Feature

빅데이터, 그 놀라움을 맛보다

앞서가는 비즈니스 기업의 빅데이터 활용법

IDC에 따르면 전 세계적으로 유용한 데이터량이 2010~2020년 사이에 20배 이상 증가하고, 기업 관련 데이터의 77%는 2015년 현재 비정형화된 상태로 남아있다고 한다. 데이터가 급증하고 데이터 종류가 많아지면서 기존의 관계형 DB와 데이터 웨어하우스 기술만으로는 해결되지 않는 정보들을 확보하기 위해 하둡(Hadoop), NoSQL 등 다른 툴로 전환하는 기업도 늘어나고 있다.

관련 연구보고서들은 빅데이터를 통한 새로운 기회가 현실이 되고 있지만 기업들은 그에 앞서 2개의 커다란 문제를 해결해야 한다고 지적한다. '어떤 방식으로 빅데이터에서 가치를 이끌어낼 것인가', '빅데이터 전략은 어떻게 수립할 것인가'가 그것이다. 이 두 가지 문제를 해결하기 위해서는 현재 빅데이터를 통해 기업이 성과를 얻고 있는 사례와 가까운 미래에 등장할 빅데이터 사례를 살펴보면 될 것이다.

빅데이터를 통해 성과를 얻을 수 있는 사례를 HDS(Hitachi Data Systems)와 펜타호(Pentaho) 솔루션 중심으로 살펴보고, 데이터를 기업의 핵심 가치로 둘 수 있는 방법을 알아보자.

01 CASE 데이터 웨어하우스 최적화

데이터 웨어하우스 최적화는 가장 일반적인 방식의 빅데이터 이용 사례로 기업은 주로 비용과 운영 성능, 두 가지 이유 때문에 이 방식을 택한다.

기업 내에 저장 및 액세스해야 하는 데이터의 양이 급증함에 따라 기존의 데이터 웨어하우스는 거의 한계에 봉착했다. 사용자들은 쿼리와 데이터 액세스의 성능 저하 현상을 피부로 느낄 것이다. 뿐만 아니라 데이터 웨어하우스용 스토리지 용량을 추가로 구입해야 할 수도 있다. 추가 구입은 고비용도 문제지만 데이터가 계속 증가한다는 점에서 근본적인 해결책이 될 수 없다.

이를 해결하기 위해 기업들은 빅데이터, 그중에서도 특히 하둡(Hadoop)을 고려한다. HDFS(Hadoop Distributed File System)에 데이터를 저장하면 기존의 데이터 웨어하우스 스토리지에 비해 상당한 비용을 절감할 수 있다. 특히 하둡 스토리지의 경우 TB당 약 1,000달러 정도에 불과하다. 이에 비해 하드웨어, 서버 등이 완벽하게 탑재된 데이터 웨어하우스 스토리지는 TB당 5,000~10,000달러 이상의 비용을 지불해야 한다.

SLA(Service Level Agreement)와 컴플라이언스 요구사항을 만족시키면서 데이터 스토리지 비용을 줄이기 위해 데이터 웨어하우스에서 사용 빈도가 적은 데이터를 하둡으로 옮길 수 있다.

접근 방식

기타 소스뿐 아니라 CRM(고객 관계 관리)과 ERP(전사적 자원 관리) 시스템의 데이터도 활용 가능하다. 하둡 클러스터를 통해 사용 빈도가 적은 데이터를 기존의 데이터 웨어하우스에서 폐기한다. 이로써 스토

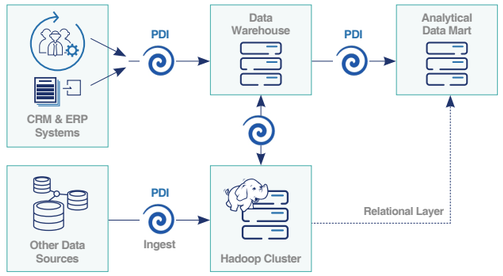
리지 비용을 절감할 수 있으며, 분석가들은 데이터 마트에서 신속한 쿼리로 정보에 액세스할 수 있다.

프로젝트 고려사항

데이터 웨어하우스 최적화는 현재 가장 보편적인 빅데이터 활용 사례 중 하나지만 이를 실행하려면 많은 시간과 노력, 계획이 필요하다. 하둡은 아직은 신기술이다. 따라서 하둡 배포에 수반되는 '독창적인' 툴을 사용하려면, 데이터 웨어하우스의 데이터를 하둡으로 옮겨 폐기하는 프로세스를 생성할 수 있는 자바 코딩 전문가가 필요하다. 하둡 개발자와 분석가는 아직 소수에 불과해 기업이 필요로 하는 인력을 채용하기가 쉽지 않다. 인건비도 SQL 및 기존의 다른 툴을 다루는 IT 인력에 비해 50% 이상 더 책정해야 한다.

펜타호는 수동 코딩을 없애 모든 데이터 개발자가 하둡에 액세스할 수 있도록 직관적인 GUI를 제공한다. 시간을 단축하고 인건비를 절감할 수 있다는 말이다. 데이터 통합 솔루션을 제공하는 기업이라고 해도 기존 데이터 소스와 데이터베이스를 하둡과 통합하는 노 코딩(no-coding) 솔루션을 보유한 곳은 없다.

(그림) 데이터 웨어하우스 최적화를 위한 펜타호의 제안



02 CASE 스트림 데이터의 수집 및 정제

정형화된 트랜잭션, 고객 데이터, 기타 데이터 등이 축적돼 데이터 양이 급증하게 되면, 기존의 ETL(Extraction, Transformation, Loading) 시스템의 속도는 급격히 저하되어 더 이상 분석 작업을 수행할 수 없는 상태가 된다. '데이터 정제' 솔루션은 하둡을 이용해 데이터를 변환하고, 대부분의 데이터 소스를 확장 가능한 빅데이터 프로세싱 허브를 통해 간소화한다. 정제된 데이터는 데이터 전반에 대한 로우 레이턴시(Low latency) 서비스 분석을 위해 분석 데이터베이스로 전송된다.

이 사례는 비용을 절감하고 최적화된 데이터 웨어하우스의 성능을 강화하기 위한 방법이다. 이 시점에 수많은 종류의 다양한 데이터가 하둡으로 로딩되며, 하둡은 수집된 비즈니스 인사이트 도출을 위한 소스로 전환된다.

이는 데이터 웨어하우스 최적화에 비해 변환 작업이 훨씬 더 수월하다. 하둡, 버티카(Vertica) 및 그린플럼(Greenplum) 등 분석 데이터베이스가 결합돼 더 빠른 쿼리, 신속한 수집, 강력한 프로세싱이 가능하므로 기업은 대량의 다양한 데이터 소스에서 유용한 분석 결과를 얻을 수 있다. 또한 데이터 분석 담당부서는 데이터 셋으로부터 예측 분석을 더 빠르게 수행할 수 있다.

접근 방식

이 사례는 개인화 서비스를 제공하는 온라인 마케팅 기업의 '정제' 아키텍처를 통해 확인할 수 있다. 온라인 캠페인, 등록, 트랜잭션 데이터가 하둡을 통해 수집 및 처리되어 분석 데이터베이스로 전송된다.

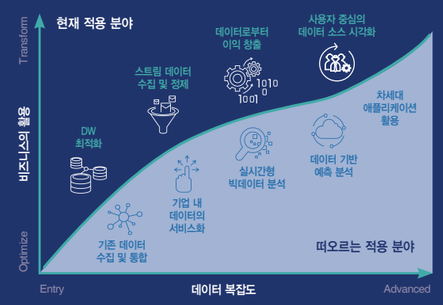
BIG DATA

빅데이터의 힘

기업 내 프로세스 최적화부터 비즈니스 모델 변화까지

아래 그림은 비즈니스의 활용 정도와 데이터 복잡도에 따라 카테고리화한 10개의 빅데이터 이용 사례다. 이들 사례가 기업에 미치는 영향은 현재의 프로세스 최적화부터 전체 비즈니스 모델 변화에 이르기까지 광범위하다. 현재 빅데이터를 통해 기업이 성과를 얻고 있는 사례와 가까운 미래에 등장할 빅데이터 사례들을 간단히 알아보자.

(그림) 빅데이터 이용 사례



현재의 적용 분야

데이터 웨어하우스 최적화

기존의 데이터 웨어하우스는 데이터의 양이 급증하면서 한계에 다다랐다. 데이터 웨어하우스 가용량을 확대하려면 상당한 비용이 소요되므로 기업들은 사용빈도가 낮은 데이터를 삭제함으로써 데이터 웨어하우스 성능이 향상될 수 있도록 빅데이터로 나아가고 있다.

스트림 데이터의 수집 및 정제

빅데이터 스토어는 이제 다양한 소스에서 취합된 데이터가 로우 레이턴시(low latency) 분석(대개는 신속한 쿼리를 위한 분석 데이터베이스)을 위해 다른 곳으로 이동하기 전에 체류 및 프로세싱되는 구역(zone)의 역할을 하고 있다. 이를 통해 ETL과 데이터 관리 비용이 대폭 절감되었으며 빅데이터가 분석 프로세스의 핵심 영역에 자리하게 되었다.

사용자 중심의 데이터 소스 시각화

고객의 모든 접점에 대해 적시적소의 분석 관점을 제공하기 위해 운영 및 트랜잭션이 진행 중인 다양한 데이터 소스를 통합한다. 뿐만 아니라 고객 접점에서 근무하는 직원들과 파트너사들도 기업의 전체 업무에 대한 일 단위 애플리케이션 정보를 활용할 수 있다.

데이터 판매

빅데이터를 기반으로 포괄적이며 익명으로 처리된 데이터 셋이 서드파티 고객 기업들에게 서비스로 제공된다. 기업은 강력한 데이터 프로세싱과 심층 분석을 통해 신규 매출원을 확보할 수 있다.

확산 가능성이 높은 분야

기존 데이터 수집 및 분석

많은 기업이 방대한 규모의 데이터를 빅데이터 저장소에 쏟아 넣고 있지만 어떤 정보가 저장돼 있으며, 어떻게 해야 이들 데이터를 생산성 있는 정보로 전환할 수 있는지에 대해서 분명한 해답을 갖고 있지 않다. 기본적인 데이터 마이닝 알고리즘을 작동시켜 데이터와 다른 소스 간에 찾아낸 패턴의 상호연관성을 연구하는 것부터 시작해야 할 것이다.

실시간형 빅데이터 분석

센서, 라우터, 셋톱박스 등에 저장된 고용량 데이터 분석은 얼마 전까지만 해도 엄청난 비용이 드는 거대 프로젝트였다. 그러나 빅데이터가 확산되고 있는 현재는 상황이 다르다. 데이터 마이닝과 짧은 대기시간(low latency) 서비스를 위해 머신 데이터 및 센서 데이터를 활용할 수 있다.

데이터 기반 예측 분석

빅데이터는 기계 학습(머신 러닝) 알고리즘[1] 최적화(교육과 평가)와 이를 활용해 성과(평점)를 예측하거나 성과에 영향을 미칠 수 있는 새로운 툴 셋을 제공한다. 빅데이터 저장소에서의 예측 분석 솔루션으로는 부정거래 탐지, 추천 엔진, 최적화 등 애플리케이션이 포함된다.

차세대 애플리케이션으로 활용

애플리케이션 벤더들은 더 강력하고 지능화된, 높은 가치를 제공하는 솔루션을 개발하기 위해 데이터/분석 아키텍처를 끊임없이 혁신 중이다. 사용자 애플리케이션에 내장된 분석 인터페이스를 통해 매출 상승 효과를 얻을 수 있다.

주문형 빅데이터 혼합

빅데이터 저장소가 생성되면 관련 부서는 기존의 데이터 웨어하우스 인프라와 관련된 업무를 위해 시간을 더 쓸 수밖에 없다. 시급을 다투는 요청이라면 데이터 웨어하우스를 완전히 우회해야 할 수도 있다. '적시 혼합(Just in time blending)'을 통해 모든 소스에서 취합된 정확한 데이터를 적시적소에 제공하므로 단계적으로 데이터를 분석할 필요가 없다.

기업 내 데이터 서비스화

데이터 수집과 액세스를 담당하는 수많은 애플리케이션 개발팀 전체에 대한 서비스를 제공하기 위해 공유 데이터베이스 서비스로 빅데이터에 접근한다. 사일로(저장장치를 물리적으로 계속 늘리는 방법) 기반 접근방식과 달리 규모의 경제와 비용절감 효과를 얻기 위해서다. 중앙 집중화된 엔터프라이즈 스택의 한 컴포넌트로 ETL과 분석 솔루션이 포함될 것이다.

1) **기계 학습 알고리즘** 인공지능의 연구 분야 중 하나로, 인간의 학습 능력과 같은 기능을 컴퓨터에서 실현하고자 하는 기술 및 기법이다.

비즈니스 분석에는 리포팅과 기업 사용자를 위한 즉각적인 분석 서비스가 포함된다.

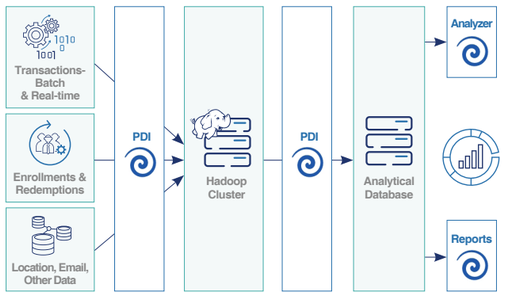
프로젝트 고려사항

이 프로젝트는 데이터 종류가 다양하고 소스가 많을수록 실행하기가 더욱 복잡해진다. 따라서 현재의 다양하고 방대한 시스템과 미래의 데이터 시스템을 유연성 있게 통합할 수 있는 데이터 통합 및 분석 플랫폼을 선택하는 것이 무엇보다 중요하다.

이 사례의 경우 데이터 개발자와 비즈니스 분석가 간 협업이 최우선적으로 요구된다. 통합 플랫폼은 데이터 연계와 비즈니스 인텔리전스를 위해 필요하며, IT 사용자와 일반 사용자가 독립적인 툴 셋을 최대로 활용할 수 있도록 조정하는 것은 훨씬 더 어려운 일이다.

분석 데이터베이스는 이 아키텍처의 핵심이다. 이러한 데이터베이스는 더 빠른 쿼리, 더 나은 확장성, 다차원 분석 '큐브' 및 인메모리 기능 등을 제공하며, 비즈니스 인텔리전스에 최적화되어 있다. 이와 비교하면 기존의 트랜잭션 데이터베이스는 원하는 수준의 쿼리와 분석 기능을 제공하지 못할 수도 있다.

(그림) 스트림 데이터의 수집 및 정제를 위한 펜타호의 제안



03 CASE
사용자 중심의 데이터 소스 시각화

데이터 웨어하우스 최적화와 스트림 데이터의 수집 및 정제가 비용과 효율성 측면에 중점을 두고 있다면, 사용자 중심의 데이터 소스 시각화는 특히 이동통신, 병원, 금융 서비스 등 고객 이탈이 잦고, 경쟁이 심한 시장에서 활용 가치가 높다. 이 분야에서 비즈니스를 성공시키는 2개의 핵심 동력은 급증하고 있는 '끼워팔기'와 '고객 이탈로 인한 리스크 최소화'다.

이 사례는 NoSQL이나 하둡처럼 빠른 쿼리가 가능하도록 거의 모든 고객 접점의 데이터를 싱글 리포지터리(repository)로 가져와 백엔드에서 활성화시킨다. '사용자 중심의 데이터 소스 시각화'를 통해 각각 독립적으로 존재하던 데이터가 혼합돼 기업 내 담당 부서들은 자사의 브랜드와 서비스에 대한 고객의 인식을 더 잘 파악할 수 있으며, 구매자의 성향을 더 잘 이해할 수 있다. 고객과의 접점에 맞닿아 있는 직원들이 이러한 통찰력을 확보하면, 더 생산적이고 높은 수익을 보장하는 의사결정을 보다 신속하게 내릴 수 있다.

접근 방식

위 사례에서 금융 서비스 기업은 다양한 소스에 존재하는 데이터를 NoSQL을 통해 단일 빅데이터 저장소에 보관한다. 데이터는 이 시점부터 고객에 대한 완벽한 파악을 위해 고객의 고유 ID로 처리된다. 이후 정확성이 더해진 정제된 고객 데이터는 콜센터 직원, 리서치 분석가, 데이터 분석가 등 각 분야 담당자들에게 전달되고 적절한 분석 결과를 제공한다.

프로젝트 고려사항

기업 입장에서는 충분히 채택할 수 있는 사례지만, 동시에 대단히 복잡하고 많은 리소스가 필요한 작업일 수 있다. '사용자 중심의 데이터 소스 시각화'는 비즈니스 관점에서 중대한 전략적 기획이 전제돼야 한다.

첫째는 특정 매출 목표를 이 프로젝트와 연계해야 한다는 점이다. 따라서 주주들이 좋은 성과를 얻으려면, 고객만족 요인 외에 고객 접점에 있는 직원들이 기대 가능한 성과 또한 정확하게 파악하고 있어야 한다. 더불어 최종 사용자(End user)도 계획 단계부터 참여해야 한다. 그래야 필요한 정보가 적시에 가장 필요로 하는 사람에게 정확한 형태로 전달될 수 있다. 또한 분석가들은 도입할 솔루션에 대해 사용자에게 충분히 설명해야 한다. 간편한 액세스와 직관적인 이해가 가능한 분석 결과를 기업의 중요한 애플리케이션에 반영하기 위해서다.

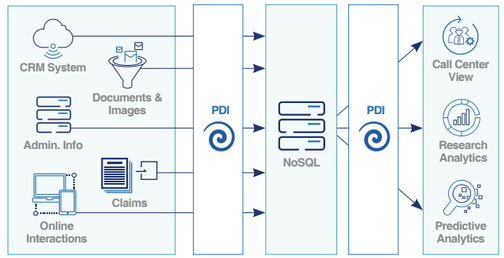
많은 고객 정보를 간편하고 신속하게 서버에 배포할 수 있는 단일 컬렉션으로의 전송 방안을 찾고 있다면 몽고DB(MongoDB) 등 NoSQL 솔루션을 빅데이터 저장소로 선택할 수 있다. 하지만 데이터의 배치 프로세스가 가능한 상황이고, 시간 순으로 저장해야 하는 경우라면 하둡이 더 나을 것이다.

고객 분석이 필요한 사용자들은 다양한 종류의 BI를 요구할 것이다.

- 경영진을 위한 직관적이고 커스터마이징 가능한 대시보드
- 분석가를 위한 고도화되고 응답 가능한 즉석 슬라이싱/다이싱(slicing/dicing) 툴
- 팀 전체의 정보 공유를 위한 분산 리포팅 기능
- 데이터 분석가를 위한 데이터 마이닝과 예측 분석 툴
- CRM, 서비스 애플리케이션과 같은 운영 소프트웨어의 분석

통합 플랫폼에 이러한 모든 기능이 갖춰져 있지 않을 경우, 대부분의 기능을 제공하는 데이터 및 분석 업체를 구해야 한다. 동시에 벤더들은 기업의 데이터 통합 시 신기술을 무리 없이 수용할 수 있어야 한다. 프로그램의 재설치를 최소화할 수 있을 뿐만 아니라 시스템의 유연성을 높일 수 있기 때문이다. 이는 사용자 중심의 데이터 소스 시각화와 같이 끊임없이 진화하는 사용자 요구에 맞게 데이터 아키텍처를 변화시켜야 하는 고도화된 프로젝트에는 특히 중요한 문제다.

(그림) 사용자 중심의 데이터 소스 시각화를 위한 펜타호의 제안



04 CASE
데이터 판매

일상적인 기업 활동에서 끊임없이 생성되는 다양한 종류의 데이터가 서드파티에게는 가치 있는 데이터가 될 수 있다. 이 경우 데이터 구매자는 주로 외부 마케터들이 될 것이다. 예를 들어 통신업체는 휴대폰 업체로부터 서로 다른 시간대의 위치 데이터를 수집해 인구통계 데이터와 결합한 후 최종적인 결과물을 유통업체에 판매하고, 유통업체는 이 데이터를 활용해 매장 계획을 수립할 수 있다. 통신업체에게는 새로운 수익원이 생기고, 오프라인 유통업체는 분석 데이터에 기반해 잠재 고객을 대상으로 효과적으로 타겟팅 할 수 있는 방안이 생긴 것이다.

접근 방식

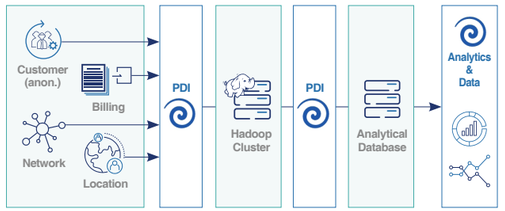
위의 사례에서 통신업체는 유동 구매 잠재력을 파악하기 위해 특정 지역과 관련이 있는 인구통계학 및 모빌리티 데이터를 통합해 서드파티 업체에 특화된 분석 서비스를 제공한다. 이 사례는 하둡과 분석 데이터베이스 모두를 최대한 활용하고 있다.

프로젝트 고려사항

가트너는 2016년까지 기업의 30%가 데이터 자산을 판매할 것으로 예측하고 있다. '데이터 판매' 사례에서 하둡은 데이터 프로세싱 플랫폼으로 활용된다. 고가의 레거시 데이터 웨어하우스 솔루션에 비해 훨씬 더 낮은 비용으로 높은 수익을 창출할 수 있다. '데이터 웨어하우스 최적화' 부문에서 언급한 것처럼 TB당 비용은 하둡이 5~10배 정도 저렴하다.

펜타호의 노코딩(no-coding) 빅데이터 통합과 비즈니스 분석 기능을 가미하면 수익성과 시간 절감 효과는 더 커진다. 이와 동시에 서드파티에 대해 분석 서비스를 제공하는 경우, 기존 웹 애플리케이션에 리포팅과 시각화 기능을 포함해야 할 수도 있다. 펜타호는 오픈 아키텍처 기반 솔루션이라는 점과 시각화에 뛰어나다는 점에서 최적의 대안이다.

(그림) 데이터 판매를 위한 펜타호의 제안



* 출처: Blueprints for Big Data Success; http://www.pentaho.com 2015년