 개방형 클라이언트/서버기능
- -다양한 하드웨어 플랫폼 지원
- -업계 표준의 네트웍 프로토콜을 통한 원격지 데이터에 대한 신속한 엑세스
- -압축된 차원 데이터를 빠르게 옮기는 기능
- -Mobile OLAP지원
- -근원 데이터 조회 기능(Reach Through) 제공

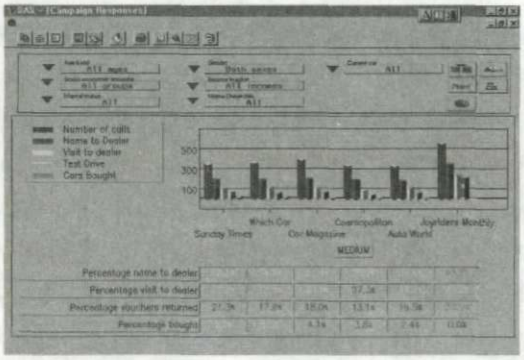
3) 웹상에서의 OLAP 구현

MDDB에 저장된 데이터를 이용한 그래프와 리포트를 웹 브라우저를 통하여 볼 수 있다. 사용자는 어떤 HTML 기반의 웹 브라우저에서도 질의나 다차원 리포팅을 자유롭게 할 수 있다.

4) OLAP 클라이언트 기능



〈그림 3〉 OLAP 클라이언트 기능



〈그림 4〉 OLAP클라이언트 기능의 다른 예

- -드래그 앤 드롭 리포트를 포함한 리포트 갤러리 제공
- -드릴 다운
- -What-if 분석
- -Exception Reporting
- -중요성공요인(Critical Success Factor)의 그래픽 표현
- -3차원 그래픽 보고서
- -지도 그래프
- -다차원 보고서

-조직도

-목표 대비 실적 차트

5) 애플리케이션 개발 환경 지원

정형화된 EIS 애플리케이션 구축을 위한 툴들을 제공할 뿐만 아니라 객체지향적 프로그래밍 기술을 이용하여 사용자의 다양한 요구에 정확하게 부합되는 OLAP 애플리케이션을 개발할 수 있다.

6) 근원 데이터 조회 기능(Reach Through)

MDDB에 요약된 데이터보다 더 상세한 데이터가 필요한 경우 MDDB서버 내의 해당 데이터를 클릭함으로써 근원 데이터베이스로부터 상세 데이터를 바로 가져와서 보여줌으로써 데이터에 대한 보다 세밀한 탐색이 가능하다.

7) OLAP과 데이터 마이닝

OLAP기법은 사용자가 데이터 내에 존재할 것으로 추측되는 관계를 예상하고 접근할 때 보다 효과적으로 정보를 끌어낼 수 있다. 즉, OLAP은 개괄적 차원에서의 분석과 데이터 내의 전반적 경향을 확인, 검증하기 위한 탐색 및 질의 도구인 반면 데이터안에 감추어진 예상치 못한 관계를 발견함으로써 보다 경쟁력 있는 의사결정을 내리기 위해서는 데이터 마이닝 솔루션이 필요하다. OLAP이 "A지역의 월별 A ,B제품 매출 실적을 비교해 보자"라는 문제에 대한 답을 주는 것이라면, 데이터 마이닝은 "A지역에서는 어떤 유형의 사람들이 우리의 고객이 될 수 있는가?"라는 예측적인 질문을 던져 답을 얻어낼 수 있는 보다 진보된 기술이라고 할 수 있다. 다차원적으로 데이터를 분석하여 문제의 골격을 파악한 후 그 다음 단계로 데이터 마이닝 기법을 통해 데이터베이스 내부에 감추어진 패턴과 경향을 발견함으로써 미래 시장의 개척하고, 고객 데이터를 이용하여 마케팅 전략을 수립하는 등 경쟁력 제고에 결정적인 역할을 할 수 있다.

데이터 마이닝

데이터 마이닝은 대용량의 데이터를 이용하여 내재되어 있는 패턴과 경향을 찾아내어 보다 진보된 의사결정을 위한 정보를 캐내는 일련의 과정이다.

SAS가 제공하는 Enterprise Miner 제품은 데이터 마이닝의 5단계 과정인 SEMMA(Sample, Explore, Modify, Model, Assess) 프로세스를 포인트 앤 클릭 방식으로 따라가며 일관되고 명료한 방법으로 데이터 마이닝을 구현할 수 있게 해주는 제품이다.

SEMMA 방법론

• 샘플링(Sampling)

데이터 웨어하우스나 데이터 마트로부터 시작하여 데이터 마이닝을 시행하고자 할 때 고려할 첫번째 사항은 모든 데이터가 실제로 필요한지를 검토하는 것이다. 특정 상설 할인 판매점이나 지점과 같은 하위 단위에 대한 구별을 하지않고 전체 데이터에서 일반적인 패턴이나 경향을 찾고자 한다면 데이터 마이닝 기법을 적용하기 전에 샘플링을 하는 것이 필수적이다. 데이터에 일반적인 패턴이 존재한다면 전체 자료를 사용하지 않더라도 통계적 잘 설계된 표본을 이용하면 보다 효과적으로 그러한 관계를 찾아낼 수 있기 때문이다.

• 데이터 탐색

데이터에서 어떤 형태의 패턴이 발견될지를 지관적으로 추측하기란 대단히 어렵다. 따라서 본격적인 마이닝에 앞서 데이터에 대한 전반적 시각이 생길 때까지 데이터를 여러 각도에서 탐색해 과정을 밟는 것이 바람직하다.

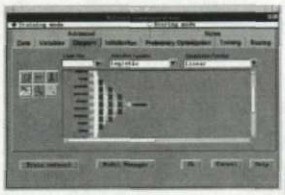
데이터를 그래픽컬하게 표현하여 직관적으로 탐색, 분석하고, 요인분석, 군집분석 등을 이용하여 데이터를 요약하고 탐색하는 과정을 반복해 나간다. 이러한 탐색 단계를 통해 얻을 수 있는 효과의 한 예는 이 단계에서 이미 명확한 구매 패턴을 보여주는 고객 집단은 그 모습이 서서히 드러나고 이들을 위한 특별한 판촉 전략을 수립할 수 있게 된다.

• 수정(Modification)

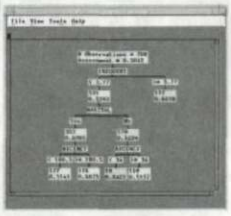
이 단계에서는 탐색 단계에서 발견된 새로운 정보가 분석에 포함되도록 데이터를 수정하고 조작한다.

• 모델링

데이터 마이닝의 핵심이라 할 수 있는 모델의 수립을 사용되는 전형적인 모델링 기법들을 소개하면 다음과 같다.



▲ 〈그림 5〉 신경망 네트워크



▶ 〈그림 6〉 데이터스 플릿

– **신경망 모형** : 신경망 모형은 복잡하고 미묘한 인과관계를 규명하는데 매우 유용한 방법이며 표본 데이터에도 잘 적용된다. SAS의 Enter-prise Miner 제품은 GUI환경에서 손쉽게 신경망 모형을 통한 모델링을 할 수 있게 해준다.

– **의사결정수 모형(decision tree)** : CHAID와 CART와 같은 방법을 사용하여 데이터의 주요 속성에 영향을 주는 또다른 속성들의 인과관계를 영향이 큰 변수를 중심으로 의사결정나무(decision tree)를 구성해 줌으로써 의사결정에 영향을 미치는 속성 또는 변수들 간의 우선순위를 검출할 수 있다.

– **통계적 모형** : 로지스틱 회귀분석, 군집분석, 판별분석 등의 전통적인 통계기법은 데이터 마이닝에 있어서 중요한 역할을 한다.

– **이외에도** : 최근에 데이터 마이닝 기법으로 새롭게 적용되고 있는 Projection Pursuit 알고리즘도 지원된다.

데이터 마이닝의 적용분야

데이터 마이닝의 적용분야로 가장 대표적인 것은 데이터베이스 마케팅이다. 금융, 통신업체에서 많이 적용되고 있는 Target Marketing, 고객이탈방지, Customer Profiling/Segmentation 등에 널리 사용되고 있으며, 이외에도 위험 관리(Risk Management), 품질관리 등에서도 데이터 마이닝 기법이 적용되고 있다. DC