

전문가제언

이젠 데이터 통합을 위한 방법론이 필요하다

김민현/ 투이컨설팅 대표

기업들이 활용하고 있는 기존의 데이터베이스(파일시스템을 포함하여)는 단위 업무시스템들을 개발해나감에 따라 단계적으로 구축되었다. 통합 데이터체계 없이, 각각의 기준에 의하여 데이터베이스를 개발해왔기 때문에, 대부분의 기업들은 몇가지 문제를 공통적으로 안고 있다.

첫째, 상당한 데이터가 중복되어 있다. 중복데이터를 찾아내서 표준화하는 노력을 기울이지 않은 기업의 경우 최소 50% 이상의 데이터중복 현상을 보인다.

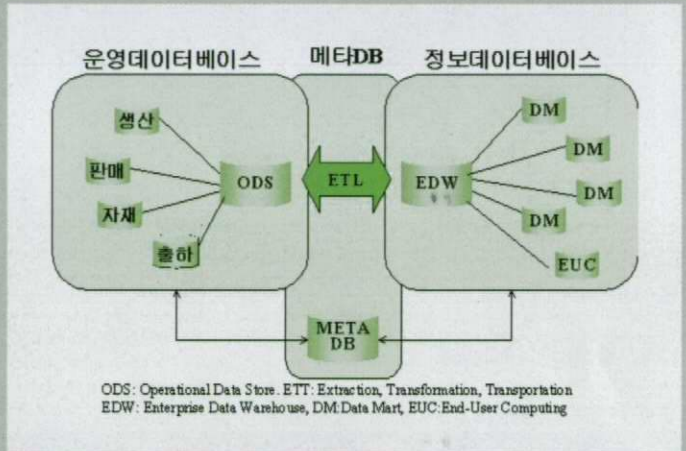
둘째, 데이터 정의가 표준화되어 있지 않다. 데이터정의는 데이터명명, 설명, 형식 등을 모두 포함한다. 데이터정의를 표준화한 경우에는 데이터딕셔너리를 통해서 관리하여야 한다. 국내 기업의 경우 데이터딕셔너리를 제대로 운용하고 있는 기업은 거의 없다.

셋째, 단위 데이터베이스간 인터페이스가 상당히 많고, 인터페이스 방식도 효과적이지 못하다. 통일된 체계가 없이 개발된 데이터베이스에 중복되어 있는 데이터들의 값을 일치시키기 위하여 인터페이스 방식을 이용하고 있기 때문이다. 심한 경우에는 전체 프로그램본수의 15%에서 20%정도가 인터페이스 프로그램이며, CPU를 포함한 전산자원의 상당부분이 인터페이스를 위하여 사용되고 있는 정도이다. 그러면서도, 애플리케이션에 의해서 인터페이스하고 있기 때문에 데이터의 신뢰성을 충분히 보장받고 있지 못한 형편이다.

이러한, 기존 데이터관리의 문제점 때문에 데이터웨어하우스 도입이 필요하다고 주장하는 경우가 있다. 그러나, 데이터웨어하우스는 데이터베이스를 별도로 구축하면서 상당한 범위의 시계열데이터를 포함하기 때문에, 데이터관리를 더욱 어렵게 한다.

데이터웨어하우스는 분석업무요건을 충족시키는 것이 기존의 데이터베이스 체계로는 비효율적이기 때문에 도입하는 것으로, 데이터관리문제를 해결하지는 않는다. 반대로, 데이터 웨어하우스가 도입됨에 따라 전사 관점의 데이터체계에 대한 필요성은 더욱 커진다.

데이터 웨어하우징이 등장하기 전까지, 정보시스템 프로젝트는 업무시스템의 개발이 주된 내용이며, 데이터베이스는 업무시스템을 지원하기 위한 수단에 불과하였다. 데이터베이스 구축 자체가 주된 목표인 정보



〈그림 1〉 데이터베이스의 3가지 유형

시스템 프로젝트는 거의 없었다. 그러나, 1990년대 중반 이후 데이터 웨어하우징 개념이 도입되면서, 데이터베이스 구축을 목적으로 하는 프로젝트가 등장하게 되었다.

데이터베이스 구축을 위한 방법론은 몇가지 제시되고 있으나, 전체를 일관되게 지원할 수 있는 방법론은 아직은 발표되지 않고 있다. 지금까지 정보시스템 프로젝트 분야에 널리 적용되고 있는 소프트웨어개발방법론으로는 구조적방법론과 정보공학방법론이 있다. 구조적방법론은 업무프로세스의 흐름에 초점을 맞춘 것으로, 프로세스의 산출과 결과로서 데이터저장소를 정의한다.

정보공학방법론은 데이터베이스관리시스템의 도입을 전제로 하지만, 시스템설계 및 구축은 업무시스템을 중심으로 수행한다. 데이터 웨어하우징 관련 업체에서 제시하고 있는 데이터 웨어하우징 방법론은 분석용 데이터베이스에 대한 구축방법만을 제시한다. 전사관점에서 데이터체계를 정의하고, 거래처리용 데이터베이스와 업무분석용 데이터베이스에 적용할 수 있는 통합된 방법론이 필요한 것이다.

종래의 시스템 개발방법론인 구조적방법론과 정보공학방법론은 운영 데이터베이스를 대상으로 하고 있다. 몇몇 벤더가 제시하고 있는 데이터 웨어하우징 방법론은 정보데이터베이스와 메타데이터를 대상으로 한다. 전사적 통합시스템을 위한 체계는 정보전략계획수립 방법이 지원하지

만, 기존의 정보전략계획 수립방법에는 데이터체계에 대한 내용을 포함되어 있지 않다. 현재 제시되고 있는 데이터 웨어하우징 방법론은 전사 데이터체계를 다루고 있지만, 정보데이터베이스에 대한 체계면 대상으로 한다.

데이터통합 방법론은 데이터베이스 단위를 정의하고, 데이터베이스 유형별로 적합한 방법론을 제공할 수 있어야 한다. 기업이 필요로 하는 데이터베이스는 세가지 유형이 있다.

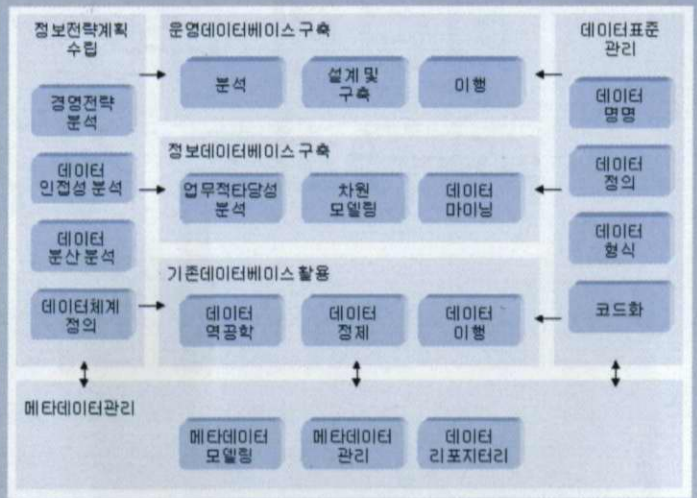
첫째는 운영데이터베이스(OPERATIONAL DATABASES)로서, 업무처리를 지원하기 위한 데이터베이스이다. 예를 들면, 고객의 계좌개설, 입금 등의 데이터가 이에 해당한다. 운영데이터베이스는 애플리케이션과 동시에 개발되며, 애플리케이션에 종속적이다. 운영데이터베이스는 업무별 데이터베이스로 구분된다. 업무별 데이터베이스는 업무시스템의 개발연혁에 따라 구축되어왔기 때문에 데이터의 정의 및 코드값 등이 상이한 경우가 많다. 제조업체의 경우, 생산, 판매, 출하 업무의 연계가 중요하다. 이러한 업무의 연계는 데이터가 공유될 때 가능하게 된다. 데이터의 공유를 위해서 기개발된 시스템을 전면 재개발하는 것은 비용 대비 효과측면에서 적합하지 않다. 이런 경우에는 운영데이터저장소를 통해서 실시간으로 전사의 관련된 데이터를 일치시키는 것이 바람직하다.

둘째는 정보데이터베이스(INFORMATIONAL DATABASE)로서, 전략수립 등의 기획 및 분석업무를 지원하기 위한 데이터베이스이다. 예를 들면 마케팅전략 수립을 위한 고객거래이력데이터 등이 있다. 정보데이터베이스는 데이터베이스 구축이 중요하며, 애플리케이션은 데이터베이스 조회 및 리포팅 도구를 이용해서 사용자가 직접 작성하도록 한다. 정보데이터베이스는 데이터 분석목적에 따라 데이터베이스를 구성하여야 한다. 이때 분석목적별 데이터베이스를 데이터 마트라 부른다. 보통 데이터 마트는 하나 이상이며, 데이터 웨어하우징이 상당히 진전된 기업이라면 보다 많은 데이터 마트를 발견할 수 있다. 전사데이터웨어하우스는 데이터 마트로 통합된 데이터를 제공하는 데이터베이스이다.

셋째는 메타데이터(META DATA)로서 전사적으로 보유하고 있는 데이터의 정의를 관리하는 데이터베이스이다. 예를 들면 고객으로 표현되는 데이터의 정의, 형식, 장소 등을 종합적으로 관리하는 데이터이다. 메타데이터는 데이터딕셔너리, 정보저장소(repository) 등의 형태로 관리된다.

운영데이터베이스와 정보데이터베이스는 데이터의 보관 형태와 주기는 다르지만, 대부분의 데이터는 동일하다. 예를 들어 운영데이터베이스의 고객데이터는 현재 거래중인 고객을 대상으로, 고객 상태를 즉시 조회할 수 있는 형태로 보관한다면, 정보데이터베이스의 고객은 과거의 고객 및 잠재고객까지를 포함하며, 다각적인 분석이 가능한 형태로 보관한다.

운영데이터베이스와 정보데이터베이스가 동일한 데이터를 보관한다



(그림 2) 통합데이터베이스 구축을 위한 방법론

면 데이터의 전송이 불가피하게 발생한다. 정보데이터베이스의 대상 범위가 커질수록 인터페이스는 복잡해지고 많아진다.

통합데이터베이스 구축을 위한 방법론은 다음 요건들을 충족해야 한다.

데이터아키텍처를 수립할 수 있어야 한다. 이때 데이터아키텍처는 기업이 필요로 하는 모든 데이터베이스를 대상으로 한다. 데이터아키텍처는 정보체계, 시스템체계와 연계해서 정의되어야 한다.

데이터베이스 단위별로 운영데이터베이스와 정보데이터베이스로 분류할 수 있어야 한다. 이를 위해서는 분산분석기법이 필요하다. 운영데이터베이스는 적용업무 구축 프로젝트의 일부로 정의한다. 정보데이터베이스는 데이터베이스구축 프로젝트로 정의한다. 각각의 프로젝트에 대해서는 업무적 타당성을 검증할 수 있어야 한다.

메타데이터를 정의하고 관리할 수 있는 기법을 제공하여야 한다. 메타데이터의 모델링방법과 메타데이터의 보관 및 검색방법 등을 포함하여야 한다.

기존에 운영하고 있는 데이터베이스를 활용할 수 있는 방법을 제시하여야 한다. 이를 위해서는 데이터의 정제(cleansing) 기법이 필요하다.

운영데이터베이스의 구축 방법은 엔티티관계 모델링 기법을 중심으로 정규화 관점에서 수행할 수 있어야 한다. 정보데이터베이스의 구축은 분석대상 사실(facts)과 분석 관점(dimension)을 모델링할 수 있어야 한다.

운영데이터베이스와 정보데이터베이스에 공통적으로 적용하기 위한 데이터 표준을 정하고 관리하는 방법을 제시하여야 한다. 데이터표준에는 코드화 기법을 포함하여야 한다.

정보데이터베이스를 활용하는 방법이 필요하다. 축적된 데이터를 분석하여 업무적으로 의미있는 결론을 도출할 수 있는 통계적 기법(data mining)을 제공하여야 한다. 