


BUSINESS INTELLIGENCE

In-Database Analytics : DB 내에서의 효율적인 정보 분석 방안

장 성 우
BI/DW팀장
한국 오라클



ORACLE

BUSINESS INTELLIGENCE

목차

- 정보 분석 개요
 - 정보 분석 트렌드
 - 효율적인 정보 분석 전략 : In-Database Analytics
- DB 내에서의 통합된 정보 분석 방안
- 요약 및 Q&A
- 별첨

ORACLE

BUSINESS INTELLIGENCE

ORACLE

정보 분석 개요

ORACLE

BUSINESS INTELLIGENCE

정보 분석 개요

Query and Reporting	OLAP	Data Mining
상세 정보의 추출	요약 및 경향 분석	숨겨진 패턴의 발견을 통한 지식 획득
“정보”	“분석”	“통찰 & 예측”
지난 3년간 펀드를 구입한 사람들은 누구인가?	펀드 구매자들의 지역별, 연도별 평균 이득은 얼마인가?	다음 6개월 동안 펀드를 구매할 것으로 예측되는 사람들은 누구이며 그 이유는?

“Business Intelligence”

ORACLE

BUSINESS INTELLIGENCE

정보 분석 관련 주요 트렌드

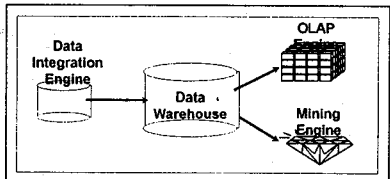
- 정보의 크기는 점점 커지는 추세
 - 3년 전 전 세계에서 가장 큰 DW : 30TB
 - 작년 세계에서 가장 큰 DW : 100TB
 - 2,3년 내로 PB급 DW가 나올 것으로 예상됨
- 정보의 저장은 이제 큰 문제가 아님. 진짜 문제는 정보 분석 방법임
 - 차원이 4000개이고 크기가 2TB인 정보를 어떻게 분석할 것인가?

ORACLE

BUSINESS INTELLIGENCE

현재의 정보 분석 프로세스

- 분석 업무의 분리
 - 서로 다른 곳에서 분석 수행 : 다른 시스템, 다른 담당자
- 개별적인 분석 애플리케이션 사용
 - 업무별 전문 패키지 사용 : 전문성은 좋으나 통합성은?
- 주요 고찰 이슈 : 구축 및 유지 비용, 실시간 대응성



ORACLE

현재 프로세스의 문제점

- 데이터의 빈번한 이동
 - 데이터의 크기가 적을 때는 별 문제가 없음. 하지만, 분석해야 할 데이터의 크기가 커진다면?(예:TB급의 고객 정보 분석)
 - 데이터의 크기가 커질수록 중요한 이슈로 대두됨
 - ✓ 데이터 중복 저장 비용
 - ✓ 데이터 이동에 따르는 시간 손실
 - ✓ 전문 패키지의 정보 확장성 및 성능 이슈
- 업무 프로세스의 분리
 - 전체 분석 프로세스의 지연
 - ✓ 실시간 분석 및 대응이 불가능

ORACLE

바람직한 정보 분석 프로세스

- 한 곳에서 정보 관리 및 분석 수행
 - 데이터 이동을 최소화하여 서버간 이동에 따르는 불필요한 시간 지연 제거
 - 중복 저장 불필요
 - 안전하고 효율적인 정보 관리
 - ✓ Security, Scalability, Availability
- 정보 분석 업무의 유기적 연결 및 차별화
 - 일반 정보 분석 업무의 상시 & 실시간화
 - ✓ 단일 SQL로 정보 분석 업무 처리
 - 필요 시 전문 패키지를 통한 고급 분석 수행

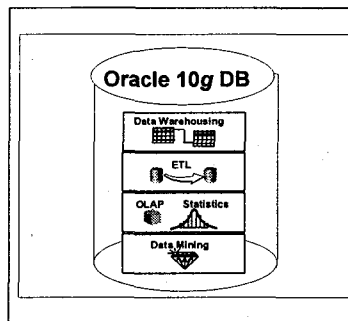
ORACLE

ORACLE

DB 내에서의 통합된 정보 분석 방안

ORACLE

In-Database Analytics

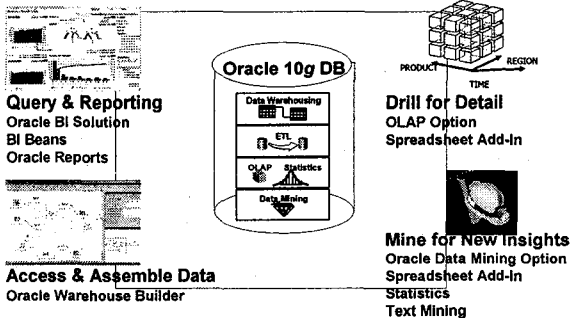


- 단일 DB 내에서 통합된 정보 분석 업무 지원
 - Data Warehouse
 - Built-in Statistics
 - OLAP Option
 - Data Mining Option

ORACLE

Oracle Business Intelligence

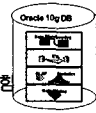
Know More, Do More, Spend Less!



ORACLE

In-Database Analytics의 장점

- 기술적인 측면
 - 데이터는 항상 적절한 제어 하에 DB 상에 존재함
 - 복합 질의를 통해 직관적인 분석 처리 가능
 - 확장의 용이성 및 우수한 처리 성능
 - ✓ Fast scoring : 단일 CPU 시스템에서 250만개의 레코드를 단 6초 만에 정수 부여 작업을 마침
- 비즈니스적인 측면
 - 실시간의 정보 분석 처리 가능
 - TCO의 절감 가능



ORACLE

In-Database Analytics : 사례

- 예제 : DVD 마케팅 캠페인 시행 결과의 통계적 검증
 - 사전 정의된 분류 방법에 의해 반응 모델이 만들어져 있을 때, 이를 이용하여 어떤 고객이 마케팅 캠페인에 응할 것인가를 예측
 - 각각의 고객들이 캠페인 시행 이전 3개월과 이후 3개월 동안 얼마만큼 DVD를 구매했는가를 분석
 - 예측된 고객들의 캠페인 성공률과 반응하지 않은 고객들의 구매율을 서로 다른 지역과 회사별로 비교하고, 이 정보들의 통계적으로 유의한지의 여부를 검증

In-Database Analytics : 사례 기존의 처리 방법

- 1단계 : 데이터 마이닝 프로그램
 - DB로부터 고객 데이터를 전달 받음
 - 프로그램 상에서 예측 작업 수행
 - 예측된 사용자 정보를 DB에 재 전송
- 2단계 : DB 검색
 - 예측된 고객 정보를 로딩
 - 해당 고객들의 캠페인 전후 구매 상황을 검색
 - 캠페인 성공 여부 정보를 검색하여 정리
- 3단계 : 통계 패키지
 - 캠페인 성공률 정보를 DB로부터 받음
 - 통계적 검증 작업을 수행

In-Database Analytics : 사례 오라클 상에서의 처리 방법

- 하나의 SQL로 수행 가능

```

select responder, cust_region, count(*) as cnt,
       sum(post_purch - pre_purch) as tot_increase,
       avg(post_purch - pre_purch) as avg_increase,
       state_t_test_paired(pre_purch, post_purch) as significance
from (
select cust_name,
       prediction(campaign_model using *) as responder,
       sum(case when purchase_date < 15-Apr-2005 then
         purchase_amt else 0 end) as pre_purch,
       sum(case when purchase_date >= 15-Apr-2005 then
         purchase_amt else 0 end) as post_purch
from customers, sales, products@PRODD8
where sales.cust_id = customers.cust_id
and purchase_date between 15-Jan-2005 and 14-Jul-2005
and sales.prod_id = products.prod_id
and contains(prod_description, 'DVD') > 0
group by cust_id, prediction(campaign_model using *) )
group by rollup responder, cust_region order by 4 desc;
    
```

통계 : 유의성 검증
마이닝 : 캠페인 예측
기본 DB 정보 검색

In-Database Analytics : 종합

- 사례를 통해 본 DB내 정보 분석의 장점
 - 데이터 이동이 전혀 없음(SQL 안에서 pipelining)
 - 분석 프로세스가 단순해 짐
 - 실시간 분석이 가능
- 고려 사항
 - DB와 DM, 통계를 모두 아는 전문가 필요
 - 일반 및 고급 분석 프로세스의 분리
 - 일반 분석 : 상시화 및 실시간 화
 - 고급 분석 : 전문화

DB 내 정보 분석 전략의 이점

In-Database Analytics	Benefit
분석 애플리케이션을 위한 플랫폼 제공	<ul style="list-style-type: none"> 데이터 이동을 없애고 보안 이슈에의 노출을 최소화 할 수 있음 빠른 정보 관리 체인 제공
넓은 범위 마이닝 및 통계 처리 알고리즘 제공	<ul style="list-style-type: none"> 대부분의 정보 분석 문제에 대한 해결 방안을 제공
복수의 HW, O/S에서 운용 가능	<ul style="list-style-type: none"> 다양한 운영 환경에서 분석 애플리케이션 수행 가능
오라클 DB 기술을 최대한 활용 가능	<ul style="list-style-type: none"> Grid, RAC, 통합 BI,... SQL & PL/SQL 사용 가능 기존의 DB 기술 최대한 활용

요약

- DB는 더 이상 단순 데이터 관리의 장소가 아니며, 실시간 정보 분석의 핵심 요소임
- 데이터 측면의 RTE 구현 방안
 - DB의 통합
 - 단순화, 표준화, 전문화
 - 정보 전달 체인의 효율화
 - 통합 DB 상에서의 정보 분석
 - 정보 분석 업무의 개선
 - 단순 분석의 실시간화, 고급 분석의 전문화

핵심 가치 : 정보 관리 및 분석의 실시간화 및 이를 통한 정보 회전율 증대

별첨 : 오라클에서 제공하는 통계 함

Ranking functions

- rank, dense_rank, cume_dist, percent_rank, rank

Window Aggregate functions (moving and cumulative)

- Avg, sum, min, max, count, variance, stddev, first_value, last_value

LAG/LEAD functions

- Direct inter-row reference using offsets

Reporting Aggregate functions

- Sum, avg, min, max, variance, stddev, count, ratio_to_report

Statistical Aggregates

- Correlation, linear regression family, covariance

Linear regression

- Fitting of an ordinary-least-squares regression line to a set of number pairs
- Frequently combined with the COVAR_POP, COVAR_SAMP, and CORR functions.

Note: Statistics and SQL Analytics are included in Oracle Database Standard Edition

Descriptive Statistics

- average, standard deviation, variance, min, max, median (via percentile), count, mode, group-by & roll-up
- DBMS_STAT_FUNCS: summarizes numerical columns of a table and returns count, min, max, range, mean, stats_mode, variance, standard deviation, median, quartile values, %n sigma values, top/bottom 5 values

Correlations

- Pearson's correlation coefficients, Spearman's and Kendall's (both nonparametric).

Cross Tabs

- Enhanced with % statistics: chi squared, phi coefficient, Cramer's V, contingency coefficient, Cohen's kappa

Hypothesis Testing

- Student's t-test, F-test, Binomial test, Wilcoxon Signed Ranks test, Chi-square, Mann Whitney test, Kolmogorov-Smirnov test, One-way ANOVA

Distribution Fitting

- Kolmogorov-Smirnov Test, Anderson-Darling Test, Chi-Squared Test, Normal, Uniform, Weibull, Exponential

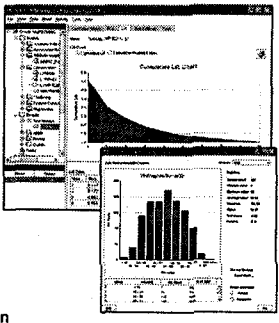
Pareto Analysis (documented)

- 80:20 rule, cumulative results table

ORACLE

별첨 : Oracle Data Mining Option

- Oracle mining platform
 - PL/SQL API
 - Java API
 - Oracle Data Miner (GUI)
 - Spreadsheet Add-In
- 지원하는 알고리즘
 - Structured & unstructured data
 - Attribute importance
 - Classification, regression & prediction
 - Anomaly detection
 - Association rules
 - Clustering
 - Nonnegative matrix factorization
 - BLAST



ORACLE

별첨 : Oracle OLAP Option

- DW 상에서 OLAP 동시 구현
 - 하나의 DB 상에서 대규모의 관계형 데이터와 다차원 데이터 집합 동시 지원
 - 별도의 데이터 구축 작업 없이 매방 작업만으로 다차원 큐브의 신속한 구축 가능
 - 압축, 파티션, 병렬 처리를 통한 신속한 Ad-Hoc 질의 처리 가능
 - 다차원 데이터 타입에 대한 일반적인 SQL 인터페이스 지원
 - OLAP API를 통한 최적화 및 확장 지원
- 주요 장점
 - 빠른 데이터 처리 성능
 - 큐브 구축 및 질의 실행 모두 빨라짐
 - 순쉬운 사용
 - SQL 및 OLAP API를 통한 순쉬운 개발 및 검색 가능

ORACLE