□ 기술해설 □

제품 기술 동향

데이타 웨어하우스를 위한 데이타베이스 기술

한국오라클 안병문

1. 서 론

의사결정권자가가 의사결정을 하는데 필요한 정보를 제공하기 위한 해결책으로 데이타웨어하우스가 부상하고 있다. 데이타베이스 업체를 선두로 하드웨어 공급업체, 시스템 통합업체. OLAP(online analytic processing)툴 공급업체등 많은 업체들이 나름대로의 솔루션을 가지고 데이타웨어하우스를 구축하겠다고 나서고 있고, 일반 기업들의 이에 대한 관심이 급격히 확산되고 있다.

데이타 웨어하우스는 기업내에서 일어나는 전단계의 의사결정 프로세스를 지원하기위해 모든 형태의 데이타를 주제별로 모아 놓은 것으로 정의할수 있다. 다시말해 의사결정지원 전용의 별도의 데이타 저장소를 가져서 원하는 때에 원하는 형태의 정보를 얻고자하는 것이다. 그러나 데이타웨어하우스는 기존의 클라이언트/서버 방식의 OLTP시스템과 비교하여 볼때 사용자, 데이타내용, 데이타구조, 하드웨어, 소프트웨어, 시스템관리 등 많은 부분에서 다르다. 또한 조직전체의 상세 데이타를 포함하고 있어 데이타를 집계해서 정형화된 프로그램을 사용하는 기존의 DSS, MIS, EIS시스템과 달리 새로운 방식의 데이타 관리기술이 필요하다는 것을 짐작할수 있다.

성공적인 데이타웨어하우스는

- 계량적인(혹은 계수적인) 결과치를 줄 수 있어야 하고
- 실행하고 유지하는데 비용면에서 효과적이어야 하며
- 운영중인 데이타, 외부 데이타, 과거 데이타 모두를 종합할 수 있도록 설계해야 하며
- 운영중인 데이타의 재연이 아닌 목적에 부합되는 데이타를 갖고 있어야 하고
- 체 아키텍쳐가 유연하고 확장될 수 있어야 한다. 다시말해 정보수용능력을 확장할 수 있어야 하며 새로운 형태의 데이타 저장방식이나 액세스 방법에도 대처할 수 있어야 한다.

이를 실현하기 위해서는 대용량 데이타의 처리에 우수한 성능을 가진 데이타베이스 엔진은 물론이고 데이타를 다양한 각도에서 분석하기 위한 데이타분석 도구, 향후 웹과의 연계성 등 다양한 요소들이 고려되어야 한다.

여기에서는 데이타웨어하우스를 효과적으로 구축, 관리, 활용하기 위한 데이타베이스의 필요한 기능에 대하여 간략하게 언급하고자 한다.

2. 데이타베이스의 필요한 기능

데이타웨어하우스의 가장 큰 이슈는 얼마나 사용자의 요구를 빠르게 보여 줄 수 있는가 이다. 이를 역으로 생각하면 데이타웨어하우스는 근본적으로 모든경우의 사용자 요구에 신속하게 응답하기 어렵다는 것을 내포하고 있음을 알 수 있다. 그 이유는 무엇보다도 액세스해야 할 대상 데이타가 수십기가에서 수테라바이트에 이르는 대용량일 수 있기 때문이다. 따라서 데이타베이스의 성능과 이를 활용하는 기술이 데이타웨어하우스를 성공하느냐, 실패하느냐의 열쇠를 쥐고 있다해도 과언이 아닐 것이다.

아래는 데이타웨어하우스에 꼭 필요한 몇가지 DB기능 및 대용량의 데이타를 액세스하기 위한 활용 기술을 나열하였고 차례로 설명을 하고자 한다.

- 병렬처리
- 비트맵 인덱스(bitmap index)
- 스타조인(star-join) 조회의 최적화
- 데이타분할 및 뷰(view)의 지원
- 데이타베이스 안정성
- 멀티미디어 데이타의 지원

2.1 병렬처리

데이타베이스에 관련된 프로세스중 조회, 데이타로딩, 인데스 생성, 백업 및 복구 등 배치성 작업들은 대부분이 CPU를 최대한 이용하여 수행속도를 향상시키는 병렬처리를 사용하고 있으며, 이러한 병렬처리기술은 대용량의 배치작업 및 복잡하고 다양한 조회를 하는 데이타웨어하우스에 꼭 필요한 기능으로 복잡한 질의어의 수행속도를 개선할 뿐 아니라 저렴한 가격으로 시스템 구성을 할 수 있는 장점이 있다. 이러한 처리기능의 내부구조는 하나의 데이타를 실행시 분할하는 동적(dynamic)분할과 실행이전 미리 데이타를 분할하는 정적(static)분할로 나누어 볼 수 있으며 작업의 성격에 따라 적절히 사용한다면 많은 성능개선을 가져올 수 있다.

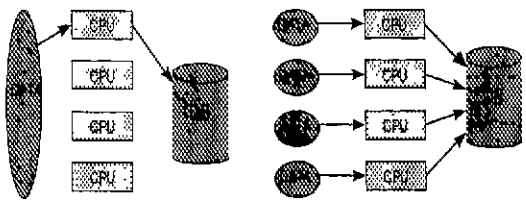

그림 1 병렬처리 구조

2.2 비트맵 인데스(bitmap index)

비트맵 인덱스는 데이타웨어하우스나 의사결정지원시스템 구축시 조회속도나 데이타 저장 및 유지보수를 용이하게 할 수 있는 인덱싱(indexing)기법이다. 이 기능은 다음과 같은 업무에 적절히 사용할 수 있다.

- 주로 배치성 작업이 많은 데이타웨어하우스 업무
- 분포도가 좋지 않은(서로 다른 레코드에 같은 값을 가지는 데이타가 많은) 컬럼의 결합 인덱스를 사용하는 경우
- 인덱스가 너무 크고 많아서 디스크를 과다하게 사용할 때
- 원시데이타를 여러 형태로 분석하고 조회해야 할 필요가 있는 업무
- 인덱스 생성시 시간이 많이 걸리는 경우

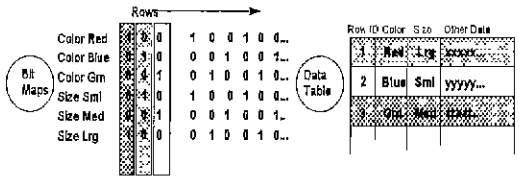

그림 2 비트맵 인덱스를 통한 데이타 추출

그림 2에서 만일 테이블의 레코드가 10,000건이고, color 컬럼에는 red, blue, green의 3가지 값만 존재하고, size컬럼에 small, medium, large의 3종류의 값만 허락한다면, 데이타건수에 비해 값의 종류가 매우 적기때문에, 이것을 우리는 분포도가 좋지 않다고 하며, 두개의 컬럼에 대하여 결합인덱싱을 한다면 비트맵 인덱싱을 하는 것이 유리하다. 이때 인덱스에 저장되는 컬럼의 값은 비트(bit)값으로 처리됨으로 디스크를 상대적으로 적게 차지한다. 이러한 비트맵 인덱스는 데이타웨어하우스에서 큰 장점을 가지는 반면, 데이타가 자주 수정되는 OLTP업무나 분포도가 좋은(같은 값을 가지는 경우가 적은)컬럼에는 사용하지 않는 것이 효과적이다.

2.3 스타조인(star-join)조회의 최적화

스타조회는 데이타웨어하우스에서 아주 중요하다. 스타스키마(star schema)는 일반적으로 데이타가 많은 1개의 상세(fact) 테이블과 이를 참조하는 데이타가 적은 여러개의 참조(dimensional) 테이블로 구성되어 있어, 사용자의 조회요구시 상세테이블과 사용자의 조회조건에 포함된 참조테이블을 조인하게 되는데, 이때 가장 효과적인 조인방법은 그림 3에서 보는 바와같이 조회조건에 포함된 참조테이블끼리 먼저 조인하여 메모리에 중간테이블(cartesian

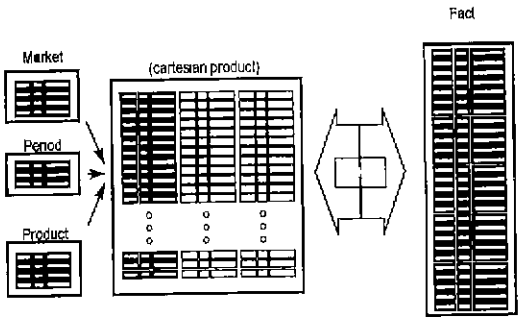

그림 3 스타조인 실행 구조

product)를 만들고 이를 상세테이블과 조인하는 방법이다. 이것을 스타조인이라고 하며, 이를 위해서 상세테이블에는 반드시 참조테이블의 키를 참조하는 키(foreign key)에 대한 결합인덱스(concatenated index)를 갖게 되는데 이때 결합인덱스의 순서는 기능향상에 많은 영향을 주게된다. 데이타웨어하우스에 사용되는 데이타베이스의 옵티마이져(optimizer)는 이와같은 스타조인을 인식하고 가장 효과적인 실행계획을 세울 수 있어야 하며, 이러한 기능은 테이블 갯수의 제약을 받지 않고 사용할 수 있어야 한다. 만일 옵티마이져가 스타조인 조회를 최적화하지 못할 경우 심각한 성능저하를 가져오게 된다.

2.4 데이타 분할 및 뷰(view)의 지원

상세테이블과 같은 큰 테이블에 대한 가용성, 운용성 및 조회 성능향상을 보장하기 위해서는, 하나의 테이블의 데이타를 어떤 기준에 의해서 분할하여 여러개의 테이블에 저장하고 사용자에게는 논리적인 하나의 뷰를 제공하여야 한다.

그림 4에서 보는 바와 같이 분할된 테이블은 개별적으로 관리할 수 있으므로 하나의 테이블

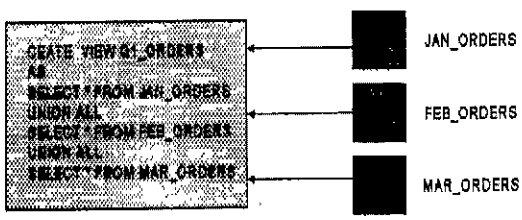

그림 4 분할테이블과 뷰의 생성방법

```
SELECT SUM(REVENUE)
  FROM Q1_ORDERS
 WHERE ORDER_DATE BETWEEN '96/01/20'
                      AND '96/02/10'
```

에 문제가 발생하더라도 전체에 영향을 주지 않을 뿐 아니라 인덱스 데이타의 재생성, 데이타 로딩, 데이타 삭제 등을 간편하게 할 수 있고, 일부 테이블에 대한 데이타 조회와 같은 부분적인 작업시에 성능향상을 꾀할 수 있다. 예를 들면, 사용자가 1월 데이타(JAN-ORDERS)와 2월 데이타(FEB-ORDERS)만 요구하는 SQL문장을 아래와 같이 작성하였다면 옵티마이져는 빠진 날짜를 알아내어 실행계획시 3월 데이타(MAR-ORDERS)는 제외시키게 된다.

2.5 데이타베이스 안정성

데이타웨어하우스의 데이타는 대용량이면서 사용자의 요구는 다양하고 복잡한 양상을 띠고 있고 이러한 데이타를 백업(backup)하고 복구하는 일은 어렵고 힘든 일이 될 수 있으며 따라서 이에 대한 실행계획을 세우는 것은 매우 중요하다. 이를 위해 데이타베이스는 기본적으로 콜드(cold)백업과 핫(hot)백업을 안정적으로 지원할 수 있어야 하고 테이블, 사용자 또는 전체를 백업단위로 해서 손쉽게 작업할 수 있어야 하며, 단위 복구시에도 신속하고 편리하게 할 수 있는 방법을 제공해야 한다. 대부분의 경우 데이타웨어하우스는 최초 많은 데이타를 로드하고 다음 데이타는 적은 양의 변경된 데이타가 추가되는데, 이러한 변경된 데이타에 대해서만 쉽게 로드할 수 있는 방법도 제시하는 것이 필요하다. 실제로 얼마 못가서 다시 데이타를 만들어야 하는 일이 빈번히 발생하거나 백업 및 복구과정이 어렵고 복잡한 일이라면 희망에 부푼 사용자의 기대에 찬물을 붓는 일이 될 수도 있다. 데이타베이스의 보이지 않는 가장 핵심적인 기술, 데이타베이스의 품질보증은 데이타웨어하우스를 성공하기 위한 가장 중요하고 기본적인 일임에 틀림없다.

2.6 멀티미디어 데이타의 지원

데이타웨어하우스의 데이타베이스는 행과 열로 표시되는 정형데이타 이외에 텍스트, 이미지, 비디오, 시공간데이타 등의 비정형 데이타를 포함하는 모든형태의 데이타를 취급할 수 있어야 한다. 이렇게 기업 내의 다양한 형태의 데이타를 포괄적으로 수용할 수 있을때 데이타웨어하우스의 진정한 역할을 할 수 있으며 앞으로 데이타 형태의 변화에 유연하게 대처할 수 있으리라 생각된다.

3. 결 론

데이타웨어하우징은 다양한 구성요소를 갖고 있으며 대부분의 경우 다양한 기술이나 제품이 포함된다. 성공적으로 데이타웨어하우스를 구현하는가는 이에 관련된 핵심기술들을 이해하고 얼마나 잘 활용하는가에 달려있다. 앞으로 데이타웨어하우징이 기업의 비지니스를 성공하기 위해서 필수적인 시스템으로 인식되면서 기업의 전산시스템의 핵심이 될 것이며, 인터넷의 웹은 물론 애플리케이션 패키지 시스템과도 연결이 되면서 많은 새로운 기술들이 나타나면 지금의 기술이 한단계 발전하리라 생각된다. 지금의 상황을 면밀히 분석하고 미래를 대비한다면 데이타웨어하우스를 단순한 의사결정을 돕기위한 별도의 시스템만으로 인식할 것이 아니라 기업의 경쟁력확보를 위한 전략적 무기로 활용하는 기술로 인식해야 할 것이다.

안 병 문

1974 서울대학교 응용물리학과 학사
1978~1993 삼성전자 컴퓨터사업본부 개발부장, 컴퓨터부문 소프트웨어 및 네트웍 개발실장 역임
1989 한국과학기술원 전산학과 석사
1993~현재 한국오라클 기술본부장(상무)