관리형 데이터 레이크 방식

빠른 분석과 실행 가능한 인사이트 확보를 위한 데이터에 대한 셀프 서비스 액세스

주요 특징

- 조직 내의 모든 사용자가 모든 범위의 데이터에 셀프 서비스 방식으로 빠르게 액세스
- 분석의 속도 및 인사이트의 정확도 향상
- 비용 효율성과 확장성을 갖춘 환경에서 방대한 양의 데이터를 원래 형식으로 저장
- 데이터 품질, 보안, 거버넌스 유지

오늘날 기업들은 방대한 양의 데이터를 수집하고 확장된 자원으로부터 데이터 중심의 새로운 인사이트를 발굴하기 위해 분석 기능을 강화하고 있습니다. 증가하는 데이터를 최대한 활용하기 위해서는 조직 전반에서 신속한 데이터 액세스가 이루어져야 합니다. 그와 더불어 장기적으로 효율적이고 실행 가능한 방식으로 데이터를 저장하고 관리해야 합니다.

관리형 데이터 레이크(governed data lake) 방식으로 이러한 과제를 해결할 수 있습니다. 데이터 레이크는 여러 저장소(repository)로 구성된 공유 데이터 환경이며 빅데이터 기술을 활용합니다. 정형 데이터 및 비정형 데이터를 모두 지원하면서 민첩하고 안전하며 잘 관리되는 데이터 환경에 대한 요구가 증가함에 따라 기업들은 본격적으로 데이터 레이크 방식에 주목하기 시작했습니다.

데이터 레이크는 데이터 웨어하우스와 달리 데이터가 사용될 때까지 원래 형식으로 유지하는 플랫 아키텍처(flat architecture)를 사용합니다. 신속하게 데이터를 가져오고 저장할 수 있으며 언제라도 제약이 없는 셀프 서비스 방식으로 데이터에 액세스하여 분석하는 것이 가능합니다. 통합 거버넌스 기능으로 손쉽게 데이터를 찾아 이해하고 중복 없이 저장할 수 있습니다(그림 1).

관리형 데이터 레이크 방식

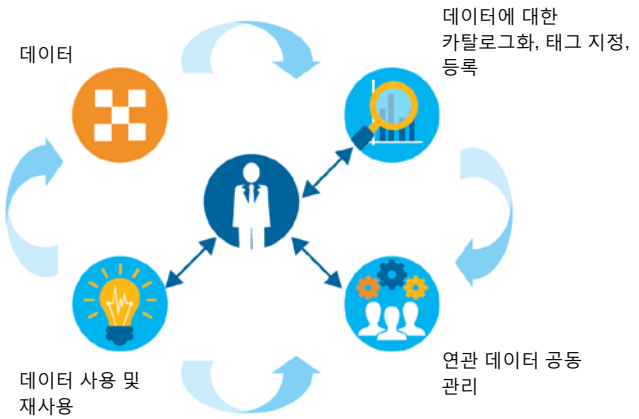

그림 1. 관리형 데이터 레이크의 각 구성 요소가 효율적으로 데이터를 처리하고 검색하므로 사용자는 더 쉽게 분석, 보고, 의사결정의 토대로 활용할 수 있습니다.

데이터 레이크로 주요 과제 해결

데이터 레이크는 기업이 전략적 비즈니스 및 IT 과제를 해결하는 데 중요한 역할을 할 수 있습니다.

- 비즈니스 사용자, 분석가, 데이터 사이언티스트, 개발자가 정확하고 유의미한 인사이트를 확보할 수 있도록 시의적절하게 모든 범위의 데이터에 대한 액세스를 지원합니다.
- 비즈니스 속도에 맞출 수 있도록 분석 및 데이터 준비를 빠르게 지원합니다.
- 사용자, 데이터 사이언티스트(심층 분석 기능이 필요할 경우), 데이터 엔지니어(데이터 레이크 기반 애플리케이션을 LOB(line-of-buisness) 사용자에게 배포할 시점) 간의 협업을 원활하게 합니다.
- 사용자에게 이해할 수 있고 믿을 만한 데이터를 제공함과 동시에 개인 정보를 보호하고 각종 규제에 대한 컴플라이언스를 유지할 수 있도록 데이터 품질, 보안, 거버넌스를 보장합니다.
- 빠르게 증가하는 데이터 볼륨을 비용 효과적으로 저장하면서 기존 자원을 최대한 활용하는 확장형 스토리지 모델을 통해 급속도로 증가하는 데이터 컬렉션을 수용합니다.

데이터 늪

가장 효과적인 데이터 레이크는 적극적인 데이터의 수집, 거버넌스, 보호, 관리를 수행하는 강력한 구성 요소로 이루어집니다. 데이터 레이크를 최적화하고 “데이터 늪(data swamp)”에 빠지지 않으려면 계획이 중요합니다. 데이터 늪은 말 그대로 복잡한 상태에서 필요한 체계, 가시성, 거버넌스를 갖추지 못하여 모든 가용 데이터를 십분 활용하지 못하는 환경입니다.

데이터 늪에서는 사용자가 데이터의 출처나 정확성을 확신할 수 없습니다. 필요한 데이터를 찾는 것은 고사하고 저장소에 그 데이터가 있는지조차 알 수 없습니다. 게다가 사용자는 데이터가 제대로 보호받고 있는지도 확신할 수 없습니다. 또한 데이터에 대한 비즈니스 컨텍스트 정보도 없습니다.

데이터 레이크 구분

데이터 레이크에 해당되는 것

- **사용자가 방대한 원시 데이터에 액세스할 수 있는 환경**
- **분석 모델을 개발하고 검증한 다음 프로덕션으로 이동하기 위한 환경**
- **인사이트를 얻기 위해 데이터를 탐색하는 분석 샌드박스**
- **사용자가 데이터를 찾고 비즈니스 용어와 기술 메타데이터를 연결하는 데 유용한 전사적 범위의 카탈로그**
- **데이터 변환 및 쿼리의 재사용을 지원하는 환경**

데이터 레이크가 아닌 것

- **기업의 모든 데이터를 수용하기 위한 데이터 웨어하우스 또는 데이터 마트**
- **대체 운영 데이터 저장소(operational data store, ODS)**
- **고성능 프로덕션 환경**
- **프로덕션 리포팅 애플리케이션**
- **특정 문제 해결을 전문으로 하는 시스템(단, 전용 데이터 마트는 데이터 레이크로부터 데이터를 가져올 수 있음)**

최적화된 데이터 레이크 구축 방법

다음에 나열된 단계는 데이터 늪의 함정에 빠지지 않고 데이터 레이크를 최적화하는 방법입니다.

데이터 저장소 구축

저장소는 기업에 정형 데이터와 비정형 데이터를 제공합니다. 각 저장소는 특별한 워크로드 기능을 지원하거나 데이터 컬렉션에 대한 특별한 관점을 제시합니다. 복수의 저장소가 다양한 데이터 유형을 수용할 수 있습니다. 기업은 필요에 따라 손쉽게 새로운 저장소를 추가하고 더 이상 사용하지 않는 저장소를 제거할 수 있어야 합니다.

데이터 레이크 서비스 수행

데이터 레이크 서비스에서는 분석가, 데이터 사이언티스트, 개발자, 비즈니스 사용자가 데이터 레이크 저장소에 액세스하는 것을 제어하고 지원합니다. 데이터의 복사본을 동기화합니다. 데이터 레이크 서비스는 데이터 카탈로그를 포함하고 있으므로 사용자는 필요한 데이터를 찾아 각자의 업무에 적합한지 확인할 수 있습니다. 데이터 카탈로그는 데이터의 출처를 확인할 수 있는 데이터 계보도 제공하여 데이터에 대한 사용자의 신뢰도를 높입니다.

통합 정보 관리 및 거버넌스 모델 개발

데이터 레이크 서비스는 정보 관리 및 거버넌스를 수행하는 전문 미들웨어의 지원이 있어야 합니다. 이 미들웨어는 다음 기능을 갖추고 있어야 합니다.

- 데이터 이동 및 변환을 위한 프로비저닝 엔진
- 데이터를 다루는 사람들 간의 협업을 활성화하는 워크플로우 엔진
- 모니터링, 액세스 제어, 감사 기능

관리형 데이터 레이크 방식의 이점

데이터 레이크를 제대로 구축하는 기업은 데이터 환경에서 진정한 비즈니스 가치를 효율적으로 창출할 수 있습니다. 데이터 레이크를 구축하면 다음과 같은 비즈니스 이점을 누릴 수 있습니다.

- **전사적 범위에서 광범위한 데이터를 더 편리하게 액세스**: 관리형 데이터 레이크에서는 사용자가 온프레미스 및 클라우드에 있는 정형 데이터와 비정형 데이터에 모두 액세스할 수 있습니다. 많은 시간을 들여 IT 팀에 요청할 필요 없이 필요한 시점에 필요한 데이터에 액세스하면 됩니다.
- **더 신속한 데이터 준비**: 데이터 레이크는 여러모로 데이터 준비의 속도를 높입니다. 예를 들어 데이터에 대한 카탈로그를 구현하면 데이터에 대한 지식 및 이해가 확장되며 이를 바탕으로 데이터 준비에 속도를 낼 수 있습니다. 게다가 하이브리드 클라우드 인프라에서 데이터 레이크를 구축할 경우 데이터의 쓰임새에 가장 적합한 플랫폼에 데이터를 저장하는 것이 가능합니다. 데이터가 이상적인 위치에 저장되면 더 빨리 찾아 액세스할 수 있으므로 데이터 준비 및 재사용 속도도 빨라집니다.
- **향상된 민첩성**: 더 빨리 데이터를 준비하면 사용자가 더 많이 탐색할 수 있습니다. 데이터 레이크의 구성 요소는 샌드박스 형태로 제공되므로 사용자가 더 민첩하게 분석 모델을 개발하여 테스트할 수 있습니다. 실험적으로 분석하고 경우에 따라서는 "빠른 실패"를 통해 일찌감치 가장 생산적인 방안을 찾게 됩니다.
- **정확한 인사이트, 우수한 결정**: 데이터 레이크는 더 많은 데이터에 대한 액세스를 지원하고 데이터 준비 속도를 높이며 사용자의 실험적인 데이터 사용을 허용함으로써 기업이 정확도 높은 인사이트를 개발할 수 있도록 지원합니다. 잘 짜여진 데이터 레이크는 데이터 계보를 추적할 수 있으므로 신뢰할 만한 데이터인지 확인할 수 있습니다. 이러한 모든 기능을 활용하여 더 쉽게 비즈니스 의사결정을 내릴 수 있습니다.

뿐만 아니라 관리형 데이터 레이크는 IT에도 중요한 이점을 제공합니다. 즉, IT 팀은 데이터 레이크 방식을 통해 지속적인 데이터 증가에 대비할 수 있습니다.
IT 팀은 데이터 레이크 아키텍처에 Hadoop을 통합하여 확장성이 뛰어나고 경제적인 환경을 구축할 수 있으며, 여기서 (아마도 속도는 다소 느려지겠지만) 대규모로 분석을 수행할 수 있습니다. 데이터 레이크는 긴급한 비즈니스 질문에 답할 수 있도록 자원을 제공할 뿐 아니라 사용 빈도가 낮은 데이터를 위한 쿼리 가능 아카이브도 저렴한 가격에 제공할 수 있습니다. 하이브리드클라우드 환경에 데이터 레이크를 구축한 기업은 설비 투자 비용을 줄이면서 신속하게 자원을 추가할 수 있습니다.

대체로 데이터 레이크를 도입하면 값비싼 엔터프라이즈 데이터 웨어하우스(enterprise data warehouse, EDW) 자원의 부담을 덜어주므로 EDW가 다른 업무를 더 성공적으로 수행할 수 있습니다. 이를테면 비즈니스 분석가의 과거 실적 모니터링 및 분석 작업을 지원할 수 있습니다. 데이터 레이크가 EDW의 서비스 레벨 계약에 영향을 주지 않고 데이터 요청을 처리하므로 셀프 서비스 분석이 가능합니다.

신속한 분석, 민첩한 비즈니스 결정 지원

관리형 데이터 레이크는 오늘날 엄청난 규모로 유입되고 있는 데이터를 효과적으로 활용할 수 있는 방법을 제시합니다. 각 기업은 관리형 데이터 레이크 구축을 위한 핵심 모범 사례를 적용하여 전사적 범위에서 광범위한 데이터에 대해 즉각적인 액세스를 제공함과 동시에 데이터의 신뢰성 및 보안을 보장할 수 있습니다. 관리형 데이터 레이크가 기업의 구체적인 니즈에 따라 최적화된다면 비즈니스 민첩성을 강화하고 더 발전된 의사결정을 지원하는 데 중요한 역할을 할 수 있습니다.

추가 정보

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

IBM Analytics
Route 100
Somers, NY 10589

Produced in the United States of America
2016년 7월

IME14033-KRKO-00