IT WORLD CIO

Big Data Special Report

DW 선택 가이드 2

- **데이터의 범람과 분석 필요성의 대두**
  - 왜 빅 데이터인가? 주목해야 할 이유
  - 빅 데이터 수혜 직종 '데이터 과학자' FAQ
- **변화하는 분석 대상과 기술**
  - 빅 데이터가 가져올 IT업계의 변화
  - 일상 생활 모니터링해 미래의 행동 예측
  - 부상하는 하둡, '아직은 RDBMS와 병행해 사용'
- **빅 데이터를 위한 선택–DW어플라이언스**
  - 우리캐피탈 | 빠르고 정확한 데이터 분석으로 경영 전략 강화
  - 코리아크레딧뷰로(KCB) | 18초만에 38억 건 데이터 조회 완료
- **빅 데이터를 위한 제언**
  - 기고 | 고성능 분석을 위한 논리적 DW 재구축 전략

Sponsored by IBM

왜 빅 데이터인가? 주목해야 할 이유

Stacy Collett | Computerworld

데이터가 폭증하면서 가장 발 빠르게 대응한 IT분야는 데이터 마이닝 전문 업체들이었다. 이들은 방대한 데이터를 거르는 방법을 더 빠르고 저렴하게 바꾸는 전략을 내놓았다. 이어서 빅 데이터가 대두됐고 다른 DW, DW어플라이언스, BI업체까지 빅 데이터에 맞는 솔루션들을 내놓고 있다.

데이터 용량에 대한 전망은 그 동안 여러 번 언급됐다. 2020년까지 디지털화돼 저장된 데이터의 용량은 35조 GB에 달해 2009년의 44배 수준이 된다고 한다. IDC의 조사에 따르면, 이미 2010년 말에 120만 PB, 즉 1.2ZB 수준을 데이터 용량을 달성했다. 이를 DVD에 저장하면 지구와 달 사이를 왕복할 만큼 쌓을 수 있다. 편도로 약 38만 km의 길이다.

초기의 데이터 마이닝 기술은 저장장치를 구축하고 더 저렴하고 빠르게 데이터를 조작하고 분석할 수 있게 했다. 향후 수퍼컴퓨팅이 확산되면 빅 데이터(Big Data) 기술은 순식간에 전파돼 기업의 업무처리 방식을 바꿔 놓을 것으로 기대된다.

컴퓨터월드는 빅 데이터에 대해 '하둡(Hadoop)을 포함하지만, 이에 국한되지 않고 비전통적인 데이터 감별 툴을 이용해 구조화된 데이터와 비정형화된 데이터의 거대한 세트에서 정보를 캐내는 것'으로 정의하고 있다.

클라우드와 빅 데이터는 그동안 광고만 무성했고 실체를 알 수 없다는 점에서 닮았다. 이제 막 실체가 드러났다는 것까지도 둘은 비슷하다. 애널리스트들과 빅 데이터 지지자들에게 빅 데이터가 미래의 데이터 발굴에 어떠한 의미를 지닐지, 그리고 기업들이 무엇을 오해하고 있는지에 대해 물었다.

빅 데이터의 등장 배경

시스템 비용이 낮아지고 처리 능력은 향상되면서 빅 데이터가 주목 받기 시작했다. 메인 메모리 가격이 내려감에 따라 기업들은 그 어느 때보다도 많은 데이터를 메모리 내에서 처리할 수 있게 됐다. 또한 서버 클러스터에 컴퓨터들을 연결하기가 더 쉬워졌다. IDC의 DBMS 담당 애널리스트 칼 올롭슨은 3가지가 조합돼 빅 데이터가 생겨났다고 말했다.

"이제는 빅 데이터를 감당할 수 있을 뿐 아니라 적절하게 처리할 수 있다"라고 올롭슨은 전했다. 과거에는 일부 대형 수퍼컴퓨터만이 다른 시스템들과 연동돼 다중으로 처리할 수 있었다. 이 수퍼컴퓨터들은 특화된 하드웨어였기 때문에 가격이 수십만 달러 이상이었다. 이제는 상용화된 하드웨어를 이용해 그런 기능성을 달성할 수 있다. 이를 통해 우리는 데이터를 더 빠르면서 저렴하게 처리할 수 있게 됐다는 게 올롭슨의 설명이다.

그러나 거대한 데이터 저장시설을 갖춘 모든 기업이 빅 데이터 기술을 사용하는 것은 아니다. IDC는 빅 데이터에 대해 기술을 감당할 수 있어야 하며, IBM이 3V(다양성(Variety), 볼륨(Volume), 속도(Velocity))라고 설명하는 3가지 요소들 중 2가지를 충족해야 한다고 정의했다.

다양성은 구조화된 데이터뿐 아니라 비정형 데이터 형태로도 들어온다는 의미다. 볼륨은 수집되고 분석되는 데이터의 양이 매우 크다는 것을 의미한다. 그리고 속도는 데이터가 처리되는 속도를 뜻한다.

올롭슨은 빅 데이터에 대해 다음과 같이 설명했다. "빅 데이터가 항상

수백 TB를 의미하는 것은 아니다. 경우에 따라 수백 GB로도 3차원적으로 속도나 시간을 고려하면 상당한 규모가 될 수 있다. 300GB를 분석하고 처리하는데 1시간이 걸리던 것을 1초 만에 처리할 수 있다면 그 결과로 기업이 할 수 있는 것은 크게 달라지기 때문에 가치를 더한다고 할 수 있다. 빅 데이터는 이들 3가지 요소들 중 최소 2가지를 충족하는 애플리케이션이다."

오픈소스 연계

"많은 사람들이 하둡과 빅 데이터를 동의어라고 생각한다. 하지만 그것은 실수다"라고 올롭슨은 지적했다. 테라데이타, MySQL, 하둡을 사용하지 않는 '클레버 클러스터링(Clever Clustering) 기술' 이행의 일부도 빅 데이터로 볼 수 있다고 그는 설명했다.

현재 가장 크게 주목 받는 것은 빅 데이터를 위한 애플리케이션 환경인 하둡이다. 하둡은 수퍼컴퓨팅에서 가장 보편적인 접근방식인 맵리듀스를 기반으로 하고 있으면서도 구글이 지원하는 프로젝트를 통해 단순하면서도 고급스럽게 바뀌었기 때문이다.

빅 데이터에 관한 3가지 오해

빅 데이터가 무엇이며 무엇을 할 수 있는지에 관한 상당한 오해가 있다. 다음의 3가지 오해에 관해 알아보자.

1. 관계형 DBMS는 매우 큰 볼륨으로 확장할 수 없기 때문에 빅 데이터 기술로 보기 어렵다. (사실이 아니다.)
2. 하둡 또는 확장에 따른 모든 맵리듀스(Map-Reduce) 환경은 작업부하와 경우에 상관없이 빅 데이터를 위한 최선의 선택이다. (이 또한 사실이 아니다.)
3. 도식적인(Schematic) DBMS의 시대는 끝났다. 개요 개발(Schema Development)은 빅 데이터 도입에 방해만 될 뿐이다. (절대 사실이 아니다.)

출처 : IDC, "빅 데이터의 중요성(The Big Deal about Big Data)," 2011년 2월 (칼 W. 올롭슨)

하둡은 맵리듀스 환경에서 자주 볼 수 있는 아파치(Apache) 프로젝트들의 완성체라 할 수 있다.

소프트웨어 개발자들은 하둡을 분석하는 모든 종류의 기술과 이와 유사한 진보된 기술들을 동원해 이에 대응했으며, 그 기술들의 대부분은 오픈소스 커뮤니티에서 개발됐다. "그들은 소위 noSQL 데이터베이스라는 다양한 데이터베이스들을 만들어 냈으며, 대부분의 데이터베이스들에는 핵심값(Key Value)이 포함돼 있는데 여러 가지 기술로 핵심값 처리량을 최적화했다"라고 올롭슨은 말했다.

그는 그러나 오픈소스 기술들의 경우, 상업적으로 지원되지 않기 때문에 "좀더 발전해야 하며 대대적인 개혁을 이끌어내는데 몇 년이 걸릴 것이다. 이런 초기의 특성 때문에 한동안 상용화에 성공하기는 어려울 것"이라고 덧붙였다.

IDC는 오픈소스 기술이 발전하는 동안 최소한 3개 IT업체들이 올해 말 하둡을 위한 일부 지원 서비스를 제공할 것으로 전망했다. 또한 데이터미어(Datameer)같은 다수의 업체들은 기업들이 자체적인 애플리케이션을 개발할 수 있도록 하둡 구성요소를 이용해 분석 툴을 출시할 것이 확실시 된다고 덧붙였다. 실제로 클라우데라(Cloudera)와 타블로(Tableau)는 이미 자사의 서비스에 하둡을 이용하고 있다.

업그레이드된 RDBMS

업계 관계자들은 관계형 DBMS를 업그레이드했다고 해서 이를 빅 데이터 기술로 보기는 어렵다고 입을 모은다. 그러나 올롭슨은 "더 빠르고, 더 크고, 더 저렴해야 한다는 3가지 요소들을 충족한다고 생각한다"라고 반박했다. 예를 들어 테라데이타는 자사의 시스템을 사용자들이 좀 더 쉽게 이용할 수 있도록 만들었으며 이는 확장할 수 있는 클러스터화된 환경이라고 그는 덧붙였다.

하지만 다른 이들은 이에 동의하지 않았다. "일반적으로 사용자들이 RDBMS와 표준 BI 툴들을 이용해 처리하는 과정은 진짜 빅 데이터라 할 수 없다"라고 가트너의 데이터 관리 애널리스트 마커스 콜린스는 지적했다. 그런 처리과정은 과거에도 존재했기 때문이라고 콜린스는 덧붙였다.

누가 빅 데이터를 분석하고 있나

1년 전만 해도 빅 데이터 기술의 주된 사용자는 클릭스트림(Clickstream) 데이터를 분석하고 싶어 하는 페이스북이나 야후같은 웹 기업들이었다. 하지만 이제는 "웹 영역 이외에 빅 데이터를 보유하고 있는 기업들도 여기에 동참하고 있다"라고 콜린스는 말했다. 그에 따르면, 은행, 공공 사업, 정보기관 등 모두가 빅 데이터를 이용하고 있다.

그러다 SNS 업체들이 소셜 네트워크 서비스를 만드는 기술로 빅 데이터를 사용하기 시작하면서 해당 기술을 보유한 기술자들이 속속 프로젝트에 투입됐다.

산업별 주요 기업들도 자신들의 가치가 기존에 생각했던 것과는 달리 정보에 기반한다는 것을 깨닫기 시작했으며 이를 통해 빅 데이터 기술이 빠르게 관심을 끌게 될 것이라고 올롭슨은 전망했다.

TV광고 · 보험사기 · 전력사용도 분석

뉴욕에 있는 TRA는 가정의 TV와 DVR 광고가 실제 구매행동으로 이어지는지를 분석해 TV 광고의 가치를 측정하는 기업이다. 이 기업은 그 상관관계들을 파악하기 위해 케이블 회사의 DVR과 식료품점의 고객 카드 프로그램으로부터 데이터를 수집하고 있다. TRA의 빅 데이터 시스템은 170만 가정의 초당 시청 습관을 담은 데이터를 처리한다. 이는 빅 데이터 기술 없이는 불가능했을 수치다. TRA는 코그니토(Kognito)의 WX2 데이터베이스를 도입했다. 이 솔루션은 데이터를 신속하게 읽어 들여 프로필을 작성하고 분석하며 DVR로부터 세부적인 광고시청 정보를 수집하고 이를 매장 정보와 통합해 보고서를 생성해 준다.

"코그니토는 인메모리 솔루션(In-memory Solution)을 보유하고 있기 때문에 전체 데이터의 절반을 메모리화 할 수 있다. 이는 고객들이 쿼리를 실행할 때 수 초 만에 대응할 수 있다는 것을 뜻한다"라고 TRA의 CEO 마크 리버맨은 말했다.

이 데이터베이스는 상용 하드웨어에서 구동되며 TRA는 닷넷 비주얼 스튜디오(.Net Visual Studio)에 내장된 프론트 엔드 애플리케이션을 사용한다. "우리는 여전히 MySQL을 일부 사용하고 있다. 그리고 사용자 인터페이스는 데브익스프레스(Dev-Express)로 개발했다"라고 리버맨은 덧붙였다.

그는 빅 데이터가 700억 달러 규모의 TV 광고 구매사업에 혁신을 불러올 수 있는 잠재력이 있다고 전했다. 전통적인 방식으로 시청률을 측정하려면 표본 추출을 위해 전국적으로 적어도 2만 가구에 특수 셋톱 박스를 설치해야 했다. 지금은 250만 개의 DVR과 케이블 박스로부터 얻은 정보를 세세하게 분석할 수 있게 됐다.

"광고주들에게 TV가 훌륭한 광고 수단이라는 확신을 주며, 700억 달러에 대한 효과를 입증해줄 것이다"라고 리버맨은 말했다. "이것은 큰 발전이다. 전적으로 빅 데이터 분석으로 가능한 것이다"라고 그는 덧붙였다.

애버딘 그룹(Aberdeen Group)의 그렉 벨킨은 TRA 및 다른 기업들이 사용하는 툴 역시 속도, 볼륨, 다양성 요건을 충족시켜 빅 데이터 솔루션으로 봐도 무방하다고 전했다. 벨킨은 "소셜 미디어 웹사이트, DVR 박스, 식료품점 회원카드 데이터 등 미쳐 발굴하지 않은 데이터 출처들만 해도 부지기수다"라고 말했다. 그에 따르면, 전통적인 데이터베이스 방식으로 분석하기에는 데이터가 너무 거대하고 복잡해 유통기업들은 빅 데이터 플랫폼을 도입하는 추세다.

빅 데이터 기술은 카탈리나 마케팅(Catalina Marketing)에도 혁신을 가져왔다. 플로리다 세인트 피터즈버그에 있는 카탈리나 마케팅은 미국 내 1억 9,000만 명 이상의 식료품 구매자들이 지난 몇 년 동안 일으킨 구매이력 데이터를 보관하고 있다. 이 데이터 량은 2.5PB에 달한다. 이 기업에서 단일 최대 규모의 데이터베이스에는 4,250억 줄의 데이터가 저장돼 있다. 또 하나의 데이터베이스에서 하루 6억 2,500만 줄의 데이터가 새로 생성된다.

카탈리나는 이 데이터를 분석해 주요 소비재 기업들과 대형 수퍼마켓 체인점들에게 누가 새로운 제품에 관심을 보이며 어떤 물건을 살 지 예측할 수 있는 정보를 제공하고 있다.

카탈리나의 부사장 겸 CIO 에릭 윌리엄스는 "데이터를 기술에 적용하는 것이 아니라 기술을 데이터에 적용하고 싶었다"라고 말했다. 다음은 그의

설명이다.

"SAS같은 마이닝 업체들이 자사의 분석 기술을 데이터베이스로 이동시키는데 빅 데이터 기술을 사용하고 있다. 이는 기업 전체를 엄청나게 변화시켰다. 이전에도 이런 일을 해 왔지만 원하는 바를 얻기에는 한계가 있었다. 우리는 자체 개발한 툴을 사용해야 했고 자체 개발한 툴들마저도 원하는 바를 이루기에는 역부족이었다. 빅 데이터 기술을 전면적으로 도입하면서 기업 전체에 변화가 생겼다."

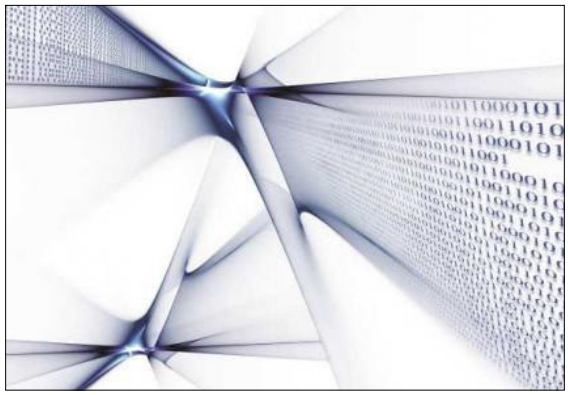
카탈리나는 자사의 시스템에서 구동하는 오픈소스 소프트웨어와 네티자 데이터 저장 애플리케이션 플랫폼을 사용하며 여기에 SAS 애널리틱스(SAS Analytics)를 얹어서 데이터를 분석하고 있다.

"기업들은 DBMS와 DW어플라이언스 위에 SAS 애널리틱스같은 별도의 솔루션을 얹어서 사용할 수 있도록 범용 하드웨어에서 구동되는 기술을 개발하고 있다"라고 윌리엄스는 말했다. "이 기술을 채택해 데이터베이스에서 직접 운영할 수 있다면, 수 주씩 소요됐던 데이터 마이닝 솔루션을 몇 시간 안에 끝낼 수 있을 것이다"라고 윌리엄스는 전했다.

2010년 10월 하둡 월드(Hadoop World)에서 전 뱅크 오브 아메리카(Bank of America)의 빅 데이터 및 분석 담당 상무였던 아브히쉑 메타는 빅 데이터는 기본적으로 뱅크 오브 아메리카의 업무처리 방식을 바꾸고 있다고 밝혔다. "지금의 하둡은 20년 전의 리눅스와 같다고 본다. 우리 모두는 리눅스가 기업 소프트웨어 영역에 어떻게 기여했는지 목격했다. 실로 엄청난 파급효과를 몰고 왔다. 하둡도 마찬가지일 것이다. 실제로 가능하냐가 아니라 언제 가능하냐의 문제일 것이다"라고 메타는 강조했다.

클릭스트림 및 거래 분석을 넘어 하둡은 뱅크 오브 아메리카가 업무 문제를 신속하게 해결하는데 기여했다. "이제 사기행각을 예방할 수 있게 됐다. 견본을 추출하고 모델링 하는데, 이 모델링에 문제가 발생하면 모델링을 새로 개발하는 대신, 사람의 5년 전 사기이력까지 조회할 수 있는 모델링을 개발할 수 있게 됐다. 번거롭던 시대는 끝났다"라고 메타는 말했다.

유틸리티 업계도 자사가 보유한 데이터를 통해



얻을 수 있는 이점을 깨닫기 시작했다. 미국 중서부의 한 유틸리티 시설은 하둡을 이용해 '스마트 미터(Smart Meter)'에서 입력된 값을 분석하고 있다. 그리고 케이블의 암페어(Amperage) 변동 등에 관한 데이터를 수집하고 있다. 올롭슨은 "이런 정보를 수집해 그 추이를 살펴보면 어떤 변압기에 문제가 발생할지 미리 알 수 있다. 또한 정전이나 암페어 변동이 발생하면 고객이 연락하기 전에 그 위치를 찾아내 조처한다"라고 말했다.

그는 앞으로 유틸리티들이 고객 서비스를 향상시키기 위해 빅 데이터를 사용하고 전기 그리드(Electrical Grid) 모니터링을 통해 운영비용을 절감하면서 문제를 발견하고 그리드를 미세하게 조정할 수 있는 능력도 갖게 될 것으로 보고 있다. 단, 여기에는 노후 인프라에 대한 상당한 수준의 업그레이드가 필요할 수 있다.

브랜드 마케팅 담당자들은 하둡을 이용해 소셜 미디어 부문의 '정성 분석'을 테스트하고 있다. 또 트위터를 분석해 제품에 관해 어떻게 말하고 생각하는지 알아내는데 하둡을 사용하는 소셜 분석 서비스 제공업체들이 속속 등장하고 있다.

주의 사항

빅 데이터 기술은 급속도로 진화하고 있다. 그러나 주의해야 할 점도 있다. 먼저 인력이 문제가 된다. 현재 빅 데이터 기술을 사용하고 있는 기업들은 IT 전문인력을 확보하고 있다.

"내부 전문가들이 많지 않다면, 클라우드같은 서비스 제공업체를 이용하라. 아니면 다양한 소프트웨어 제품과 서비스가 제공될 수 있는 시점까지 기

다리라"고 올롭슨은 제안했다.

또 데이터 마이닝은 끝없이 변화하고 있지만 빅 데이터 기술이 현재의 데이터 저장 및 데이터 마이닝 툴들을 완전히 대체하지는 못할 것이라는 전망도 있다.

가트너의 콜린스는 "현재의 데이터 마이닝 기술은 적은 데이터를 이용하기 때문에 정교한 모델을 구축할 수 있다"라고 말했다. 콜린스에 따르면, 현재 빅 데이터는 엄청난 양의 데이터를 제공하기 때문에 더 이상 정교한 모델이 필요치 않을 수도 있으며 이는 데이터 마이닝을 처리하는 방식의 변화를 의미한다.

빅 데이터는 스토리지에도 영향을 미칠 것으로 관측된다. 올롭슨은 "빅 데이터가 결국은 스토리지 시장을 확장시키게 될 것"이라고 주장했다. "하둡의 상용화나 확장에 관계없이 맵리듀스 같은 기술을 이용해 이전에는 절대 가질 수 없었던 흥미로운 기업 정보 데이터를 생성할 것이다. 그러면 이를 재사용하고 패턴을 추적하기 위해 데이터를 저장할 것이다. 결국 데이터의 스토리지 사용을 확대하게 될 것이다"라고 올롭슨은 설명했다.

빅 데이터를 어떻게 활용할 지에 대한 매뉴얼이 필요하다는 주장도 제기됐다. 콜린스는 "이 기술을 도입하고 사용하는데 아직 정해진 패턴이 없기 때문에 사례를 통해 배우게 되며 또 다른 문제들도 초래할 수 있다"라고 지적했다.

빅 데이터 관련 기술들이 발전하고 있지만, 사용성에서 좀더 쉽게 변모해야 한다는 의견도 등장했다. 콜린스는 "패키지 형태의 툴들이 나오면서 일부 기술 이슈들을 해결하고 있지만 그 기술이 여전히 프로그래밍 인터페이스(Programming Interface)에 가깝기 때문에 기업 정보수집 활동에서 오히려 걸림돌일 수 있다"라고 전했다. "하둡은 매우 기술적인 시스템이다. 그리고 BI는 사용하기 쉽고 편리한 방향으로 발전해 왔다"라고 그는 덧붙였다.

또한 콜린스는 빅 데이터 기술이 IT부서 이외의 현업 사용자들에게도 제공돼야 효과가 극대화될 것이라고 강조했다. CIO

●Stacy Collett는 컴퓨터월드에 IT기사를 기고하는 프리랜서다.

빅 데이터 수혜 직종 '데이터 과학자' FAQ

편집부 | CIO 코리아

BI나 DW에 IT예산을 쏟아 붓고도 원하는 분석 결과를 얻지 못하는 기업들이 부지기수다. 이유는 무엇일까? 분석 도구가 문제일까? 이에 대해 업계 전문가들은 데이터 분석 결과를 읽고 이를 해석할 수 있는 능력을 지닌 '데이터 과학자(Data Scientist)'가 없다는 것이 좀더 근원적인 문제라고 입을 모은다.

데이터 과학자란 통계 모델을 사용해 거대한 정보로부터 쓸만한 결론을 이끌어 내는 전문가를 의미한다. 아직 데이터 과학자들에 대한 수요는 그리 많지 않다. 그러나 데이터 마이닝 및 데이터 과학 분야의 전문가들은 이러한 추세가 곧 변할 것이라고 예측하고 있다.

포레스터 리서치의 수석 애널리스트 브라이언 홉킨스는 기업들이 점차 방대한 양의 정보를 잘 사용하기 위해서는 기존의 데이터 마이닝이나 BI만으로는 부족하다는 점을 깨달을 것이며, 이에 따라 데이터 과학자들에 대한 수요가 늘어날 것이라고 예측했다.

정보 관리 기술에 초점을 맞춘 전사 아키텍처(EA) 분야의 전문가인 그는, "기업들은 항상 경쟁사보다 더 많은 정보를 얻고자 한다. 개중에는 예측 기능이 있는 분석 도구를 사용하면 데이터에 대한 이해가 저절로 따라올 것이라고 생각하는 사람도 있는데, 이는 잘못된 생각이다. 데이터 과학자들이 있어야 '통계' 모델을 만들고 사용할 수 있으며 그 결과를 사람들에게 이해시킬 수도 있다"라고 말했다.

데이터 과학자가 부상하는 이유로는 시대적인 배경도 있다. 기존 BI 소프트웨어 한계와 함께 병렬 컴퓨팅의 발전, 정교한 데이터 모델링 도구의 출현 등이 그것이다. 방대한 데이터를 축적할 수 있고, 이를 분석할 수 있는 처리 능력과 도구가 등장함에 따라 이를 다룰 수 있는 인력이 필요하게 된 셈이다.

현재 데이터 과학자들이 분석에 사용하는 하드웨어와 소프트웨어들은 높은 수준의 완성도를 보여준다. 가격 또한 과거와 비교할 수 없을 정도로 저렴하다. 다양한 기업에서 다양한 용도로 사용할 수 있는 기반이 마련된 것이다. 데이터 과학자의 역할과 이들을 특히 필요로 하는 산업 분야, 데이터 과학자에게 요구되는 역량과 지위 등에 대한 FAQ를 정리했다.

Q. 데이터 과학자란 정확히 무엇인가?

A. 급격히 발전한 컴퓨터의 분석 및 모델링 기술을 활용해 대용량 데이터(Big Data)로부터 의미 있는 통찰을 이끌어낼 수 있는 전문가를 의미한다. 빅 데이터에 데이터 마이닝 기술, 통계 기술, 모델링 기술 등을 적절하게 적용해 기업에게 필요한 결론 및 정보를 도출해내는 작업을 한다. 소셜 네트워크 상의 방대한 비정형 데이터로부터 트렌드나 패턴 등을 발굴하는 것이 한 예다.

Q. 특히 어떤 종류의 기업과 분야에서 데이터 과학자를 필요로 할까?

A. 빅 데이터를 입수할 수 있거나 다루는 산업은 모두 해당된다. 현재로서는 주로 거대한 양의 소비자 정보를 관리하는 대기업에서 그 수요가 크다. 앞으로는 웹 기업이나 광고 업체, 통신 업체 등으로 확장될 전망이다. 영업이나 마케팅 데이터를 추적하는 소기업들도 데이터 과학자에 대한 필요성이 증가할 것으로 예상된다.

Q. 데이터를 분석하는 직업은 과거에도 있었다. 새롭게 등장한 직종으로 볼 수 있을까?

A. 코카콜라나 P&G, 나이키 등과 같은 소비재 기업들은 과거에도 리서치 전문가나 통계 전문가를 두고 소비자의 데이터를 분석하려고 했다. 즉 새롭게 등장한 직종은 아니다.

관건은 활용 분야다. 인터넷과 컴퓨터가 보급된 이후 양상이 달라졌다. 소비자, 센서, 인터넷 접속 기기 등으로부터 데이터가 폭증하고 있으며 분석 기술들도 더 다양해지고 있다. 과거 직접적 소매 데이터를 활용할 수 있는 산업에서만 필요로 하던 기술들을, 이제는 더 다양한 분야의 기업들이 사용하려는 추세다. 즉 분석할 데이터가 있고 이를 가능케 하는 도구들이 있다. 단순히 지난 4분기나 월이나 년 단위의 실적을 알고 싶은 것이 아니라, 전체적인 소비 패턴을 알고 왜, 어떻게 그런 패턴들이 생겨났는지를 알고 싶어하는 것이다.

Q. 기존 BI나 DW도 바로 이런 측면에 초점을 맞췄던 것 아닌가? 데이터 과학자에 대한 수요가 증가한다는 것은, 이들 기술의 실패를 의미하는 것인가?

A. 실패보다는 한계라는 표현이 더 적합하다. BI는 일반적으로 과거 데이터의 합산을 통해 과거의 특정 지역에서, 특정 기간 동안, 특정 그룹의 소비자들에게 어떤 일이 있었는지를 분석한다. 물론 기업 운영에 있어 대단히 중요한 정보들이다.

그러나 기업들이 이제 원하는 것은 트렌드에 대한 이해와 소비자 욕구에 대한 예측이다. 기업들이 이러한 필요성을 느끼게 된 데에는 BI의 역할이 크다. BI는 이런 측면에서 결코 실패작이 아니다. 사업 데이터를 분석하기 시작하면, 데이터 분석의 유용성을 실감할 수 있기 때문이다. 그리고 더욱 나은 방안을 찾아 나서게 된다.

DW 역시 한계가 있다. 일단 규제가 너무 심하고 느리다는 것이다. 또 너무 크고 다루기 힘들다. DW는 주로 정확한 숫자가 요구되는 컴플라이언스 리포팅이나 연말 리포팅에 사용된다. 반면 데이터 과학자들이 자유롭게 데이터를 탐색하기에는 적합하지 않다. 예측성 질문에 대한 답변을 필요로 한다면 좀 더 융통성 있는 데이터 소스가 필요하다.

Q. 데이터 과학자가 주목 받는 이유에는 경제적 요인도 있을까? 예를 들어 신성장 동력 발굴에 대한 필요 및 경쟁 격화, 비용 절감 압박 등의 이유 등이 요인인가?

A. 관련이 있다. 최근 몇 년간 기업들의 방향성 중 하나는 최적화였다. 직원과 예산이 한정된 상황에서 성장을 위해서는 최적화가 필요하다. 이는 대단히 복잡하고 어려운 작업이다.

기업들은 이를 위해 올바른 정보가 필요했고 비즈니스 프로세스에 대한 깊은 이해가 요구됐다. 데이터 분석에 핵심적인 요건들이다.

기술의 발전도 한 요인이다. 10여 년 전만 해도, 데이터 분석에 사용되는 도구는 기본적인 것들이었다. 엑셀을 사용할 때도 있었으며, 확장이 어려운 통계 도구들이 주류를 이뤘다.

지금은 확장이 가능하고 훨씬 더 고도화된 통계 도구들이 대거 등장한 상태다. 데이터 과학자가 제대로 일을 할 수 있으려면 병렬 조작이 가능한 컴퓨터와 정교한 모델링 도구, 빠른 데이터 조작 처리 기술이 있어야 한다. 이러한 조건들이 제대로 갖춰지기 시작한 시점은 최근 10년 안쪽이다.

Q. IT 전문가와 데이터 과학자의 접점은 무엇인가? IT 전문가가 데이터 과학자가 되기 위해서는 어떤 역량이 필요한가?

A. 일단은 통계와 모델링, 그리고 수학적인 지식이 있어야 한다. 가장 기반이 되는 역량이다. 실제로 물리학자들이 우수한 데이터 분석 역량을 보이는 경우가 많은데, 실제 상황에 수학을 적용시켜 모델

을 만들어 내는 데 익숙하기 때문이다. 생물정보학을 연구하던 사람들도 데이터 과학자에 적합한 경우가 많은 것도 같은 이유다.

데이터 과학자에게 필요한 다른 역량으로는 사업 분야나 업종에 대한 깊은 이해다. 데이터 과학자들은 가설을 세우는 것으로 일을 시작한다. 그 분야에 대한 충분한 경험 없이는 가설을 세울 수 없다.

다른 사람들 말에 귀 기울이는 역량도 필요하다. 소프트웨어 엔지니어들이 고객의 요구를 잘 알고 있어야 하는 것처럼, 데이터 과학자들도 경영진이나 관리직 임원들의 질문을 잘 듣고 답할 수 있어야 하기 때문이다.

좋은 데이터 과학자가 되기 위한 또 다른 요건은 '데이터를 얼마나 잘 가시화시켜 설명해내는가' 다. 대부분의 기업 경영진들은 데이터 과학자들이 내놓는 정보를 다 이해하기에는 역부족이다. 데이터 과학자들은 사업에 가장 필요할 만한, 가장 큰 영향력을 가질 만한, 그리고 실제로 사업 구상이 가능한 정보를 추려내야 한다. 또 결과를 시각적으로 설득력 있게 표현해 내는 것도 중요하다.

Q. 데이터 과학자들이 기업의 IT부서에서 일하는 경우가 있을까?

A. IT부서에서 일하는 경우도, 다른 부서에서 근무하는 경우도 있다. 확률적으로는 IT부서보다는 마케팅이나 재무 등의 실용 부서가 높다. 가끔은 IT와 비즈니스의 중간점이라고 할 수 있는 연구 부서에 있는 경우도 있다. 데이터 과학자의 필수 역량 중 하나가 사업에 대한 이해라는 점을 반영하는 부분이다. CIO

빅 데이터가 가져올 IT업계의 변화

Tim Lohman | Computerworld

빅 데이터의 등장은 IT업계에 새로운 변화를 가져올 것으로 관측된다. 과거에는 하지 못했던 데이터를 기반으로 한 행동 분석이 가능해 지면서 파생 기술을 가진 기업들의 출현도 기대된다. 업계 전문가들에 따르면, 현실 세계의 모든 일상 데이터를 마이닝하고 각 데이터에 꼬리표를 붙여 관리하게 될 것이다. 또한 데이터 보호 수준을 전문적으로 평가하는 기관이 등장할 것이며 데이터 폭증에 따른 스토리지 수요도 함께 늘어날 전망이다.

일상에서 얻는 데이터 '현실 마이닝'

스토리지 네트워킹산업 협회 호주/뉴질랜드 지부(Storage Networking Industry Association A/NZ)와 컴퓨터월드 호주 지부가 공동으로 개최한 행사 '정보 인프라 구축 심포지엄(Implementing Information Infrastructure Symposium)'에서 연설을 맡은, 전략 조언가이자 작가며 미래학자인 로스 도슨은 '현실 마이닝(Reality mining)'을 언급했다. 그가 말한 현실 마이닝이란, 사람들의 행동 데이터를 수집하는 것을 뜻한다. 현실 마이닝은 빅 데이터에서 탄생했으면서 동시에 데이터 용량을 늘리는 데 기여하는 주요 트렌드라고 그는 밝혔다.

그는 "사무실 환경을 둘러보면, 살펴볼만한 정보들이 엄청나게 많이 있다. 사람들은 어떠한 행동을 취하는가, 그들은 어디를 보고 있으며 어떠한 대화를 나누는가, 서로 이야기를 나눌 때 얼마나 많이 웃는가, 등등이 예가 될 수 있다"라고 설명했다.

"일상에서 쏟아내는 데이터들은 단 몇 시간 만에 TB 규모가 될 수 있다. 그 데이터는 생산성을 올리기 위해, 협력을 증진시키고 조직 내에서 가치를 창출해낼 수 있는 새로운 방법을 고안하기 위해 수집된다"라고 그는 덧붙였다.

도슨은 데이터를 기반으로 의사결정을 내리는 기업들이 그렇지 않은 기업들보다 생산성이 5~6% 높다는 연구 결과를 인용하며 "이러한 종류의 현실 마이닝을 더욱 강조해야 한다"라고 말했다.

로스 도슨이 현실 마이닝과 빅 데이터와의 관계를 설명하고 있다.

비정형 데이터에 이름표 붙이는 '데이터 태깅'

비정형 데이터를 대량으로 수집하고자 하는데 있어 데이터 그 자체뿐 아니라 수집 시점의 정보까지도 함께 저장해야 한다.

그는 "우리가 가지고 있는 대다수의 정보는 정형화되지 않은 것이기 때문에 소스에서의 데이터 태깅이 매우 중요하다. 데이터 태깅(Data Tagging)은 데이터가 폭발적으로 증가할 때, 데이터를 분류해 좀더 체계적으로 저장하는 것이다. 데이터 태깅은 비정형 데이터가 생성되는 것과 동시에 이뤄져야 한다. 이를 차후로 미루면 비용이 훨씬 많이 든다"라고 전했다.

도슨은 데이터 태깅 비용에 대해 비디오의 예를 들어 설명했다. "영상물 안에서 말하는 사람, 주제, 말하는 사람의 감정까지도 태깅돼야 현실 마이닝에서 사용될 수 있다"라고 도슨은 강조했다.

그에 따르면, 그러한 태깅 작업은 주로 기계가 처리하겠지만 '인공 지능'과 같은 기능을 적용시키려면 결국 전문 인력이 필요하다. 또는 아마존의 '미

케니컬 터크(Mechanical Turk)' 와 같은 서비스들을 통해서도 가능하다.

데이터 평가 기관 등장

도슨은 데이터 태깅의 걸림돌을 퍼블릭 클라우드의 보안 문제로 보고 있다. 그는 "미국의 신용 평가 기관인 S&P(Standard&Poor's)나 무디스처럼 클라우드 제공업체들의 보안 순위를 매기는 전문 평가 기관들이 등장할 것"이라고 내다봤다.

그는 "'평판의 경제(reputation economy)' 와 일맥상통한다"라고 운을 뗀 후, "지난 10년 동안 우리가 하는 모든 것들을 움직인 것은 평판이었다. 개인에 대한 평판, 소셜미디어에 대한 사람들의 영향력 같은 것들 말이다"라고 말했다. "클라우드 기업이라는 평판은 좋은 이미지가 아니다. 클라우드에도 S&P나 무디스와 같은 평가 기관들이 필요하다"라고 도슨은 설명했다.

전문 평가 기관들과 마찬가지로 클라우드를 기반으로 분산 데이터베이스 등과 같은 서비스를 제공하는 기업들이나 저장공간이나 처리 능력을 거래하는 조직인 클라우드 브로커(Cloud broker)들도 등장할 것으로 도슨은 전망했다.

플래시 스토리지, 대량으로 등장

행사에서 연설을 맡았던 또 한 사람인 에너지 전략 그룹(Enterprise Strategy Group)의 설립자이자 수석 애널리스트 스티브 듀플레시는 플래시 기반의 스토리지가 향후 12개월 안에 빠르게 시장을 잠식하게 될 것이라고 말했다.

그는 "지난 15년간 스토리지는 오로지 속도에만 관심을 기울였고 지금까지는 그것이 별 문제가 되지 않았지만, 지금 대부분의 사람들이 서버 공간을 최소 40% 이상 가상화 하고 있다. 서버 가상화는 그동안 도외시했던 입출력 증가 문제를 야기시킬 것이다"라고 설명했다.

다음은 듀플레시가 강조한 내용이다.

"기업들은 얼마 전까지만 해도 디스크 드라이브를 구매해서 사용했다. 하지만 이제는 별 가치 없는 1,500rpm의 파이버 채널(fibre-channel) 디스크 드라이브를 사고 싶어하지는 않을 것이다. 데이터

스티브 듀플레시는 "스토리지의 속도뿐 아니라 용량 자체가 이슈가 될 것"이라고 설명했다.

센터에 있는 모든 것들은 디스크 드라이브만 제외하고는 솔리드 스테이트(solid-state)다. 이는 매우 흥미로운 점이다."

듀플레시는 기업들이 솔리드 스테이트를 구매하지 않는 유일한 이유에 대해 '비용 때문' 이라고 밝혔다.

그는 "만약 기업에게 1,000달러짜리의 TB 장비와 1,000달러짜리의 SSD를 살 수 있다고 알려주면 그 기업은 아마도 SSD를 구매할 것이다. 머잖아 그런 상황이 올 것이다"라고 말했다.

기업들이 모든 지난 데이터에 같은 가치를 부여하면서 데이터 백업은 논란 거리 중 하나가 됐다.

듀플레시는 "손목시계보다 더 적은 전력을 쓰는 효율적인 컴퓨터 전력으로 인간을 달로 보내고 다시 되돌려오는 법을 알아냈지만 백업에서는 저전력 솔루션을 아직 만들지 못했다"라고 덧붙였다. CIO

일상 생활 모니터링해 미래의 행동 예측

Lucas Mearian | Computerworld

기업들은 빅 데이터로 과거 어느 때보다 더욱 정확하게 고객의 활동, 행동양식 및 위치를 파악해 고객이 앞으로 어떻게 행동할 지를 예측할 수 있을 것이다 예측의 정확성은 얼마나 많은 데이터를 분석했느냐에 달려 있으며 기업들은 고객들의 사소한 움직임까지도 놓치지 않으려 하고 있다.

IBM의 수석 엔지니어인 제프 조나스는 "우리들이 기존에 가졌던 프라이버시에 대한 개념이 바뀌게 된다"며 "이제 감시 사회를 피할 수 없는 정도라 아니라 이에 저항할 수 없을 정도가 될 것"이라고 설명했다. 조나스는 2011년 3월 23일 '스트럭처 빅 데이터 2011(Structure Big Data 2011)' 컨퍼런스에서 수백 명을 대상으로 발표하는 자리에서 이같이 말했다.

진정한 개인화 서비스 기대

기업들은 사람들이 지리적으로 어디에 위치해 있는지 파악함으로써 광고 및 마케팅 정보를 맞춤화해 제시할 수 있다. 예를 들어 특정 고객 한 명이 카리브해의 아열대 기후 휴양지인 아루바(Aruba)에 위치해 있다는 사실을 알고 있다면 뉴욕에 소재한 레스토랑을 소개하지는 않을 것이다. 대신 스쿠버 다이빙 장비나 썬 태닝 로션을 광고해 팔 것이다.

사람들의 위치 정보를 파악하게 되면 잠재 고객에 대한 정확성을 높일 수도 있다. 이름과 생년월일이 같지만 사는 도시가 다른 다섯 명의 고객이 있다고 가정해보자. 이 경우, 위치 정보를 토대로 개인을 구별할 수 있다.

조나스는 이에 대해 "과거 10년 동안의 주소지 기록을 살펴보는 것만으로도 동일 인물인지 아닌지를 판단할 수 있다. 한 사람이 동시에 같은 주소지의 집에서 살 수는 없기 때문"이라고 설명했다.

조나스에 따르면, 미국에서는 매일 6,000억 건에 달하는 온라인 트랜잭션이 발생한다. 그리고 이중 상당수에는 휴대폰이 생성하는 위치 데이터가 담겨 있다. 따라서 무선 통신 기업들은 이러한 데이터를 실시간으로 확보해 관리하고 있다.

이제 기업들은 몇 년 동안의 데이터를 관찰해 사람들이 일상을 어떻게 보내는지, 어디에서 업무를 보는지, 누구와 교류를 하는지를 파악할 수도 있다.

조나스는 "이러한 정보들이 빅 데이터 분석을 위한 방대한 자료"라며 "이런 분석을 통해 87%의 확률을 가지고, 특정 인물이 다음주 목요일 오후 5시 35분에 어느 위치해 있을지 예측할 수 있다"라고 주장했다.

야후의 클라우드 아키텍처 부문 부사장인 토드 파파이오나우에 따르면, 빅 데이터 분석은 기업이 고객들을 감시해 이용하려는 수단이 아닌, 맞춤화된 웹 경험이 될 것이다.

파파이오나우는 "소비자가 관심 갖는 정보를 전달해 준다면, 그게 누가 됐건 개의치 않을 것"이라고 강조했다.

야후는 올해 새로 업그레이드한 검색 엔진인 서치 다이렉트(Search Direct)를 출시했다. 서치 다이렉트는 구글 인스턴트와 유사한 방식으로 과거 검색 이력을 토대로 더욱 풍부한 콘텐츠를 사용자에게 전달한다. 예를 들어 검색 창에 뉴욕(New York)' 이라는 단어를 입력하면, 해당 단어가 포함되어 있고, 사람들이 가장 즐겨 찾는 검색어들이 순서대로 목록에 나타난다. 이 경우 '뉴욕 타임즈' 가 가장 먼저 뜬다.

점점 방대해지는 데이터 창고

기업 입장에서 풀이하자면, 빅 데이터란 대형 DW다. 컴퓨터 시스템 로그 파일, 금융 서비스 전자 거래 정보, 웹 검색 스트리밍, 이메일 메타 데이터,

검색 엔진 쿼리, 소셜 네트워킹 활동 같은 데이터 등을 들 수 있다. 2010년 한해 동안만 1.5 제타바이트(zetabyte)에 달하는 데이터가 생성됐다. 그리고 이중 대부분은 컴퓨터 같은 기계에서 만들어진 것들이다. 클라우드 소프트웨어 공급업체인 조이엔트(Joyent)의 설립자겸 수석 과학자인 제이슨 호프먼에 따르면, 지난해 기업들의 데이터센터 저장돼 있는 이런 데이터의 양은 16엑사바이트(exabyte)에 달한다고 한다.

분석 엔진 업체인 클라우드스케일(Cloudscale)의 CEO인 빌 맥코일은 "현재까지 빅 데이터 분석에는 오프라인 쿼리나 구글이 개발한 '맵리듀스' 알고리즘이 이용되고 있다"라며 "그러나 기업 DW 사용자의 90%는 실시간 분석 방안을 요구하고 있다"라고 전했다.

맥코일은 "경쟁사보다 더 빨리 데이터로부터 정보를 추출하는 기업이 시장을 선점할 수 있다"라고 주장했다.

MPP(Massively Parallel Processing) DW어플라이언스 업체인 네티자의 창업자겸 CEO 짐 바움 역시 맥코일의 의견에 동의했다. 바움은 "만약 기업 사용자들이 분석 쿼리에 많은 시간을 소비하지 않는다면 정보로부터 더 많은 가치를 창출해낼 수 있을 것"이라고 주장했다.

바움은 이에 대해 "실시간으로 답을 얻을 수만 있다면, 다음 질문, 또 다음 질문을 계속해서 물을 수 있다. 이렇듯 실시간으로 답을 얻는 것은 아주 중요하다"며 "이는 우리가 빅 데이터를 이용해 얻고자 하는 것이기도 하다"라고 설명했다.

유명 서점인 반즈앤노블의 부사장 마크 패리시는 "전자책 판매가 본격화되면서 컴퓨터가 생성하는 데이터가 폭증하고 있다"고 말했다. 아마존의 경우, 지난해 전자책 판매가 종이책 판매를 넘어섰다고 밝혔다.

패리시는 "고객들이 e북 리더기와 e북을 이용하는 방식에 대한 웹로그 데이터는 현재 35TB에 달한다"며 "올해에는 25TB가 더 늘어날 전망"이라고 설명했다. 패리시는 "이 데이터를 이용해 고객들의 행동양태를 판단할 수 있다. 예를 들어 좋아하는 작가를 기준으로 책을 구매하는 고객의 비율 같은 정보가 여기에 속한다. 따라서 고객들의 생각을 어떻게 포착하고 이를 어떻게 분석하며 발전시킬 지를 결정해야 한다"라고 강조했다.

다른 기업들 또한 빅 데이터 분석을 이용해 자사 웹사이트 콘텐츠 활용을 추적하고 있다. 고객의 취향에 한층 부합하기 위한 목적에서다.

내셔널 퍼블릭 라디오(National Public Radio)의 통계 분석가 손드라 러셀은 "웹 사이트 사용자의 트렌드를 실시간으로 추적하도록 해주는 수단이 필요하다"라고 말했다. NPR은 웹 사이트를 통해 팟캐스트와 라이브 스트림, 온디맨드 스트리밍, 기타 라디오 관련 콘텐츠를 전달하고 있다. 이 회사는 웹 분석 엔진인 옴니츄어(Omniture)를 이용해 왔다.

하지만 NPR은 쿼리 지연 시간이 짧게는 6~12시간, 최악의 경우 수주 정도 걸리면서 고민에 빠졌다. 결국 로그와 매트릭스, 다른 애플리케이션과 서버, 네트워크의 데이터와 인덱스 검색이 가능하도록 스플렁크(Splunk)의 리포팅 툴을 도입했다.

러셀은 "특정 시간대에 누군가 프로그램을 얼마나 반복해 청취하는지 알고 싶었을 뿐"이라며 "스플렁크를 이용하고 나서 데이터 쿼리와 리포팅이 빨라졌다. 따라서 정확한 수치를 반영한 그래프를 얻기 위해 몇 주간을 기다릴 필요가 없어졌다"고 말했다.

하둡, 카산드라, 맵리듀스의 등장

IBM의 조나스는 빅 데이터를 퍼즐 조각에 비교했다. 책상 위에 올려놓고 조각을 맞추기까지는 그게 뭔지를 알 수 없다는 점에서다. 이는 하둡과 카산드라(Cassandra), 그리고 다른 분석 엔진들이 등장한 이유이기도 하다. 하둡은 구글의 맵리듀스를 기반으로 한 분산형 소프트웨어 파일 시스템으로, 대규모 서버 클러스터 전반에 걸쳐 대용량 연산(배치 프로세싱)을 병렬로 처리해준다. 또 구조적이든 비구조적이든 사용자나 컴퓨터가 생성한 데이터를 대상으로 이런 연산을 처리할 수 있다. 하둡은 비정형 무작위 데이터 세트에서 가장 큰 위력을 발

휘한다. 즉 분석 엔진이 정보를 한층 빠르게 수집할 수 있다는 뜻이다.

맵리듀스 시스템은 기존의 전통적인 데이터베이스와는 다르다. 데이터, 파일, 블록의 형식과는 상관없이 배치 프로세스에서 데이터를 빠르게 사전 분류할 수 있다는 점에서 그렇다. 또 C++, C#, 자바, 펄(Perl), 파이썬(Python), 루비(Ruby) 같은 많은 언어들과 호환이 가능하다.

데이터를 분류한 후, 구체적인 쿼리를 수행하기 위해서는 정교한 분석 애플리케이션이 필요하다. 기존 데이터베이스는 테이블과 테이블을 비교해 분석하기 때문에 속도가 느리다. 또 확장도 어렵다.

예를 들어 구글의 연구 및 특별 이니셔티브 부문 부사장인 알프레드 스펙터의 '서버 클라우드 클러스터는 언젠가 하나의 MPP DW를 생성하는 1,600만개의 프로세스를 포함하게 될 것'이라는 전망도 이와 맥락을 같이한다.

스펙터는 "이를 구현하는데 필요한 좋은 엔지니어링 자원을 제외하고 여타의 제약은 없다. 무어의 법칙과는 상관없이 사실상 무한대의 연산 능력을 보유하고 있다"라고 설명했다.

스펙터는 "분산형 컴퓨팅 시스템이 완전히 투명한 프로세싱을 웹 개발자들에게 제공하는 날이 도래할 것"이라고 예측했다. 빅 데이터 분석 엔진은 언어에 상관없이 사용자의 프로파일 이력을 토대로 파일이나 블록 데이터를 분석하고, 웹 사이트에서 스팸들을 파악해 이를 걸러내는 웹 사이트 조절자로서의 역할을 하는 시스템이다.

스펙터는 "예측 API를 통해 사용자들이 이와 같은 역량을 사용할 수 있게 되기를 바란다. 데이터 세트를 제공하고, 이들 데이터 세트를 토대로 컴퓨터의 알고리즘을 훈련시키는 것이다"라고 설명했다.

현재까지 하둡에 가장 많이 공헌한 기업은 야후다. 야후는 코드의 약 70%를 만들어냈으며 모든 사업 부문에서 이를 활용하고 있다. 또 아파치 하둡을 기반으로 표준화했다.

파파이오나우의 설명에 따르면, 야후는 현재 4만 3,000여 서버를 보유하고 있다. 그리고 이중 상당수는 하둡 클러스터로 설정이 되어 있다. 파파이오나우는 "서버팜의 수가 올해 말까지 6만여 개로 늘어날 것"으로 전망했다. 현재 200PB의 데이터를 저장하고 있는 상태에서 매일 50TB씩의 데이터를 생성하고 있기 때문이다.

파파이오나우는 "우리는 어떤 데이터도 버리지 않는 것을 원칙으로 한다"라고 강조했다. 이는 다른 기업도 바라는 방식으로 '모든 데이터를 분석해 차별화된 경쟁우위로 활용하는 것'을 의미한다. CIO

부상하는 하둡, '아직은 RDBMS와 병행해 사용'

Jaikumar Vijayan | Computerworld

빅 데이터와 관련해 가장 많이 쓰이는 세 가지 기술로는 하둡, 카산드라, 맵리듀스가 있다. 이 가운데 오픈소스 아파치 소프트웨어인 하둡이 각광받고 있는 것으로 알려졌다.

벤타나 리서치(Ventana Research)의 최신 보고서는 하둡이 대세인 것은 분명하나 전통적인 관계형 DBMS(RDBMS) 플랫폼을 대체하는 것은 아니라고 밝혔다. 이 보고서에 따르면, 기업들은 하둡과 RDBMS를 함께 사용하는 것으로 나타났다.

하둡은 기업이 PB급 데이터를 관리하고 처리하도록 설계됐다. 하둡의 강점은 매우 큰 데이터 세트를 더 빠른 프로세싱을 위한 상용 하드웨어의 클러스터로 전달하는 더 작은 데이터 블록으로 나누는 것이다.

페이스북, 아마존, e베이, 야후 등 얼리 어답터들은 전통적인 RDBMS로는 쉽게 관리할 수 없는 PB급 비정형 데이터를 분석하고 저장하는 데 하둡을 사용하는 것으로 조사됐다.

160개 이상의 기업들을 대상으로 조사한 벤타나의 보고서는 점점 더 많은 기업들이 같은 이유로 하둡을 함께 사용하기 시작했음을 보여준다.

조사에 응한 기업의 절반 이상은 대규모 정형 및 비정형 데이터를 분석함으로써 비즈니스 통찰력을 수집하기 위해 하둡을 사용하기 시작했다고 답했다.

"대부분의 기업들은 기존 기술을 대체하기보다는 새로운 기능을 더해 하둡을 이용하고 있다"라고 벤타나 보고서의 저자 데이비드 메닝거는 전했다.

벤타나에 따르면, 하둡을 이용하는 기업들 대다수가 소셜 미디어 사이트의 로그나 이벤트 데이터, 검색 엔진 데이터, 텍스트 및 동영상 데이터와 같은 거대한 규모의 비정형의 기계로 정리된 데이터를 분석하고 취합하는 데 하둡을 주로 사용하는 것으로 나타났다.

"하둡이 전통적인 RDBMS의 능력을 능가하고자 하지만, 아직은 트랜잭션 데이터, 고객 데이터, 전화 기록 데이터와 같은 정형 데이터를 분석하는데 많이 쓰일 것 같지 않다"라고 메닝거는 말했다.

메닝거는 하둡이 새로운 종류의 데이터 분석 능력을 구현하기 때문에 사람들이 이를 사용하는 것이라고 밝혔다. "응답 기업의 2/3가 고도화된 분석과 과거에 하지 못했던 형태의 분석을 위해 하둡을 사용하고 있다고 답했다"라고 그는 덧붙였다. CIO

"DW어플라이언스에 주목해야 하는 이유"

편집부 | CIO 코리아

빅 데이터는 데이터의 양적 증가뿐 아니라 비정형 데이터라는 질적인 변화도 의미한다. 빅 데이터에서 CIO와 IT임원들이 주목해야 할 것은 대용량 데이터를 저장하는 아카이빙 기술과 비정형 쿼리를 분석하는 솔루션이다. 기존의 데이터마트와 DW는 정형 데이터를 저장하고 분석하는데 쓰였기 때문에 비정형 데이터에는 적합하지 않다.

비정형 데이터를 분석하는데 필요한 DW는 바로 DW어플라이언스다. 이는 DW를 위해 최적화시킨 솔루션으로 속도와 성능을 유지하면서 대용량 데이터를 처리하기 때문이다.

DW어플라이언스는 기존 IT투자를 보호해 준다는 장점도 있다. 기업이 빅 데이터를 분석하기 위해 기존 IT인프라를 버리고 완전히 새롭게 구축하는 것이 아니라 기존의 인프라와 DW어플라이언스를 함께 사용할 수 있다.

기업이 현재 사용하는 인프라가 데이터마트라면, 이를 폐기하지 않고 DW어플라이언스를 추가로 도입해 사용할 수 있다. DW를 사용하고 있는 기업도 마찬가지다. 신규 IT투자 규모를 줄이면서 성능과 속도 향상, 더 나아가 과거에 하지 못했던 비정형 데이터의 분석까지도 가능하게 해주는 솔루션이 바로 'DW어플라이언스'다.

우리캐피탈

빠르고 정확한 데이터 분석으로 경영 전략 강화

한국IBM

우리캐피탈은 데이터 경영을 위해 빠른 데이터 처리 속도는 물론 관리 편의성을 갖추면서 비용 절감 효과까지 제공하는 DW 어플라이언스 제품에 주목했다. 우리캐피탈은 네티자 NPS10100을 도입해 대용량 데이터를 빠르고 정확하게 분석할 수 있게 됐다. 그리고 이를 적재적소에 활용함으로써 비즈니스 경쟁력을 높였다.

Project Overview

- **도입 배경** | 전사데이터웨어하우스(EDW) 시스템 구축
- **도전 과제** |
  - 많은 시간을 요구하거나 다량의 데이터를 일괄 처리하는 배치작업이 많아 실시간으로 들어오는 방대한 데이터를 정확하고 빠르게 처리해야 함
  - 관리 포인트가 적은 시스템
  - 현업에서도 손쉽게 활용할 수 있는 시스템
- **도입 효과** |
  - 비즈니스 경쟁력 확보 – 개인신용데이터를 철저하게 분석해 구체적인 비즈니스 전략 수립 및 전략 신뢰성을 확보, 상품 완성도 높임. 고객 성향을 고려한 크로스셀링 영업을 강화로 할부나 리스 쪽 고객을 개인 대출 고객으로 끌어들일 수 있는 기반 마련
  - 업무 성능 향상 – 데이터 분석 시간을 크게 단축돼 (ex. 2박 3일→10분) 데이터 분석에 들였던 시간을 보다 생산적인 업무에 투자
  - 업무 편의성 향상 – 빅 데이터를 적시 처리할 수 있는 고성능 시스템을 통해 물리적인 데이터마트를 별도로 구축할 필요 없음. 빠른 응답 속도로 개발 단계에서 튜닝이나 인덱스 의존도 낮아짐,
  - IT 총소유비용(TCO) 20~30% 절감
  - 관리 효율성 향상
- **도입 제품** | IBM Netezza NPS 10100 (현 IBM Netezza Platform Software)

지난 1995년 주택할부금융사로 출발한 우리캐피탈은, 창립 이후 14년 연속 흑자를 기록한 여신전문 금융기관이다. 2005년에는 대우자동차판매주식회사와 손잡으면서 자동차 할부금융시장으로 진출하기 위한 미래 성장의 발판을 마련했다.

우리캐피탈은 이에 안주하지 않고 경제활동인구 4,000만 명의 신용정보를 기반으로 분석 업무를 수행, 고부가가치 금융상품과 획기적이고 참신한 리스상품을 지속적으로 개발해 나가고 있다.

이를 위해 우리캐피탈은 2009년 4월 네티자의 DW어플라이언스를 도입해 빅 데이터를 처리할 수 있는 EDW 시스템을 구축했다.

우리캐피탈은 앞으로 '분석' 측면에서 진행한 EDW시스템을 위험관리시스템, 종합수익시스템까지 확대해 비즈니스 경쟁력을 확보할 방침이다.

저비용 고성능 최대 만족

과거 주택할부금융서비스만 제공했던 우리캐피탈은 2005년, 대우자동차판매의 자회사로 편입되면서 자동차할부와 리스, 개인 대출 서비스를 함께 제공하게 됐다. 이후 정보시스템에 심각한 문제가 발생했다. 기존 계정계 시스템으론 감당하기 힘들 정도로 데이터 량이 폭주했고, 급기야 시스템 과부하라는 난제에 봉착한 것이다.

정보시스템의 문제는 비즈니스에도 악영향을 미쳤다. 고객 연체 성향 분석해서 채권 회수율을 높이는데 어려움을 겪는 등 다양한 상관분석을 원활히 수행하지 못했다. 이처럼 정보시스템이 비즈니스 확대에 걸림돌이 되는 것을 더 이상 방치할 수 없었던 우리캐피탈은 내부 회의를 거듭한 끝에 EDW시

스템 구축으로 해결책을 찾기로 했다.

이후 다양한 고민을 거듭한 끝에 지난 2009년 EDW 플랫폼으로 네티자 제품을 도입키로 결정했다. 고성능 데이터 분석 시스템이 필요했던 우리캐피탈은 네티자의 12.5TB급 규모의 DW어플라이언스 장비인 NPS10100(Netezza Performance Server 10100)를 도입해 EDW 시스템을 구축했다.

우리캐피탈이 다양한 DW 솔루션 중 네티자 NPS10100을 최종 선택한 이유는 고성능, 저비용이란 두 가지 이점 때문이다.

자체 테스트 결과, 이 제품은 서버, 스토리지, DB를 별도 도입하는 종래 방식 DW와 비교했을 때, 최소 10~100배 이상 빠른 데이터 처리 속도를 보였다. 또한 HW/SW 일체형이기 때문에 설치가 매우 간편하고, 사용과 관리 또한 용이했다. 아울러 오픈 시스템보다 장비의 가격이 상대적으로 저렴한 데다 운영비용 또한 거의 들지 않았고, 응답 속도가 뛰어나기 때문에 개발단계에서 튜닝이나 인덱스에 신경 쓸 필요가 없었다.

이런 네티자 장비의 장점들은 전반적인 TCO(총소유비용)를 크게 줄였는데, 종래 방식 대비 TCO는 30%나 낮은 것으로 조사됐다.

네티자에게 여러 장점이 있었지만, 우리캐피탈에게 DW어플라이언스 적용은 모험이었다. 도입 결정을 내릴 때만해도 국내에선 DW어플라이언스를 기반으로 EDW를 구축한 사례가 매우 드물었기 때문이다.

이런 고민을 덜어준 것은 해외의 선진 성공사례들이었다. 특히 일본의 성공사례는 우리캐피탈이 DW어플라이언스를 남보다 먼저 도입하는데 확신을 갖게 했다.

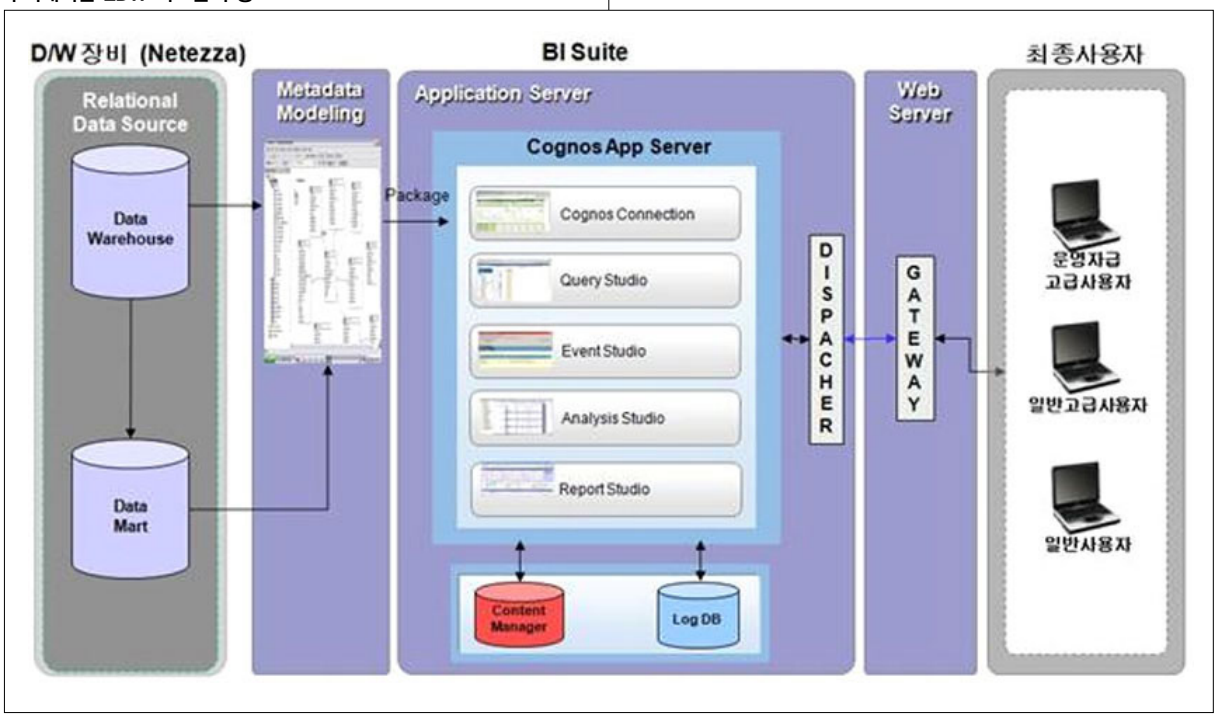
네티자는 까다롭기로 소문난 일본에 이미 60여 군데나 고객사를 확보했고, 일본의 여러 사이트에서 탁월한 성능과 우수한 사용자 편의성, 낮은 TCO 등 DW어플라이언스가 가진 고유한 차별점을 검증 받은 상태였다.

비즈니스 경쟁력 확보

NPS10100을 기반으로 EDW를 구축한 우리캐피탈은 개인신용데이터를 신속 정확하게 분석하고, 이를 적재적소에 활용해 기업의 비즈니스 경쟁력을 확보할 수 있었다.

우선 우리캐피탈은 EDW를 이용해 과거 특정 시점부터 현재까지의 대출 및 연체 추이를 분석할 수

우리캐피탈 EDW 시스템 구성도



있게 되면서, 연체나 대출을 많이 하는 고객군을 정확히 알게 됐다. 그리고 이를 바탕으로 구체적인 비즈니스 전략을 세울 수 있게 됐다. 즉, 과거엔 막연한 추측으로 진행했던 사업들을 이제 확신을 가지고 추진할 수 있게 된 것이다.

이렇게 비즈니스 전략의 신뢰성을 확보한 우리캐피탈은, 상품의 완성도를 높이는 효과를 덤으로 얻게 됐다. 시스템 구축 후 연체 추이, 신용 상담 횟수, 신용 상담 고객 성향 등 상품에 반영해야 할 요소를 폭 넓게 산출할 수 있게 돼, 보다 완성도 있는 상품을 개발할 수 있었다.

또 다른 효과는 데이터마트를 더이상 구축하지 않게 된 점이다. 기존 계정계시스템에선 고객신용정보, 연체성향 등을 분석하기 위해 그 용도에 맞는 데이터마트를 먼저 만들어야 했다. 또한 현업으로부터 추가적인 분석 요구사항이 발생할 때마다 시스템 담당자는 분석결과의 유연성 및 정확성을 확보하기 위해 데이터마트를 계속 생성해야 하는 불편을 겪었다.

하지만 DW어플라이언스 도입 후 이런 문제점은 완전히 사라졌다. 네티자 DW어플라이언스는 빅데이터를 적시에 처리할 수 있는 강력한 기능을 갖춰 굳이 데이터마트를 만들 필요가 없었다. 또한 현업이 새로운 분석 요청으로 데이터마트의 제작을 요구할 경우 관리자는 먼저 쓰던 테이블 위에 뷰만 보태주면 돼, 마트 생성이 원천적으로 필요하지 않았다.

업무 효율 향상

데이터 처리 속도가 빨라지면서 업무 효율이 크게 높아진 점도 주목된다. 금융기관은 다량의 데이터를 일괄 처리하는 배치작업이 많고 해당 작업을 수행하는데 적지 않은 시간을 투입해야 한다.

하지만 이 회사의 계정계 시스템은 빅 데이터 처리에 한계를 보였다. 처리 자체가 아예 불가능하거나, 처리를 한다해도 그 속도가 매우 느려 직원들은 요구사항 하나를 분석하는데 며칠씩 애를 먹었다.

우리캐피탈은 분석 시간을 2박 3일에서 10분으로 크게 줄일 수 있었다. 분석 결과를 신속히 얻을 수 있게 된 우리캐피탈은 다양한 분석 자료를 토대로 새로운 비즈니스 계획을 만드는 등 보다 생산적인 업무에 집중할 수 있었다.

또한 시스템 도입 후 데이터간 상관관계분석이 가능해지면서, 우리캐피탈은 크로스셀링(Cross Selling) 영업을 강화할 수 있었다. 크로스셀링이란 금융기관들이 대형화 · 겸업화에 속도를 내면서 최근 나타난 영업전략으로, 고객이 요구하는 각종 금융상품을 단일 금융회사 창구에서 모두 제공하는 원스톱 서비스를 일컫는다.

우리캐피탈은 2005년 대우자동차판매의 자회사로 편입되면서 크로스셀링 영업을 강화하려 했지만, 기존 계정계 시스템으론 크로스셀링을 할 수 없었다. 크로스셀링을 하려면 개인고객에 대한 성향과 데이터 간의 상관관계를 분석해야 하는데, 데이터 처리 성능이 낮은 종전 계정계시스템에선 불가능했기 때문이다.

EDW 구축 후 우리캐피탈은 체계적이면서 빠르게 고객 성향을 분석하게 됐고, 이를 기반으로 할부나 리스 쪽 고객을 개인 대출 고객으로 유입할 수 있는 기반을 마련했다.

관리 효율성 또한 높아졌다. 우리캐피탈은 EDW를 구축한 후 계정계 관리자들이 정보계를 함께 관리하게 했다. 전통적인 방식의 개방형 DW시스템은 스토리지, 서버, DB를 별도로 두고 관리해야 하기에 관리 인력을 추가로 둬야 한다. 그러나 HW/SW 일체형인 네티자 DW어플라이언스는 관리 포인트가 적어 시스템 관리자를 추가로 쓰지 않아도 된다.

EDW 활용도 높이는 데 주력

우리캐피탈은 앞으로 EDW시스템을 향후 확대 적용할 계획이다. 지금은 개인신용데이터를 분석하는데 머물고 있지만 앞으론 위험관리(RM), 종합수익시스템으로 확대해 비즈니스 경쟁력을 높이고자 한다.

아울러 직원들의 시스템 활용도 높일 방침이다. 우리캐피탈은 모든 직원이 새 시스템을 잘 활용해 필요한 자료를 자유롭게 뽑는 것을 목표로 정했다. 새 시스템에 들어 있는 많은 분석결과물을 직원들이 쉽게 사용하도록 시스템 정비에 주력하고 있다. CIO

우리캐피탈

"데이터 분석 능력이 곧 기업 비즈니스 경쟁력"

interview

남두현 | 우리캐피탈 IT기획팀 팀장

✻금융권에서 데이터 분석 능력이 중요한 이유는? "데이터 분석 능력이 곧 비즈니스 경쟁력인 시대다. 금융회사는 4,000만 명이란 개인신용거래 정보를 기반으로 금융 서비스를 제공하고 새로운 상품을 수시로 만들기 때문에, 신뢰성이 낮은 데이터를 활용하면 비즈니스에 악영향을 받을 수 있다. 고객의 연체 성향을 잘못 분석하면 채권 회수율은 나빠지고, 고객의 성향을 잘못 분석하면 만든 제품의 상품성이 떨어지게 된다."

✻네티자 제품을 도입한 이유는? "고성능과 저비용이라는 제품 선택 요건을 모두 만족시켰다. 자체 테스트한 결과, 네티자 NS10100은 종래 방식 DW 시스템에 비해 데이터 처리 속도가 10~100배 빨랐고, TCO는 30%나 낮았다. 아울러 처리 속도가 빠르다 보니 개발 단계에서 튜닝이나 인덱스에 특별히 신경 쓸 필요가 없었다."

✻도입 결정 당시는 DW어플라이언스가 생소했을 때다. "물론 우리로선 모험이었다. 하지만 일본에선 이미 DW어플라이언스가 기존 DW시스템을 대체하는 트렌드로 정착했다는 소식을 접한 후 제품의 성능 및 안정성에 확신을 갖게 됐다. 특히 네티자는 일본에 이미 60여 군데나 고객사를 확보하고 있었고, 일본의 여러 사이트에서 탁월한 성능과 우수한 사용자 편의성, TCO 등 DW어플라이언스만의 고유한 차별점을 검증 받은 상태였다. 위험 부담이 있긴 했지만 기업의 비즈니스 경쟁력을 고려했을 때, 선택을 미룰 순 없었다."

✻향후 추가 투자 계획은? "네티자 도입 후 지금까지 개인신용데이터 분석에만 썼던 EDW시스템을 향후 RM, 종합수익시스템으로 확대할 계획이다. 또한 단기적으로는 직원들의 시스템 활용도를 높이는 데 주력할 방침이다. EDW를 구축하긴 했지만 아직은 직원들의 활용도가 낮다. 특정 몇 사람이 아니라 모든 직원이 새 시스템을 잘 활용해 필요한 자료를 자유롭게 사용하도록 하고자 한다."

코리아크레딧뷰로(KCB)

18초만에 38억 건 데이터 조회 완료

한국IBM

개인신용평가 전문 기업인 코리아크레딧뷰로(KCB)는 설립 후 지난 5년여 동안 '금융 강국을 만드는 선진 신용 사회 실현'을 모토로, 금융회사에는 보다 입체적이고 과학적인 신용 평가 서비스를, 개인 고객에게는 보다 체계적인 신용 정보 관리 서비스를 제공하기 위해 신용 정보 분석의 고도화를 지속적으로 추진하고 있다. 이 일환으로 네티자의 DW어플라이언스를 도입한 KCB는 고도화된 데이터 분석을 통해 선진 신용 정보 인프라 구축에 한 발 다가서고 있다.

Project Overview

- **도입 배경** | 정보계시스템 구축
- **도전 과제** |
  - 대용량의 다양한 형태 데이터 분석 고도화 및 처리 성능 개선
  - 다양한 요구를 수용하는 유연한 데이터 웨어하우징 및 데이터마트 환경 구현
- **선정 이유** |
  - 방대한 데이터의 빠른 처리 성능
  - 데이터 압축을 통한 효율적인 가용 용량 확보, 유연성과 확장성 탁월
  - SW/HW 일체형의 관리 용이성
- **도입 효과** |
  - 샘플링 아닌 전수 분석 가능
  - 2일에서 1~2시간 이내로 분석 소요 시간 단축
  - 분석 시간 단축에 따른 고도화된 분석 기법 적용 가능
  - 로우 데이터에 직접 접근해 현업 사용자들이 쿼리 실행(유연한 분석 가능)
  - 데이터 압축 기술을 통해 최대 7배의 데이터 용량 절감 및 성능 개선 효과
- **도입 제품** |
  - 2008 : IBM Netezza Platform Software(현 IBM Netezza Platform Software),
  - 2010년 : IBM Netezza 1000-12(현 IBM Netezza 1000-12)

KCB는 우리나라 경제 인구 3,900만 명의 신용을 종합적으로 평가하는 전문 기업이다. 2005년 2월에 선진 신용 정보 인프라 구축을 목표로 국내 19개 대형 금융회사들이 공동 출자해 설립됐다.

기존의 불량 정보 중심의 개인 신용 관리에서 벗어나 선진국과 같은 대출 상환 실적, 카드 사용 실적 등 우량 정보까지 다각도로 분석, 평가해 선진화된 금융 인프라를 구축하고 개인의 신용도 관리를 전문화하자는 것이 KCB의 설립 취지다.

KCB는 정확하고 다각적인 신용 정보 분석을 위해 시장 및 고객 분석, 선진 리스크 관리 기법 등에 대한 다각적인 연구를 수행하고 있다. 이를 기반으로 국내 최초로 포지티브 정보를 포함한 신용평가 방식을 채택해 연체정보뿐만 아니라 대출상환 실적, 카드사용 실적 등 다양한 형태의 우량 정보를 활용해 개인에 대한 균형 잡힌 신용 평가 및 신용 관리를 실현하고 있다. 이러한 KCB의 비전과 목표 실현을 위해 IBM 네티자의 DW 어플라이언스가 채택됐다.

적재에서 분석으로, 선진 신용인프라 구축

그 동안 개인신용평가는 주로 연체 정보 등 부정적인 정보에 의존해왔기에 개인들이 과거 성실히 쌓아온 우량 정보는 제대로 평가 받지 못했다. 이러한 한계를 극복하기 위해, KCB는 부정적인 정보뿐만 아니라 대출상환 실적, 카드사용 실적 등 다양한

형태의 우량 정보를 활용해 균형 잡힌 개인신용평가 방식을 국내 처음으로 도입했다. 아울러 개인 신용도 변화를 파악하고자 동태적 우량정보까지 포괄하는 개인신용평가 시스템을 구축하게 된다.

하지만 KCB는 여기에 만족하지 않았다. KCB는 세계 수준의 선진 신용인프라를 갖추기 위해선 기존의 '적재' 중심의 서비스 제공방식에서 벗어나 '분석' 중심의 데이터 서비스를 고객에 제공해야 한다고 생각했고, 이후 정보계시스템을 구축하기로 결정했다. 정보계시스템을 구축하기 위해 KCB는 1차로 신용거래나 금융거래 등 다양한 개인신용 정보를 확보해야 했다. KCB는 우선 금융회사로 구성된 KCB 주주회사들로부터 축적된 데이터를 제공받았다. 그리고 5년 분량의 개인 거래데이터를 시계열 방법으로 분석했다.

제반 데이터를 확보한 KCB는, 이후 고객의 우량 및 불량 점수평가(스코어링), 권유 상품 분석 등을 진행하기 위해 각각의 용도에 맞는 데이터마트를 만들었다.

기존에 구축한 데이터마트도 있었지만, 이를 그대로 활용할 경우 날짜가 왜곡되거나, 변화하는 시장과 고객 상황을 실시간으로 반영할 수 없는 한계가 있었다. 또한 고객 요청사항이 추가될 때마다, 시스템 담당자는 분석의 정확성을 확보하기 위해 물리 및 논리적 설계나 데이터마트를 반복 생성해야 하는 상황이 발생했다.

KCB가 DW어플라이언스 제품을 도입한 이유는, 로우 데이터에 직접 접근해 데이터 마이닝 작업을 수행하고, 다양한 요구를 수용할 수 있는 다이나믹 마트를 확보하기 위해서다.

다이나믹 마트를 확보하기 위해서는 고성능 및 대용량의 스트리밍 처리가 필요했고, 시장조사 끝에 찾아낸 해답이 바로 DW어플라이언스였다.

KCB는 곧바로 시중에 나와 있는 DW어플라이언스 제품들을 대상으로 벤치마킹테스트(BMT)를 진행했다. 그리고 데이터 처리 성능, 운영 편의성 등을 고려해 총합산한 결과 2008년 국내 업계 최초로 DW 어플라이언스 제품인 네티자 시스템을 구축했다.

KCB는 지난 2008년에 네티자의 DW어플라이언스 제품인 NPS10200을 기반으로 정보계시스템을 구축했다. 이어 2010년 1월엔 NPS 압축 엔진인 'NPS 릴리스 4.5'를 추가로 도입했으며, 2011년에는 NPS의 4세대 서버 모델인 네티자 트윈핀(TwinFin)으로 정보계 시스템을 업그레이드 했다.

KCB는 세계 수준의 선진 신용인프라를 갖추기 위해선 기존의 '적재' 중심의 서비스 제공방식에서 벗어나 '분석' 중심의 데이터 서비스를 고객에 제공해야 한다고 생각했고, 이후 정보계시스템을 구축하기로 결정했다.

KCB가 네티자의 DW 어플라이언스 제품을 도입한 이유는 크게 3가지였다. 방대한 데이터의 빠른 처리 성능, 데이터 압축을 통한 효율적인 가용 용량 확보, 유연성과 확장성 등의 효율적인 운영 가능 등이 그것이다.

우선 KCB의 비즈니스 특성상 방대한 데이터의 수집 및 분석이 매우 중요한데, 기존의 DW 구조로는 불가능했던 대량 데이터의 정보 분석이 가능해졌다. 네티자의 DW 어플라이언스 제품은 비대칭 초병렬 데이터 처리 방식인 'AMPP (Asymmetric Massively Parallel Processing)' 구조를 취하고 있어, 데이터의 입출력(I/O) 과정이 생략돼 데이터 처리 시간이 크게 단축된다.

전수 분석 가능

실제로 KCB는 4,000만 명에 달하는 경제 활동 인구의 이력 데이터를 모두 분석하고 있다. 과거 인프라에서는 전수 분석이 불가능했기 때문에 100만 명 미만을 추출해 분석할 수밖에 없었다. 그러나 네티자의

DW어플라이언스 도입 후에는 샘플링이 아닌 전수 분석 방식으로 전환해 분석 정보의 정확도를 크게 높일 수 있게 됐다. 게다가 전수 분석을 하더라도 추가로 수 차례의 테스트를 수행할 수 있게 됐다.

KCB의 IT 서비스부 장용혁 차장은 "전수 분석이 가능하다는 점은 우리에게 매우 큰 의미가 있다. 일반적인 DW 구조에선 100만 건 이상의 데이터 분석은 사실상 불가능하고 수행한다 하더라고 2일 정도의 시간이 소요된다. 그러나 네티자의 제품은 5,000만 명의 데이터 분석도 2~3시간 내에 처리해 분석 데이터의 품질과 정확도를 100% 보장할 수 있다"라고 설명했다.

KCB에 따르면, 자사가 보유한 데이터 중 최다 건수인 38억 건의 데이터를 조회하는데 14초가 걸렸다. 장 차장은 "이 결과가 네티자 DW어플라이언스의 빅 데이터 처리 성능을 단적으로 보여주는 예"라고 강조했다.

KCB는 네티자의 DW 어플라이언스 솔루션이 가진 데이터 압축 기능에 대해서도 매우 만족감을 나타냈다. 초기 구축 당시 KCB의 데이터 양은 20TB였다. 지난해에 50TB로 용량이 늘어나면서 KCB는 NPS 압축 엔진인 'NPS 릴리스 4.5'를 추가로 도입해 사용 공간을 40% 가량 늘렸다. 이 압축 엔진 소프트웨어 업그레이드만으로 데이터 처리 성능이 30% 향상됐다.

최근 최신 제품인 네티자 트윈핀으로 DW 어플라이언스를 업그레이드한 KCB는 더 높은 압축 효과를 체험했다. 기존 NPS 사용 당시와 비교하면 약 3배, NPS에 압축 엔진을 더해 사용했을 경우와 비교하면 약 2배의 데이터 압축 효과가 있다는 설명이다.

특히 일반 DW 환경과 비교하면, 이 압축 효과는 더욱 극명하게 나타난다. 일반적인 DW 환경을 사용한다고 가정하면 인덱스를 제외하고 150TB, 인덱스를 포함할 경우 200TB의 용량 산정이 필요하다는 것이다. 즉 네티자의 트윈핀을 통해 5~7배의 데이터 압축 효과를 보고 있는 것으로 평가하고 있다.

유연한 사용 환경과 관리 용이성 등도 DW 어플라이언스 도입 시 중요한 검토 요건이었다.

네티자의 DW어플라이언스는 분석 요건에 따라 그때마다 물리, 논리 모델 설계나 데이터마트를 반복적으로 생성할 필요가 없다. 현업의 분석 담당자들은 로우 데이터에 직접 접근해 직접 쿼리를 작성하고 실행할 수 있는 다이나믹한 분석 환경 구현이 가능해졌다.

또한 관리 측면에서도 인덱스를 설정할 필요가 없어서 데이터 처리 속도나 가용 용량 확보 면에서 효율적인 인프라 운영이 가능하며, 데이터 튜닝 포인트가 제로에 가깝기 때문에 데이터 아키텍처 등의 비즈니스 요구 사항을 유연하게 적용할 수 있다. 이 밖에 HW/SW 일체형 제품이라 운영 관리가 수월하고 시스템 개발 과정이 매우 단순해 별도의 운영 인력이나 관리에 필요한 시간도 최소한으로 투자할 수 있다. 정보 서비스 부서도 시스템 운영에 매달리지 않고, 데이터 이관 시간을 줄이는 방법이나 데이터의 가공, 처리 성능을 강화하는 방법을 고민할 수 있는 시간적 여유가 생겼다.

통계 솔루션과 GIS로 확장 가능

2012년 2월에 네티자 DW 어플라이언스 신제품인 네티자 트윈핀으로 정보계 시스템을 업그레이드했다. KCB가 트윈핀으로의 업그레이드를 결정한 이유는 성능 개선도 중요한 요인이었지만, 더욱 중요한 이유는 KCB의 비즈니스 때문이다.

네티자 트윈핀은 네티자의 4세대 DW 전용 어플라이언스로, 가용성, 성능, 관리 용이성 측면에서 모두 이전 제품보다 크게 개선됐다. 그러나 무엇보다 KCB에게 매력적으로 평가된 점은 확장성이다. 트윈핀은 SPSS, i-Class 등의 통계 솔루션뿐 아니라 스파이럴(Spiral) 등의 지리정보시스템(GIS) 등 다양한 소프트웨어 모듈을 선택해 적용할 수 있다.

KCB는 DW의 기능과 성능이 비즈니스 경쟁력에 밀접하게 연관된다. 네티자의 DW 어플라이언스 도입을 통해 샘플 분석이 아닌 전수 분석이 가능해지

고 분석 시간이 기존 1~2일에서 2시간 이내로 단축된 것이 중요한 경쟁력이다. 이는 실제로 회원사와 개인 대상의 온라인 분석 서비스 제공이라는 KCB만의 차별화된 서비스 개발로 연결됐다. 현재 KCB의 회원사는 온라인으로 분석 정보를 신청하고, 1시간 이내에 결과를 확인할 수 있는 서비스를 이용하고 있다.

이처럼 KCB는 새로 도입한 트윈핀에 통계 시스템과 지리정보 시스템을 추가 적용해, 비즈니스 경쟁력을 확보한다는 계획이다. 통계 시스템의 경우, 잠재적 대출 수요 등을 예측할 수 있어 회원사 대상의 다양한 CRM 데이터 분석 서비스를 수행할 수 있다. 또한 GIS는 지역 단위의 성향과 현황을 분석할 수 있는 기반이 된다. CIO

KCB

신규 비즈니스 집중할 수 있는 환경 구현

interview

서정진 | KCB IT 서비스부 부부장

정보계 시스템의 구축 목표는? "비즈니스 관점으로 볼 때 KCB의 사업 방향은 명확했다. 개인신용정보 데이터를 기반으로 고객들이 원하는 의미 있는 정보로 재구성해 제공하는 것이 최우선 목표다. 정보계시스템의 구축 목표 또한 매우 명확했다. KCB는 기존의 배치시스템 환경을 획기적으로 개선하고, 데이터 및 업무규칙 검증의 효율화를 추구했으며, 데이터와 서비스의 일관성 및 정확성 유지함으로써 종래엔 없었던 DW/BI 환경의 토대를 마련하는데 노력을 집중했다."

기존 DW시스템의 한계는? "기존 DW시스템의 개발은 매우 복잡한 과정을 거쳐야 했다. 먼저 질의 처리 및 결과의 최적 경로 비용을 찾아내는 옵티마이저를 설치해야 한다. 또한 DW 프로젝트의 성패를 가르는 물리 및 논리모델링 작업을 진행해야 한다. 이외에도 상당 기간의 테스트 과정과 시간을 들여 최적화된 인덱스 설정하는 것을 비롯해, 업무에 적합한 마트 설계 등의 과정을 거쳐야 한다. 여기서 더 큰 문제는 이런 복잡한 과정을 거쳐도 업무가 정형화됐을 때만 DW 프로젝트의 성공을 보장할 수 있다는 것이다. 결국 과거의 방식으로 DW를 구축하면 현업의 다양한 요구사항을 적절하고 신속하게 반영할 수 없다."

기존 DW 시스템과 대비했을 때 네티자 제품의 장점은? "네티자 제품은 인덱스 설정이 필요 없고 물리 모델에서 자유롭다. 특히 데이터 튜닝 포인트가 제로에 가깝기 때문에 데이터 아키텍처 등의 비즈니스 요구 사항을 유연하게 받아들이고 물리적 데이터 조회 속도도 빠르다. 또한 뛰어난 확장성과 고성능 분석기능 역시 만족스러운 제품이다."

향후 비즈니스 계획은? "2008년 첫 네티자DW어플라이언스 도입 이후 2012년 2월 네티자의 신제품으로 정보계 시스템을 업그레이드 했다. KCB는 새로 도입한 트윈핀에 통계 시스템을 적용해 잠재적 대출 수요 등의 예측 능력을 높이고 회원사 대상의 다양한 CRM 데이터 분석 서비스를 수행할 예정이다. 또한 GIS를 통해 거주 지역별 고객 성향 분석 등 특정 지역의 고객 성향 분석 서비스 등의 새로운 비즈니스 경쟁력을 확보할 계획이다. 고도화된 분석이 보다 정확해지고 빨라지면서 생긴 여유는 새로운 비즈니스 혁신과 서비스 고도화를 위해 투자를 위한 기반이 됐기 때문이다."

기고 | 고성능 분석을 위한 논리적 DW 재구축 전략

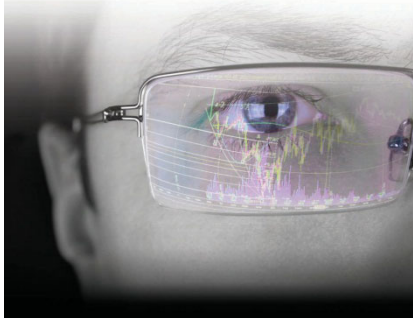
Mike Kearney | 제품 마케팅 책임자, IBM Netezza

비즈니스 인텔리전스는 DW를 토대로 구축된다. 이 DW 시스템은 주로 고객, 잠재 고객, 공급사, 비즈니스 파트너, 경쟁사, 규제당국 등과의 상호작용을 기록하는 거래/업무 처리 시스템과 같은 내부 데이터 소스에서 데이터를 가져와 통합한다. 일반적으로 DW는 중앙집중형 시스템으로 구현된다. 즉, 컴퓨터 시스템 한 대에서 한 가지 데이터 관리 기술을 운용하는 방식이다. 이러한 방식에서는 데이터베이스의 규모가 매우 커진다. 웨어하우스는 일반적으로 기업에서 가장 큰 데이터베이스다.

DW는 온라인 트랜잭션 처리 용도로 설계된 DBMS를 기반으로 구축된다. 온라인 트랜잭션 처리 시스템은 통상적으로 분석 처리에 사용되는 것보다 훨씬 적은 데이터 세트를 관리한다. 따라서 온라인 트랜잭션 처리에 최적화된 DBMS는 데이터 분석 용도로는 최선의 선택이 될 수 없다.

이제 BI는 단순한 보고서와 현황 정보가 아닌, 의사 결정을 위한 예측과 통찰을 제공하는 분석 애플리케이션으로 확대됐다. 보고서와 분석 애플리케이션은 서로 매우 다른 워크로드를 만들어낸다. 보고서 워크로드는 SQL로 표현되는 비교적 간단한 읽기 작업을 통해 이행되는 반면, 분석 애플리케이션의 운용에는 C, C++, 자바 등 언어를 이용한 대량의 컴퓨팅이 필요하다.

이메일과 로그 파일, 그리고 소셜 미디어 웹사이트에서 사용자가 생성한 콘텐츠와 같은 새로운 데이터 소스, 이미지, 오디오 등의 새로운 유형의 데이터는 1990년대 DW의 태동기에는 전혀 고려되지 않았던 요소들이다. 그러나 이 데이터들이 오늘날 BI와 분석론의 성장을 위한 기회를 제공하고 있다.



기업에서 비즈니스 인텔리전스 포트폴리오를 확대하고, 데이터에서 더 큰 가치를 끌어내기 위해 분석 애플리케이션을 구축하면서 DW의 중요성은 그 어느 때보다 커졌다. 애초에 오프라인 보고서와 초보적인 분석 쿼리를 지원할 목적으로 개발된 이 시스템은 두 가지 문제에 직면했다. 하나는 엄청난 양의 데이터이고 다른 하나는 지식 근로자 커뮤니티에서 만들어내는 다양한 종류의 분석 워크로드다. 중앙집중형 EDW 시스템이라는 기존 모델은 지나치게 경직돼 있어 새로운 분석 애플리케이션 수요를 따라잡을 수 없다. 따라서 IBM은 DW 아키텍처에 대해 재평가하기를 권고한다.

가트너는 지난 2011년 5월 2일 DW/BI 심포지엄을 통해 단일 EDW는 이미 과거의 개념일 뿐 아니라, 결코 실현된 적도 없었다는 의견을 피력했다. 가트너의 산업 분석 전문가들은 논리적 DW를 언급했다. 이 백서에서는 이 용어를 채택하고, 단일의 중앙집중형 물리 시스템에서 분산형 아키텍처로의 방향 전환이라는 발전을 권고한다. 개별 시스템에서 계산 능력을 제공하되, 각 노드는 특정 워크로드에 최적화된 형태를 말한다.

분산형 아키텍처에서 이루어지는 작업과 활동은 자동화된 소프트웨어로 통합된다. 시스템 관리자는 세부적인 모든 사건에 일일이 관여해야 하는 부담을 덜고, 관리 정책 수립이라는 좀더 총괄적인 업무에 집중할 수 있다.

이러한 방식에 따라 소프트웨어 도구로 데이터의 품질을 확보하고, 계보를 기록하면서 데이터를 이동하며, 전달 방향을 재지정하고, 접근 및 보안 통제 수단을 적용하면서 저렴한 비용으로 웨어하우스를 관리할 수 있는 것이다.

탄력적인 이 아키텍처는 애플리

IBM Netezza High Capacity Appliance(HCA)는 빅 데이터를 수용하기 위한 최적의 DW어플라이언스로 2011년 6월에 발표됐다.

케이션에 주는 영향을 최소화하면서 새로운 처리 노드를 추가하고 오래된 노드를 제거할 수 있다. 외부 환경은 기회와 위험을 함께 가져다 주며 이 두 가지의 부침과 변화에 따라 새로운 분석 애플리케이션을 신속히 배포할 수 있게 된다. 그 결과 민첩성이 향상돼 현업에도 혜택이 구현된다. 또한 웨어하우스 아키텍처 분산으로 간단한 두 가지 요건이 충족되므로 사업적 타당성을 확보할 수 있다.

이 변화를 통해 비즈니스 단위에 제공되는 데이터의 가치가 향상되는가? 그리고 이 접근 방식은 데이터 관리 및 분석 비용을 줄여주는가? 이러한 논리적 DW의 구현에 관한 IBM의 비전은 깊이와 일관성을 갖춘 제품 포트폴리오를 통해 뒷받침되므로, CIO는 이 두 질문에 자신 있게 "예. 그렇습니다"라고 답할 수 있다.

데이터 양적 증가에 따른 중앙집중형 기업 DW의 한계

현대의 웨어하우스와 분석 인프라는 매우 다양한 워크로드를 지원해야 한다.

- OLTP 용도의 데이터를 지식 근로자에게 유용한 형식으로 변환
- 일관성 있는 마스터 데이터 확보
- 보고서 실행
- 상황 정보 업데이트
- 운영상의 필요를 위한 데이터 분석
- 사기/부정 감지 등 사건에 대한 실시간 대응
- 임시적, 반복적 데이터 마이닝 활동 지원
- 소셜 미디어 및 웹 로그 데이터의 마이닝

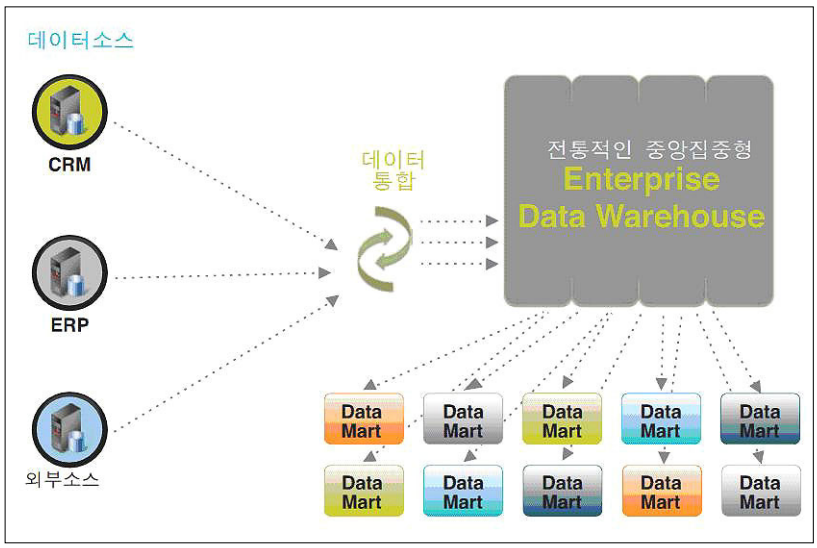
데이터의 양이 계속 늘어남에 따라 단일 컴퓨터 시스템으로 단일 DBMS를 운용하는 방식으로는 다양하고 광범위한 작업을 효과적으로 처리할 수 없게 됐다. 이러한 수요에 대응하는데 중앙집중형 기업 웨어하우스를 고수한다면 대부분의 워크로드에서 최적의 성능을 기대하기 어려울 것이다.

기존의 중앙집중형 EDW 모델

단일의 중앙집중형 EDW로는 오늘날과 같은 대량의 다양한 데이터를 제대로 처리할 수 없다. 따라서 각 비즈니스 단위에서 특별 임시 솔루션에 의지하게 됨으로써 데이터마트가 무질서하게 확산돼 효과적인 거버넌스가 불가능해지고 복잡성과 비용이

데이터-로드 작업 속도와 쿼리 성능이 떨어지면 우선 튜닝과 파티션 분할로 대응하게 된다. 이러한 방식은 결코 정확한 처방이 아니며, 관리자들을 웨어하우스 관리 업무에 끊임없이 묶어두는 결과만 초래한다. 따라서 기술 전문가들은 비즈니스 단위의 동료들과 제대로 협력할 수 없게 되며 전체 BI 비용이 늘어난다.

기존의 중앙집중형 EDW 모델



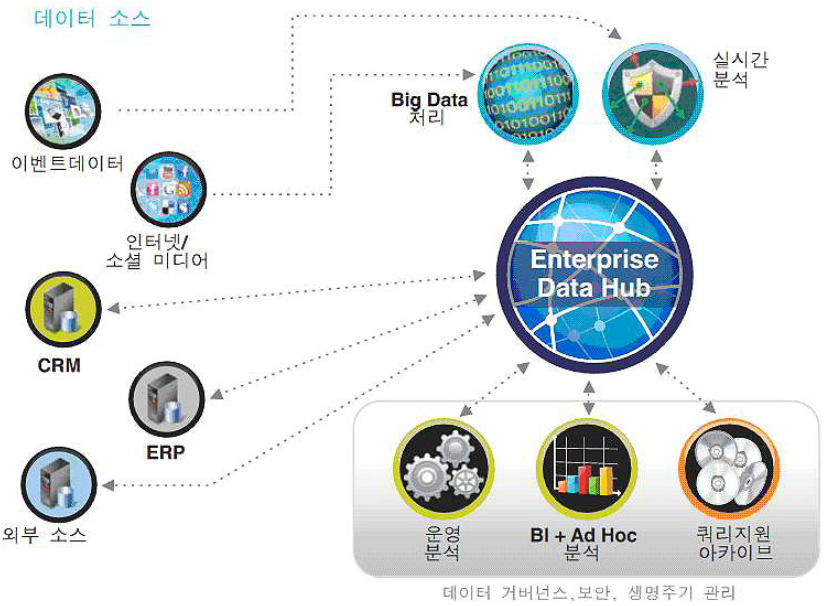
크게 늘어나게 된다.

모든 데이터를 단일 시스템으로 관리한다는 관점으로 보면, 데이터의 양과 유형이 지속적으로 늘려야 하는지 의문을 제기할 수 있다. 온라인 비즈니스 모델을 통해 웹 로그 형태로 생성되는 방대한 데이터 세트를 생각해보자. 이 파일은 풍부한 정보를 포함하고 있지만, PB 규모의 원시 데이터를 기존 웨어하우스로 로딩해 분석하는 방법은 최적의 방식이 아닐 것이다. 이처럼 비교적 적은 양의 구조화 데이터 용도로 설계된 추출-변환-로딩 처리 방식과 방대한 양의 웹 로그 데이터 사이의 부조화는 성공을 가로막는 첫 번째 장애물이다.

튜닝으로 중앙 웨어하우스의 성능을 높이는 데 실패할 경우, 잘 알려진 처방은 데이터 복제본을 만들어 제 2의 컴퓨터 시스템 또는 데이터마트에 옮기는 것이다. 그러나 문제 하나를 해결하려다 여러 개의 새로운 문제가 만들어지는 상황이 발생한다. 고립된 데이터 저장소(silo)가 늘어나면서 전사적인 분석에 제약을 받게 된다. 거버넌스의 실현은 어려워지고, 데이터 추출로 이미 과중한 부담을 지고 있는 EDW의 작업 부담은 더욱 늘어난다. 또한 비용과 복잡성은 수직 상승한다. 데이터 복제 방식에는 권장할 만한 요소가 많이 있지만, 효과적인 성공을 위해 구조적 접근의 전환과 새로운 소프트웨어 도구에 대한 투자가 필요하다.

스마트 통합: 새로운 접근

컴퓨팅을 중앙에서 관리하지만 그 실행은 각각의 특정 워크로드에 최적화된 개별 시스템에서 담당하는 아키텍처에서는 성능, 거버넌스, 확장성, 민첩성이 향상된다.



DW와 분석론 재고 : 기본 원칙 가이드

물리적 단일 시스템에서 탈피해 분산형 컴퓨팅 인프라에 소프트웨어 유틸리티를 결합하는 논리적 웨어하우스로 진화하는 방법을 제시하고자 한다. 이러한 진화는 세 가지 구조적 원칙을 지침으로 한다.

제 1원칙: 인프라의 통합 – 분석의 단순화

동일한 워크로드의 특성을 가진 작업/활동을 그룹화 하면 해당 워크로드에 특화된 전문 시스템과 분석 어플라이언스에 웨어하우스 처리 작업을 분산시킬 수 있다. IBM의 첫 번째 원칙은 무질서하게 확산된 다수의 데이터마트를 훨씬 더 적은 수의 분석 어플라이언스에 통합시켜 범용 서버를 기반으로 구축된 기존의 데이터마트와 달리 데이터마트 관리할 부담을 획기적으로 줄이면서도 분석 쿼리에 최고의 가격 대비 성능을 제공하는 것이다. 그러면 IT 리소스의 부담이 줄어 새로운 분석 애플리케이션에 더욱 집중할 수 있다.

제 2원칙: 목적에 맞는 가이드에 따라 데이터와 컴퓨팅 분산

기존 설계 목적이 아니었던 분석 처리에 대한 부담을 벗겨 줌으로서 중앙 DW는 더 많은 계산 리소스를 보다 많은 운영 활동에 집중 투입할 수 있을 뿐만 아니라, 기업에 중요한 데이터의 정확성과 계보를 확실히 관리할 수 있다. 결과적으로 전사 데이터 웨어하우스(EDW)는 이러한 전환을 통해 메타데이터 관리 작업, 적절한 거버넌스 및 계보의 확보,

간단한 어플라이언스 기반 접근 방식에 따라 데이터를 적절한 분석 엔진에 배급하는 역할 등 기업 데이터 허브로서 새로운 역할을 수행하는 단계로 발전한다. 또한 데이터 관리의 중앙집중화로 복잡성과 비용을 줄이고, 엄밀한 데이터 거버넌스를 간단히 구현할 수 있다.

이 외에도, 운영 분석용 어플라이언스와 쿼리 지원 아카이브로서의 어플라이언스, 전송 중인 데이터(예: 디지털 센서에서 송출되는 데이터, 이메일 네트워크상의 데이터)의 실시간 분석을 위한 스트림 처리 시스템, 웹 클릭스트림 및 통화 상세 내역 기록과 같은 빅 데이터 분석을 위해 하둡 분산 파일 시스템(Hadoop Distributed File System)을 운용하는 그리드 등, 데이터 관리 및 계산 작업을 더욱 분산할 수 있는 다양한 방법이 존재한다. 논리적 DW는 역동적인 시스템으로서 빅 데이터의 분석 결과를 더욱 심층적인 분석이 가능한 분석 어플라이언스로 옮길 수 있다.

제 3원칙: 논리적 웨어하우스 전역에서 관리 조율

정책 기반 관리와 어플라이언스 중심 접근 방식을 통해, 논리적 DW는 지식 근로자의 입장에서 매우 간단해진다. IBM의 제품 포트폴리오에는 공급사가 다른 경우까지 포함해 다양한 컴퓨터 시스템을 다룰 수 있는 이기종 유틸리티가 포함돼 있다. 이 유틸리티들은 공통의 관리 틀 안에서 메타데이터 관리, 신속한 데이터 이동 및 통합, 데이터 거버넌스, 라이프사이클 관리 등 다양한 기능을 제공한다. 또한 이 분산형 인프라 접근을 가상화하면 데이터 관리 및 처리 장소에 대한 자세한 사항을 사용자들이 파악하지 않아도 되며 인프라에 새로운 컴퓨팅 노드 또는 어플라이언스가 추가되더라도 쿼리 전달 방향을 재지정할 수 있는 장점을 가질 수 있다.

데이터 분석 발전을 촉진하는 논리적 DW

MIT 슬론 매니지먼트 리뷰(Sloan Management Review, 2011년 겨울호)에 게재된 '빅 데이터, 분석론 및 통찰을 가치화하는 방법(Big Data, Analytics and the Path from Insights to Value)'이라는 제목의 논문에 따르면, 비즈니스 정보 및 분석론의 활용이 업종 내 경쟁력 차별화에 기여한다는 주장에 적극적으로 동의하는 기업일수록 최고의 실적을 거둘 확률이 반대로 낮은 실적을 기록할 확률에 비해 두 배 높았다.

그러나 많은 기업들은 비즈니스에서 분석 기술을 활용하는데 있어 가치 실현 속도를 보다 높여야 할 필요가 있다는 사실도 인식하고 있다. 데이터웨어하우징 인스티튜트(The Data Warehousing Institute)가 2011년 6월 발행한, '셀프서비스 비즈니스 인텔리전스: 사용자의 통찰 지원(Self-Service Business Intelligence: Empowering Users to Generate Insights)'라는 보고서를 통해 콜린 화이트와 클라우디아 임호프는 응답자의 74%가 분석 기술 시스템의 가치 실현 속도를 높여야 할 필요가 있다는 최근의 설문 조사 결과를 내놓았다.

이미 많은 기업들이 과중한 부담을 지고 있는 중앙의 웨어하우스에서 DW어플라이언스로 분석 처리를 옮기는 방법을 찾아냈다. 논리적 DW 개념을 중심으로 설계된 IBM의 스마트 통합(Smart Consolidation) 전략은 이러한 성과를 바탕으로 데이터 관리의 강화, 복제 자동화, 거버넌스 통제력 강화를 실현한다. 데이터 관리를 그 구체적 활용으로부터 분리함으로써 가치 실현 속도는 빨라지고, 비용이 절감되며, 향후 새로운 분석 기술이 등장하면 처리 노드를 추가해 적용할 수 있는 탄탄하고 유연한 아키텍처를 구성할 수 있는 것이다.

IBM은 비즈니스 분석론에 140억 달러 이상을 투자했다. IBM 네티자 DW어플라이언스는 기업의 논리적 분산형 DW 구현 과정에서 중요한 역할을 담당할 핵심 요소로서 다양한 데이터 유형에 대한 폭넓은 분석에 최적화돼 있다. IBM은 최적화된 어플라이언스와 시스템을 포함해 스마트 통합 전략을 실현하는데 필요한 모든 포트폴리오를 완벽히 갖춘 유일한 기업이다. CIO

NETEZZA®
an IBM® Company

IBM Netezza

비즈니스 통찰력을 위한 DW 분석전용 어플라이언스

간단하고 빠르고 쉬운 비지니스 분석으로
DW 인프라에 대한 고정관념을 완전히 바꾸어 드리겠습니다.

Exceptional Performance

Unparalleled Simplicity

Astonishing Value