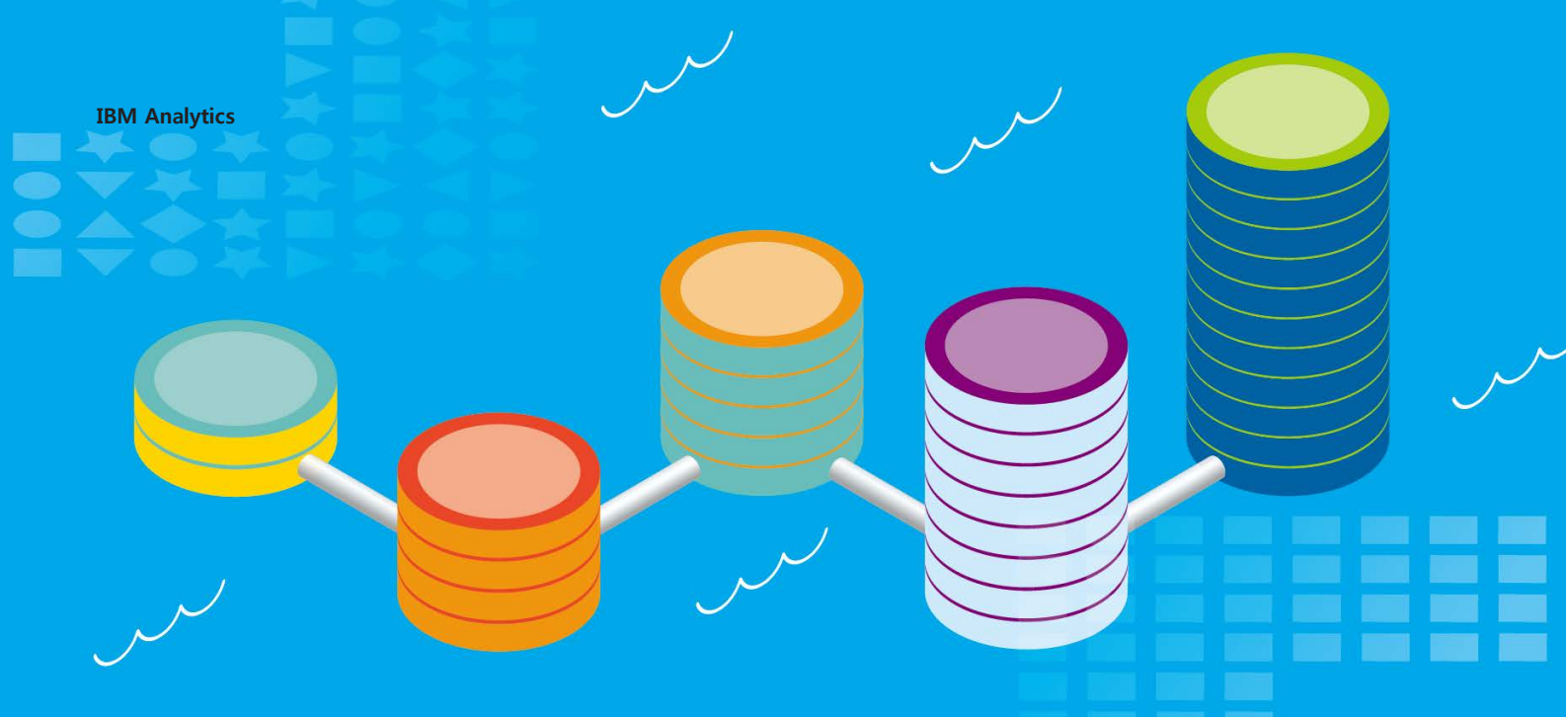
IBM Analytics



데이터 길들이기

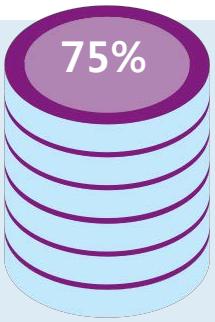
데이터를 어디서나 수집하고, 어디서든 관리하며, 모두를 위해 가치를 창출하는 방법

데이터의 변화

기업의 디지털 혁신과 함께 데이터 볼륨이 무서운 속도로 증가하고 있습니다. IDC에 따르면 지구의 인구 1인당 1초에 1.7MB의 데이터가 생성됩니다. 관계를 확장하고 디지털 및 모바일 플랫폼에 서비스를 프로비저닝하며 고객 관계에서 사용자가 생성한 컨텐츠를 활성화하는 과정에서 데이터가 급증하고 있습니다. 이제 기업은 새로운 환경에서 데이터를 처리, 관리, 분석하고 결정하려면 먼저 데이터에 대한 주도권을 확보해야 합니다.

데이터 볼륨의 증가는 새로운 소식이 아닙니다. 수십 년간 데이터 관리에서 다뤄온 주제입니다. 빅데이터의 시대에서 더 어려운 점은 데이터 소스가 갈수록 다양해진다는 것입니다. 제품, 고객, 매출, 거래에 관한 통상적인 정형 데이터가 주를 이루던 내부 생성 데이터는 그 범위가 확대되어 고객이 보낸 문자 및 메시지에 대해 고객 서비스 팀이 작성한 메모부터 사내외 브랜드 플랫폼에서 모든 디지털 접점의 지속적인 클릭스트림을 통해 제작되는 이미지 및 동영상까지 포함합니다.

뿐만 아니라 데이터 사이언티스트 팀의 등장과 같은 새로운 비즈니스 인텔리전스 및 예측 분석 프랙티스 때문에 검색하고 분석해야 할 데이터 세트의 범위가 더욱 확대되고 있습니다. IBM은 외부로부터 유입되는 데이터의 가치에 주목했으며 최근 The Weather Company(미국에서 4번째로 많이 쓰는 모바일 앱을 운영하는 회사)를 인수하고 IBM Cloud에서 데이터 서비스를 시작했습니다. IBM은 지리정보 공급자인 Mapbox 및 소비자 데이터 브로커인 Acxiom과 손잡고 이들의 큐레이션된 서드파티 데이터를 IBM 솔루션에 있는 퍼스트파티 고객 데이터와 즉시 통합할 수 있도록 했습니다.



2019년에는 분석 솔루션의 75%가 세컨파티 파트너 또는 서드파티 공급자로부터 입수한 외부 데이터 소스를 10가지 이상 포함할 것입니다(출처: Gartner).

기업들은 효과적인 의사결정 및 실시간 고객 지원을 위해 그 어느 때보다 많은 데이터를 생성하고 어느 때보다 빠른 속도로 사용하고 있습니다. 따라서 데이터 관리 프로세스 및 기술의 측면에서 이러한 새로운 수요를 해결하려면 커다란 변화가 불가피합니다.

IBM

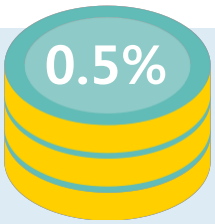
전혀 새로운 데이터 인프라

데이터 웨어하우스는 확고한 기반을 다진 안정적이고 성숙한 환경이며, 여기서 기업들은 거래 및 재무 정보, 고객 기록, 대화, 인구 통계 정보 등 비즈니스에 중요한 데이터를 저장할 수 있습니다. 전사적 범위나 각 운영 부문 범위에서 구축된 데이터 웨어하우스는 코어 비즈니스 프로세스에 데이터를 공급하고 보고 및 분석을 지원하며 부서별로 생성되는 각기 다른 보기를 통해 해당 기업이 현황을 제대로 이해할 수 있게 해줍니다.

그러나 이 데이터 인프라는 실시간 대용량 비정형 데이터 플로우에 최적화되지 않았습니다. 새로운 데이터 소스(특히 서드파티)는 리드 타임이 길어질 수 있습니다. 게다가 대개 방대한 데이터 볼륨을 탐사하기 위해 검색하고 분석하는 데이터 사이언스는 반복 가능하고 연속적이며 명확하게 정의된 비즈니스 프로세스 지원에 최적화된 엔터프라이즈 데이터 웨어하우스에는 큰 부담으로 작용할 수 있습니다.

그에 따라 데이터 인프라에 대한 새로운 접근 방식이 대두했는데, 분산형 스토리지 및 클라우드 기반 또는 범용 하드웨어 컴퓨팅 솔루션(대개 오픈소스 애플리케이션을 실행하고 있음)을 활용하는 것입니다.
특히 데이터 사이언티스트의 입장에서 이러한 솔루션의 매력 중 하나는 시간 제약이 있거나 목적이 한정된 프로젝트를 위해 구축했다가 목적을 달성하면 종료할 수 있다는 것입니다.

한편 이러한 솔루션은 "섀도우 IT(shadow IT)"로 운영될 위험도 있습니다. 즉 제어, 거버넌스, 감독을 포함하여 기술 프로젝트의 모든 영역이 엔터프라이즈 프레임워크에 속하지 않는 것입니다. 정형 데이터와 비정형 데이터를 통합 분석하여 가치를 창출할 인사이트를 얻거나 이 실험적 환경에서 개발된 모델을 운용하는 것도 제한될 수 있습니다.



기업은 생성하는 데이터의 80%를 저장하지만, 그 중 분석 대상이 되는 데이터는 0.5%에 불과합니다(출처: MIT).

이러한 문제를 해결하려면 두 환경의 장점을 융합하는, 즉 기존 레거시 시스템을 유지함과 동시에 액세스 가능한 데이터 기반을 확대하는 새로운 데이터 아키텍처가 필요합니다. 데이터 레이크(data lake)라고 부르는 이 아키텍처는 어떤 소스의 데이터 플로우도 받아들여 공통 플랫폼에서 사용 가능한 상태로 만듭니다. 데이터는 정제되지 않은 원시 상태로 저장하고 필요에 따라 찾아서 가공하고 정제하고 추출합니다.

IBM은 여기서 한 발 더 나아가 데이터 레이크에 저장되는 데이터를 카탈로그화하고 분류하여 저장 상태의 데이터 및 사용 중인 데이터 모두에 대한 데이터 거버넌스를 보장합니다. 그러면 IT 팀과 데이터 사이언티스트는 다음과 같은 중요한 비즈니스 이점을 누리게 됩니다. 프로젝트의 니즈에 민첩하게 대처하면서 비용을 낮추고 비즈니스 크리티컬 데이터 프로비저닝에서 탄력성를 유지하며 거버넌스 범위에 포함되지 않은 데이터 환경이나 용도가 갑자기 나타나는 것을 방지할 수 있습니다. 그리하여 범위를 알 수 없는 예측 불가능한 데이터 늪(data swamp)이 아니라 개방적이고 모니터링 및 유지 보수가 가능한 신뢰할 수 있는 데이터 세트를 얻게 됩니다.

비즈니스 활용 사례

각 기업은 시장 입지, 기술 도입, 성숙도, 자원 등 다양한 이유를 반영하여 데이터 레이크를 구축합니다. 공통적으로 나타나는 여러 동인 중 몇 가지를 정리해보면 다음과 같습니다.

모바일 앱 – 실시간 의사결정 및 검증이 아니라 일괄처리를 위해 설계된 기존 데이터 인프라는 모바일 디바이스를 통한 고객 서비스와 거래를 지원하는 데 어려움이 있습니다. 데이터 레이크에서는 컨텍스트 정보(예: 디바이스 ID, 위치)를 정형 데이터(예: 계좌 번호, 비밀번호)와 조합하여 안전하고 탄력적이며 신뢰할 수 있고 복제 가능한 모바일 앱 기반 서비스를 구현할 수 있습니다.

일반적으로 데이터 사이언티스트는 데이터를 준비하는 데 최대 80%의 시간과 노력을 씁니다(출처: Forrester Research)

예측 분석 – 차선책 또는 제품 추천과 같이 성향(propensity) 모델을 기반으로 하는 의사결정에서는 고객별 데이터(예: 구매 이력, 기호)를 포괄적인 컨텍스트(예: 구매 상황, 제품 구매 가능 여부)나 심지어 외부 참조 데이터(예: 날씨, 휴일 달력)와 조합하여 정확도 높고 유의미한 개별 맞춤형 커뮤니케이션을 제공해야 합니다.

사기 탐지 – 사이버 범죄자가 유효한 고객 인증 정보를 사용하여 계정에 접근한 다음 계정을 탈취하거나 비즈니스 시스템에 침투하거나 중요한 기밀 정보를 빼내는 경우가 늘고 있습니다. 최신 사기 탐지 기술은 보고된 도용 또는 분실 내역과 비교하면서 인증 정보를 점검할 뿐 아니라 소셜 및 디지털 데이터도 조사하여 모순이 되거나 이례적인 시도(예: 알려지지 않은 IP 주소를 통한 액세스 또는 해외로부터의 액세스)가 있는지 확인합니다. 가급적 거버넌스 환경에서 최대한 광범위한 데이터 세트에 대해 실시간으로 이 작업을 수행하면 사기 적발 가능성이 높아지고 오탐지가 줄어듭니다.

데이터 확장 – 기업이 상호 작용 및 거래의 전 범위에서 생성하는 고객에 대한 관점과 함께 포트폴리오 전반에서 고객에 대한 관점을 제시하는 외부 데이터 소스를 도입하는 것이 유익합니다. 데이터 레이크에 이러한 소스를 프로비저닝하면 데이터 및 소스 자체를 탄력적으로 변경할 수 있습니다. IBM은 The Weather Channel을 인수했고 Mapbox 및 Acxiom과 같은 여러 서드파티 데이터 오너들과 제휴하면서 검증 및 큐레이션을 거친 정보를 데이터 레이크와 같은 분석 및 의사결정 환경에 사전 통합하는 데 앞장서고 있습니다.

데이터 사이언스 – 내부 데이터와 큐레이션을 거친 외부 소스를 조합한 최대 규모의 데이터 세트를 조사하고 의미 있는 패턴을 밝혀냄으로써 고객 행동, 마케팅 기회, 제품 기능, 서비스 과제에 대한 새로운 인사이트를 얻는 경우가 늘고 있습니다. 이러한 데이터가 조합, 카탈로그화, 거버넌스가 완료된 상태로 제공되면 데이터 탐사를 위해 검색하고 조합하는 데 드는 시간(데이터 사이언티스트의 시간 중 최대 80%)이 단축되고 더 빠르게 비즈니스 가치를 창출할 수 있습니다.

데이터 수익화 – 데이터 가용성 확대에서 특히 마케팅과 관련하여 중요한 의미를 갖는 새로운 차원이 바로 더 우수한 연관성을 찾아 고객 관계를 발전시키고 확장된 데이터, 파생된 변수 또는 대상 미디어 기회를 재판매하면서 부가 가치를 창출할 수 있다는 점입니다.

장애물 극복

데이터 레이크는 새로운 프로세스 및 가치 창출 활동을 가능하게 하면서 큰 변화를 가져올 수 있습니다. 그와 동시에 데이터 레이크 관련 기술이 아직 성숙 단계에 이르지 못했기 때문에 기존 프랙티스에 큰 부담이 될 수도 있습니다. 아직도 자신들이 특정 데이터 유형의 오너라고 생각하는 일부 부서는 더 큰 통합 자원을 구축하기 위한 데이터 풀링에 소극적일 수도 있습니다. 이러한 데이터 사일로를 해체하는 데 당위성을 부여하려면 최고 경영진의 지지를 확보해야 합니다.

미국 기업이 데이터 품질 문제로 지출하는 비용이 연간 6천억 달러에 달합니다(출처: Data Warehousing Institute).

데이터 레이크가 신뢰받는 데이터 소스가 되려면 거버넌스가 이루어져야 합니다. 그렇지 않으면 데이터의 타당성, 가치, 유효 기간(데이터가 퇴화하거나 변화하는 속도, 즉 수명)을 고려하지 않고 데이터를 저장하는 짜임새 없는 저장 공간이 될 뿐입니다. LOB(line of business), 법률 및 컴플라이언스, IT, 분석 분야의 대표자들로 구성된 전담팀을 조직하여 지속 가능한 정책 및 데이터 정의를 마련함으로써 앞으로의 문제를 예방할 수 있습니다.

이러한 정의에 합의하고 메타데이터 관리 계층을 통해 배포함으로써 장기적으로 데이터 레이크에 추가되는 데이터의 일관성 및 사용 편의성을 유지할 수 있습니다. 이 프로세스는 전사적 범위에서 기존과 다른 중요한 차이점이 있습니다. 이를테면 고객 또는 매출의 정의, 재고 기록에 사용하는 단위, 주소 또는 전화 번호 서식 등입니다. IBM은 데이터 레이크에 속한 모든 소스에 일관성 있는 규칙 및 제어를 적용하는 정보 거버넌스 솔루션을 통해 이러한 프로세스를 지원합니다. 이 방식의 중요한 특징은 코어 엔터프라이즈 데이터 웨어하우스 및 확장된 데이터 레이크 모두에서 동일한 표준 및 제어를 적용하고 온프레미스 및 클라우드 기반 시스템 모두에 거버넌스를 수행할 수 있다는 것입니다.

새롭게 확장된 데이터 세트의 사용 현황을 파악할 수 있도록 액세스 제어 및 모니터링도 적용해야 합니다. 일부 실무자는 허용된 액세스 권한을 넘어서 액세스하려고 할 수 있습니다. 즉, 마스킹되거나 가명 처리된 데이터를 사용해야 하는데 데이터 사이언티스트가 개인 정보 데이터 세트 전체를 사용하려고 할 수 있습니다. 이러한 문제에 대한 감시와 감독을 철저히 하면 컴플라이언스 요건을 이행하는 데에도 도움이 됩니다.

비즈니스 이점

데이터를 길들이면 전사적으로 생산성 향상, 매출 신장, 마케팅 실효성 증가 등 중대한 효과를 거둘 수 있습니다. 데이터 레이크 프로젝트의 잠재 규모 및 복잡성을 고려하면 이 프로젝트의 ROI를 종합적으로 파악할 수 있도록 최대한 많은 영역에 메트릭을 적용해야 합니다. 다음과 같은 비즈니스 이점을 기대할 수 있습니다.

데이터 웨어하우스 비용 절감 – 데이터 레이크에서는 현재 데이터 웨어하우스에 저장되어 있지만 코어가 아닌 데이터 요소의 상당 부분을 더 저렴한 분산형 스토리지로 이동하여 비용 부담을 덜 수 있습니다.

Fortune 1000대 기업에서 데이터 접근성이 10% 향상되면 순수익이 6,500만 달러가 늘어나는 효과가 있습니다(출처: Baseline)

시스템 통합 비용 절감 – 부실한 IT 프로젝트에서 외부 컨설턴트의 실제 비용을 감추는 경우가 있습니다. 대개 중앙 IT 예산이 아닌 부서 예산에서 집행되기 때문입니다.

외부 분석 컨설팅 수요 감소 – 데이터 사이언스 팀에 데이터 및 기술 자원을 프로비저닝함으로써 외부 전문가를 고용하지 않고도 인사이트를 개발하고 예측 분석을 수행할 수 있습니다.

분석 팀의 업무 생산성 향상 – 일반적으로 데이터 레이크를 사용하면 더 신속하게 모델의 결과 및 인사이트를 얻고 이를 바탕으로 마케팅 및 세일즈 성과를 높일 수 있습니다.

데이터 품질 비용 절약 – 중복되는 기록, 배송 불가 주소 등 부실한 데이터 품질로 낭비되는 숨은 비용을 절약할 수 있습니다.

참조

IBM은 데이터 인프라 관리, 첨단 분석, 데이터 변환, 품질 관리, 거버넌스, 데이터 보호 분야에서 축적한 최고의 전문성과 경험을 바탕으로 귀사와 함께 기술 검증, 개념 검증, 데이터 랩, 현장 프로젝트 수행 등 다양한 프로젝트 활동을 수행할 수 있습니다.

성공적인 데이터 레이크와 관련된 비즈니스 이점 및 기회에 대해서는 다음 사이트에서 자세히 알아보십시오.
ibm.biz/data_lake

IBM United Kingdom
PO Box 41
North Harbour
Portsmouth
Hampshire
PO6 3AU

Produced in the United Kingdom
2016년 9월