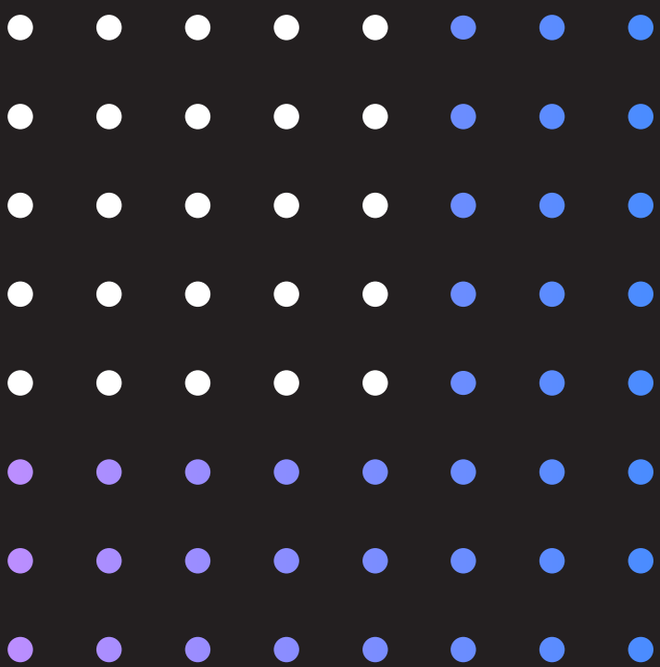
똑똑한 데이터 카탈로그 작성 및 데이터 레이크 거버넌스를 통해 비즈니스에 즉시 사용 가능한 데이터 제공

IBM Watson Knowledge Catalog 는 데이터 레이크의 문제 해결에 도움을 주는 머신 러닝 기반 데이터 거버넌스 플랫폼입니다.



IBM

목차

주요 요점

- 현재 신뢰할 수 있는 통찰을 위해 데이터를 구축하고 분석하고자 구축한 데이터 레이크에서 기업이 기대하는 가치를 온전히 누리는 기업은 많지 않습니다.
- DataOps는 데이터를 액세스, 준비, 통합 및 작성하고 소비자가 사용 가능하도록 만드는 과정에서 조직이 겪고 있는 비효율성 문제를 해결하는 동시에 기업 및 규정 당국의 정책을 준수할 수 있도록 해 줍니다.
- 데이터 레이크와 관련한 일반적인 어려움으로는 새로운 데이터 소스를 데이터 레이크로 가져올 때의 어려움 및 비용, 내부와 외부 데이터 세트 통합 능력의 부재, 데이터 거버넌스 관련 신뢰성 부족, 셀프 서비스 데이터 준비 도구 액세스 불가능, 데이터 레이크 내의 데이터를 찾고 이해하는 능력 부족 등이 있습니다.
- 카탈로그 작성, 데이터 품질 및 데이터 검색 기능을 갖춘 엔터프라이즈 데이터 거버넌스 플랫폼은 실패로 끝날지 모를 데이터 레이크 프로젝트를 비즈니스 가치 실현을 위한 원천으로 변모시켜 줍니다.
- IBM Watson® Knowledge Catalog (IBM Cloud Pak™ for Data 기반)는 데이터 검색, 데이터 카탈로그 생성, 데이터 품질 및 거버넌스를 지원하는 머신 러닝(ML) 카탈로그입니다. 데이터 사용자는 이를 활용해 데이터 자산, 데이터 세트 및 분석 모델을 신속히 검색하고, 큐레이션하고, 범주화하고, 공유할 수 있습니다.
- 조직에서 보유한 데이터를 면밀히 이해하지 못할 경우 이를 신뢰하기란 더욱 어려울 뿐만 아니라 ML 및 딥 러닝을 비롯한 모든 형태의 인공 지능(AI)으로 이 정보를 활용할 수도 없습니다.

DataOps 접근법으로 데이터 레이크 관련 문제 해결

10년 전, 모든 엔터프라이즈 데이터 클라우드가 상주하는 중앙 데이터 저장소를 구축하기 위한 유연하고 다양한 접근법을 찾는 데서 이 여정이 시작되었습니다. 그 해결책이 바로 모든 유형의 데이터를 저장할 수 있는 범용 데이터 스토리지인 데이터 레이크였습니다. 비즈니스 분석가와 데이터 과학자는 데이터 레이크를 이용해 적절한 분석 엔진과 도구 대부분을 각 데이터 세트에 위치 변경 없이 적용할 수 있습니다.

일반적으로 이러한 데이터 레이크는 Apache Hadoop 및 Hadoop Distributed File System(HDFS)을 Apache Hive 및 Apache Spark와 결합하여 구축되었습니다. 데이터 레이크의 규모가 커지면서 몇 가지 문제점이 드러났습니다. 기술 자체는 방대하고 다양하게 수집된 구조화/비구조화 데이터를 캡처, 저장 및 분석할 수 있었지만 이러한 기능을 비즈니스 워크플로에 실질적으로 도움이 되는 방식으로 적용하지 못했습니다.

2022년까지 80% 이상의 데이터 레이크 프로젝트가 가치를 실현하는 데 실패할 것으로 예상되며, 그 원인은 데이터를 찾고, 인벤토리에 저장하고, 큐레이션하는 작업이 분석 및 데이터 과학의 성공에 가장 큰 걸림돌이 되기 때문입니다.[1] 따라서 "데이터 레이크에 어떤 데이터를 저장해야 하는가?", "사용 대상은 누구인가?", "좀 더 쉽게 데이터를 찾을 방법은 무엇인가?", "이 데이터의 출처는 어디인가?", "데이터의 오용을 막을 방법은 무엇인가?"와 같은 의문이 묻혀 버린 경우가 많았습니다. 사람, 프로세스, 기술 문제를 해결하는 데 존재하는 이처럼 중대한 한계는 데이터 레이크 구현의 실패를 불러왔습니다.

오늘날 많은 조직이 실패를 깨닫고 데이터 레이크 구현을 이끄는 팀을 교체했으며 데이터 레이크의 성공적인 구현을 위한 두 번째, 세 번째 혹은 네 번째 시도를 이번에는 DataOps 데이터 운영을 통해 이어 가고 있습니다.

이 백서에서는 데이터 레이크에서 발생하는 일반적인 문제를 평가하고, DataOps와 같이 데이터 레이크를 데이터가 한 번 빠지면 나오지 못하는 늪에서 **비즈니스에 즉시 활용 가능한 데이터를 공급하는 파이프라인**으로 바꿀 수 있는 새로운 접근법을 제시합니다.

DataOps는 조직 전반에서 데이터 관리자와 데이터 소비자 간의 의사소통, 통합 및 데이터 플로 자동화에 중점을 둔 협업 기반 데이터 관리 절차입니다.

DataOps 소개

DataOps 는 여러 이해관계자들 간 데이터 플로를 협업 기반으로 개발 및 유지관리하는 방식으로 DevOps, 데이터 관리 및 데이터 거버넌스의 모범 사례를 공용 프레임에 도입한 것입니다. DataOps는 데이터를 액세스, 준비, 통합 및 작성하고 소비자가 사용 가능하도록 만드는 과정에서 조직이 겪는 비효율성 관련 문제를 해결하는 동시에 기업 및 규정 당국의 정책을 준수할 수 있도록 설계되었습니다.

이러한 비효율성은 사업부, 분석 팀 또는 운영 프로세스에도 존재할 수 있습니다.

이 방법을 따르려면 사람, 프로세스 및 기술 문제를 해결해야 하며 이러한 문제의 해결에 따라 바로 데이터 레이크 구현의 성패가 결정됩니다. 기술 측면에서 DataOps는 통제된 데이터 레이크 생성을 위해 데이터 수집 및 통합, 데이터 품질, 데이터 거버넌스 및 데이터 소비를 지원하는 완전히 통합된 엔드 투 엔드 플랫폼 사용의 중요성을 강조합니다. 데이터 품질 검증 규칙은 기업 전체에서 지속적인 데이터 파이프라인이 유지되도록 수집 프로세스의 일부로 자동 실행되어야 합니다. 수집 프로세스는 파이프라인의 핵심이 되는 데이터 카탈로그에 완전히 통합되어야 합니다. 데이터 소비자는 데이터 카탈로그에서 데이터 품질 점수 및 데이터 프로파일링 결과에 액세스할 수 있어야 하며 조직이 맥락에 따라 동일한 데이터를 사용하고 있다고 믿을 수 있어야 합니다.

데이터가 늘어나는 속도가 조직이 데이터에서 가치를 얻는 능력을 넘어서고 있습니다. 통찰 시스템 사용 시에 경험하는 가장 큰 어려움을 묻는 질문에 대한 조직의 응답은 1) 분석을 위해 기존 비즈니스 프로세스를 소스 데이터에 병합하는 것 40%, 2) 데이터의 규모 확장에 따르는 데이터 소싱, 수집, 관리 및 거버넌스 39%였습니다.[2] 오늘날 이것은 단순히 데이터 레이크 기술에 이미 투입된 막대한 시간과 리소스 투자를 보호하는 문제가 아니며, 사실은 다른 대안이 없는 것이 현실입니다. AI 구현에서 종합적인 분석 실행에 이르기까지 가능한 한 많은 데이터에 대한 총체적이고 종합적인 시야가 핵심이며, 이를 위해 모든 데이터를 한곳에서 보관하고 분석하며 통제하는 능력을 갖춘 아키텍처가 필요합니다. 많은 경우에 거버넌스를 갖춘 데이터 레이크가 이 모든 요구사항을 충족하는 유일하고 현실적인 옵션입니다.

현재는 DataOps를 위해 비즈니스에 활용할 수 있는 데이터 파이프라인을 확실히 지원함으로써 데이터 레이크에서 가치를 추출하는 방법을 찾아낼 수 있습니다. 그리고 반드시 찾아야 하기도 합니다.

엔터프라이즈 데이터 레이크 사용 시의 해결 과제

데이터 공유
기업 내의 한 팀이 데이터 세트를 새로 만들거나 수신할 경우, 가치 있는 데이터인 동시에 그와 관련한 민감성이 존재할 수 있습니다. 업무상 기밀 정보, 개인 식별 정보(PII) 또는 고객 데이터 등이 포함되어 있을 경우 팀에서는 해당 정보의 사용 허용 여부를 인지하고 팀원 중 누구도 이를 잘못 사용하지 않도록 조치를 취할 것입니다.

또한 팀 외부의 잠재적 데이터 사용자는 데이터의 가치 또는 데이터를 올바르게 사용하지 못했을 때의 위험을 잘 알지 못할 수 있다는 점도 의식할 것입니다. 이러한 위험으로 인해, 데이터를 공유하거나 팀원의 통제를 벗어난 위치에 저장하게 되는 경우 극도의 주의를 기울이게 됩니다.

이는 데이터 레이크에 부정적으로 작용합니다. 기업에서 데이터 레이크를 단순히 데이터를 쏟아부어 두고 관리되지 않는 장소로 생각하게 되면 가치 있는 데이터를 보관하기가 꺼려질 것입니다. 따라서 다른 부서에서 해당 데이터를 활용할 수가 없게 되고 기업 데이터 공유를 위한 데이터 레이크를 셀프 서비스 리포지토리로 사용하려는 계획은 무너집니다.

데이터 통합
팀에서 데이터를 데이터 레이크에 통합하는 데 동의했다 하더라도 통합 과정에서 어려움을 겪을 수 있습니다. 데이터 레이크의 원래 개념은 일반적 데이터 웨어하우스처럼 데이터를 복잡하게 추출, 변환 및 로드(ETL)하지 않고 원시 형태 그대로 가져오는 것입니다. 하지만 현실적으로 거의 모든 데이터 소스는 일정 수준의 재처리를 거쳐야 각종 의미 있는 분석에 유용하게 사용할 수 있습니다.

따라서 새로운 데이터 소스를 데이터 레이크로 통합하기 위해 흔히 몇 개월이 소요되곤 합니다. 이 데이터의 대부분은 이전에 엔터프라이즈 시스템이 아닌 소규모 운영 사일로에 보관되어 있었기 때문에 전체적으로는 통합할 소스가 수십 또는 수백 개에 달할 수 있습니다.

이는 곧 많은 경우에 비즈니스 분석가 또는 데이터 과학자가 필요로 하는 정보가 아직 데이터 레이크에 추가되지 않았을 수 있으며, 앞으로도 몇 개월 또는 몇 년이 지나도록 추가되지 않을 수 있다는 의미입니다. 이것 또한 데이터 레이크 도입을 저해하는 요소가 될 수 있습니다.

데이터 저장
지난 몇 년간 일반 스토리지 및 컴퓨팅 리소스의 비용이 크게 감소하기는 했지만 Hadoop 클러스터는 무료가 아닙니다. 막대한 양의 데이터를 데이터 레이크에 저장하는 것이 고성능 데이터 웨어하우스 기기에 저장하는 것보다 훨씬 가성비가 뛰어나다고는 해도 여전히 막대한 비용이 소요됩니다.

뿐만 아니라 데이터 웨어하우스에 저장된 일반적인 데이터와는 달리, 데이터 레이크에 보관된 빅 데이터는 볼륨 대비 가치가 상대적으로 떨어집니다. 가치 있는 바늘 고작 몇 개를 찾느라 대량의 건초 더미를 보관하는 형국이 될 수 있습니다.

데이터 과학자에게 정말 유용하고 가치 있는 데이터 세트가 무엇인지 알지 못할 경우, 데이터 레이크의 가장 밑바닥에 가라앉아 결코 사용되지 않는 데이터를 통합 및 저장하는 데 상당한 비용을 투자해야 할 수 있습니다.

데이터 검색
저장할 가장 중요할 데이터 세트를 파악해 이해관계자들에게 해당 데이터를 공유하도록 설득하고 이를 데이터 레이크 안에 통합하는 데 성공했다 하더라도, 다른 사용자가 해당 데이터를 적절하게 찾고 이해하고 사용할 수 있도록 하는 과정이 아직 남아 있습니다.

엔터프라이즈 데이터 레이크 사용 시의 해결 과제

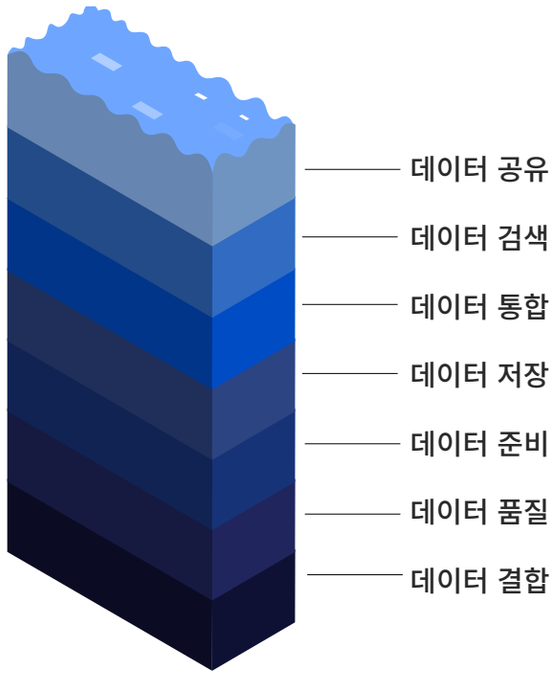


그림 1. 데이터 레이크 기술을 도입한 기업은 이러한 일반적인 문제 중 한 가지 이상을 경험할 수 있습니다.

데이터 레이크에 있는 데이터의 품질은 또 다른 문제입니다. 데이터의 품질이 높은지 낮은지 불확실한 상황에서도 데이터는 끊임없이 레이크에 쌓입니다.

안타깝게도 대부분의 데이터 레이크에서 데이터의 품질을 유지하기란 쉽지 않습니다. 아무런 맥락 없이 저장되는 데이터가 많아, 신규 사용자가 최초 소유자에게 확인하지 않고는 데이터를 이해하기가 어렵거나 불가능합니다. 흔히 용어가 영역별로 매우 달라서 한 비즈니스 영역에서 사용되는 지표가 다른 영역에서는 완전히 다른 이름으로 사용되거나 미묘하게 다른 방식으로 정의되어 있을 수도 있습니다. 혼란이나 오역이 발생할 가능성이 매우 커서 이러한 상황에 익숙하지 않은 분석가에게는 많은 데이터 세트가 사실상 아무 가치가 없거나 심지어 위험하기도 합니다.

내부와 외부 데이터 결합
마지막으로, 가장 큰 데이터 레이크라 할지라도 회사의 데이터 과학자가 사용하고자 하는 모든 가능한 데이터 세트를 보관하려고 해서는 안 됩니다. 예를 들어, 데이터 과학자가 특정 지역 관련 분석을 수행하거나 날씨 데이터 또는 주식 가격을 알고리즘에 통합하고 싶다고 해서 데이터 레이크에 Google Maps, Weather.com® 또는 Bloomberg의 완전한 복제본을 가져오는 것은 있을 수 없는 일입니다.

비즈니스 분석가가 분석에 필요한 모든 데이터가 데이터 레이크에 들어 있지는 않기 때문에 여러 애플리케이션에서 이를 검색하는 데 시간을 소비해야 할 것입니다. 유용한 분석의 대부분은 내부 데이터와 외부 데이터 세트를

결합했을 때 가능한 만큼, 이는 또 한 번 데이터 레이크에 대한 진입 장벽을 높이고 사용자가 인식하는 데이터 레이크의 가치를 떨어뜨리고 맙니다.

데이터 준비
데이터가 저장된 위치를 알아내는 것부터 데이터의 형식을 지정하는 것까지, 데이터 준비를 어렵게 하는 요인이 많습니다. 분석에 사용하기 위해 데이터를 준비하는 작업은 데이터 사용자에게 가장 비효율적이고 시간이 소요되는 일입니다. 데이터 사용자는 데이터 분석, 모델링, 비즈니스에 미칠 영향에 대한 통찰을 얻는 데 집중하는 대신 정보를 찾고 정리하고 형식을 지정하느라 대부분의 시간을 보냅니다.

준비 단계에서 통제된 데이터 세트에 대한 액세스 제한도 IT에 대한 과도한 의존을 불러옵니다. 이 제한된 액세스는 전사적으로 셀프 서비스 기능 및 데이터 활용 능력을 개선해 이 문제를 완화할 필요가 있다는 신호입니다.

데이터 품질
데이터 레이크에 데이터를 버리다시피 쏟아부으면 데이터가 쓸모없어집니다. 데이터 레이크에 유입되기 전에 데이터에 적용되는 데이터 품질 또는 검증 규칙이 없기 때문에 신뢰 및 사용 가능한 데이터를 마련하지 못합니다. 고품질 데이터는 의사결정을 내릴 때 데이터를 신뢰할 수 있는지 여부를 결정하는 필수 요소입니다. 데이터는 소중한 자산이며 조직 내에서 이동하기 때문에 관리가 필요합니다. 정보 소스는 점점 더 다양해지고 방대해지는 동시에 규제 준수 요건은 점차 까다로워지기 때문에, 이처럼 서로 다른 소스에서 일관성 있고 신뢰할 수 있으며 재사용 가능한 방식을 통한 정보의 통합 및 액세스가 매우 중요합니다.

종합적인 접근법을 통한 통제된 데이터 레이크 구축

대부분의 데이터 레이크는 데이터 스토리지 계층 및 분석 엔진에 Apache Hadoop과 그 주변의 폭넓은 오픈 소스 에코시스템을 활용합니다. Hadoop 관련 오픈 소스 커뮤니티는 당연히 현재 데이터 레이크 구현에서 발생하는 문제를 인식하고 있으며 최근에 다양한 문제를 개별적으로 해결하기 위한 여러 프로젝트가 출범했습니다. 마찬가지로, 동일한 문제를 해결한다고 말하는 수많은 독점 도구들이 시장에도 출시되어 있습니다.

데이터 레이크의 단편적인 문제 발생 시 이를 개선해 준다는 말은 유혹적입니다. 데이터 세트의 수가 관리하기 벅찰 정도로 늘어나면 카탈로그 지정 도구를 추가합니다. 사용자가 필요한 데이터를 찾을 수 없다는 불만을 제기하면 검색 기능을 갖춘 사용자 도구를 덧붙입니다. 데이터 담당자가 더 이상 데이터의 출처 또는 사용자를 추적 및 관리할 수 없을 경우 데이터 계보 도구 및 데이터 거버넌스 프레임워크를 배포합니다.

이러한 말들은 간단하게 들리지만 실제로 이렇게 단편적으로 접근하다 보면 특히나 데이터 레이크의 규모와 범위가 확장되면서 복잡성이 엄청나게 증가하고 유지보수가 어려워지기 마련입니다. 데이터 레이크에 새로운 데이터 소스를 추가하면 ETL 요구 사항의 복잡성이 늘어나는 것과 마찬가지로, 새로운 도구를 추가하면 데이터 레이크의 비기능적 요구 사항의 복잡성이 증가하게 됩니다.

데이터를 통합할 수 있는 통합 엔드 투 엔드 플랫폼을 구현하고 데이터에 대한 품질 관리를 수행하고 비즈니스 분석가가 효율적으로 사용할 수 있도록 데이터에 카탈로그를 지정하는 대신, 일반적으로 각 도구에서 오류를 관리하는 고유한 방법과 로그를 기록하는 자체적인 방식이 있다는 것을 알게 됩니다. 따라서, 문제 해결에 많은 시간이 소요될 수 있습니다.

또한, 단편적인 접근 방식의 더 중대한 단점은 데이터 레이크가 일반적으로 겪는 문제를 기술적인 측면보다 개념적인 측면을 중심으로 볼 때 더욱 명확히 드러납니다. 핵심은 확장성, 검색성, 통합, 데이터 품질 및 거버넌스는 별개의 문제가 아니라 서로 연관된 불가분의 관계라는 것입니다. 이와 관련한 문제를 해결하기 위해서는 훨씬 더 종합적인 접근법이 필요합니다.

확장성, 검색성, 통합, 데이터 품질 및 거버넌스는 별개의 문제가 아니라 서로 연관된 불가분의 관계입니다. 이와 관련한 문제를 해결하기 위해서는 정보 관리에 대해 훨씬 더 종합적인 방식으로 접근해야 합니다.

IBM Watson Knowledge Catalog 데이터 검색, 데이터 카탈로그 작성 및 데이터 품질

IBM Watson Knowledge Catalog (IBM Cloud Pak for Data 기반)는 데이터 사용자가 데이터 자산, 데이터 세트, 분석 모델 및 조직의 다른 구성원과의 관계를 신속히 찾고 큐레이션하고 범주별로 분류하고 공유하는 데 도움을 줍니다. 데이터 거버넌스 팀이 비즈니스 용어, 정책 및 규칙을 정의하고 거버넌스를 위한 고급 워크플로를 제공할 수 있도록 지원합니다. 카탈로그는 믿을 수 있고 확신을 가지고 사용할 수 있는 데이터에 대해 셀프 서비스 액세스를 확보할 수 있도록 데이터 엔지니어, 데이터 담당자, 데이터 과학자 및 비즈니스 분석가를 위한 단일 정보 소스가 되어 줍니다.

IBM Watson Knowledge Catalog(IBM Cloud Pak for Data 기반)와 같은 솔루션은 오늘날 데이터 레이크에서 발생하는 주요 문제들을 해결하는 데 필요한 모든 기능을 포괄적인 단일 플랫폼으로 제공할 수 있습니다. 카탈로그는 이러한 상호 연관된 문제, 데이터 레이크의 광범위한 오류들의 근본 원인을 찾아 메타데이터를 캡처, 저장 및 관리하고 데이터 계보를 추적하는 효과적인 도구를 제공할 수 있습니다.

여러 면에서 데이터 레이크의 가치는 데이터 그 자체의 중요성만큼이나 그 안에 들어 있는 메타데이터에 의해 결정됩니다. 메타데이터에 데이터 세트의 출처, 데이터 생성자, 포함된 내용, 사용 권한의 범위, 사용 방법이 지정되지 않을 경우 실질적으로 데이터는 쓸모가 없습니다. 사용자는 데이터를 찾을 수 없고, 찾는다 하더라도 데이터가 무슨 내용인지 이해할 수 없으며 신뢰할 수 없고 사용 방법을 알 수 없게 됩니다.

Watson Knowledge Catalog
믿을 수 있고 의미 있는 데이터 제공

데이터 구성

확인

데이터는 완전하고 적용이 가능하며 어디에서나 액세스할 수 있어야 합니다. 모든 유형의 데이터를 검색, 분류, 이해하십시오.

데이터 관리

신뢰

신뢰할 수 있는 셀프 서비스 액세스를 확대하려면 데이터가 안전하고 잘 정리되어 찾기가 쉬워야 합니다. 데이터의 출처와 품질을 확인하십시오.

자유로운 데이터 검색

사용

기업의 성장을 위해 셀프 서비스 검색을 유도하고 의사결정을 자동화하는 기능이 필요합니다. 필요한 사용자에게 모든 정보가 담긴 뷰를 제공하고 액세스 권한을 제공하십시오.

그림 2. IBM Watson Knowledge Catalog는 데이터 검색, 데이터 카탈로그 지정 및 데이터 거버넌스를 위한 기능을 폭넓게 제공합니다.

단일 정보 소스 및 단일 액세스 포인트

IBM Watson Knowledge Catalog(IBM Cloud Pak for Data 기반)는 메타데이터를 최우선으로 하여 이러한 문제를 해결합니다. 이 제품의 핵심은 데이터 레이크, 데이터 웨어하우스 또는 트랜잭션 시스템, 일련의 스프레드시트를 비롯해 데이터가 상주하는 위치에 관계없이 기업이 액세스해야 하는 모든 데이터 세트 및 분석 자산에 색인을 생성하는 강력한 카탈로그 엔진입니다. 구조화 또는 비구조화 데이터 여부 또는 온프레미스 저장 또는 클라우드 호스팅 여부에 구애를 받지 않습니다. 또한 회사에서 구독 중인 독점 데이터 서비스 또는 개방형 데이터 API와 같은 외부 데이터 세트 및 소스도 카탈로그에 포함할 수 있습니다.

데이터 카탈로그는 모든 데이터 세트에 관한 단일 정보 소스를 제공할 뿐만 아니라 단일 액세스 포인트도 제공합니다. AI 기반 검색 및 제안 기능은 비즈니스 분석가, 데이터 과학자, 데이터 품질 엔지니어, 데이터 거버넌스 팀이 좀 더 쉽게 자산을 찾고 사용할 수 있는 메타데이터를 제공함으로써 사용자가 찾은 내용을 이해하고 유용한 정보인지 여부를 평가할 수 있도록 지원합니다.

기본 제공되는 셀프 서비스 데이터 준비 기능이 분석 및 AI 애플리케이션에서 실제 사용을 위해 데이터를 변환하는 데 소요되는 시간을 단축해 주기 때문에 비즈니스 분석가 및 데이터 과학자가 데이터 준비 및 분석에 시간을 낭비할 필요가 없습니다. IBM® InfoSphere® Advanced Data Preparation과 같은 전사적 데이터 준비 솔루션과 통합하면 카탈로그를 통해 생성된 통제된 데이터 세트가 비즈니스 사용자를 위한 비즈니스 통찰 및 실행 방안을 이끌어 낼 대부분의 맥락에 부합하도록 할 수 있습니다. 더 나아가 이 통합으로 데이터 파이프라인에서의 협업을 확장할 수 있습니다.

확장성, 검색성, 통합, 데이터 품질 및 거버넌스는 별개의 문제가 아니라 서로 연관된 불가분의 관계입니다. 이와 관련한 문제를 해결하기 위해서는 정보 관리에 대해 훨씬 더 종합적인 방식으로 접근해야 합니다.

카탈로그는 또한 데이터 세트 태그 지정 및 분류, 계보 및 사용 자동 추적, 데이터 전체에서 비즈니스 용어 표준화를 위한 기본 제공 비즈니스 용어집을 통해 최고 데이터 책임자(CDO) 사무실에 속한 데이터 담당자를 지원할 수 있습니다. 이를 통해 담당자는 각 데이터 세트에 포함된 내용이 무엇인지, 민감한 데이터 또는 PII는 어디에 있는지, 액세스를 허용할 대상은 누구인지 손쉽게 파악할 수 있습니다.

조직 내부 및 외부의 여러 데이터 소스를 위한 단일 카탈로그

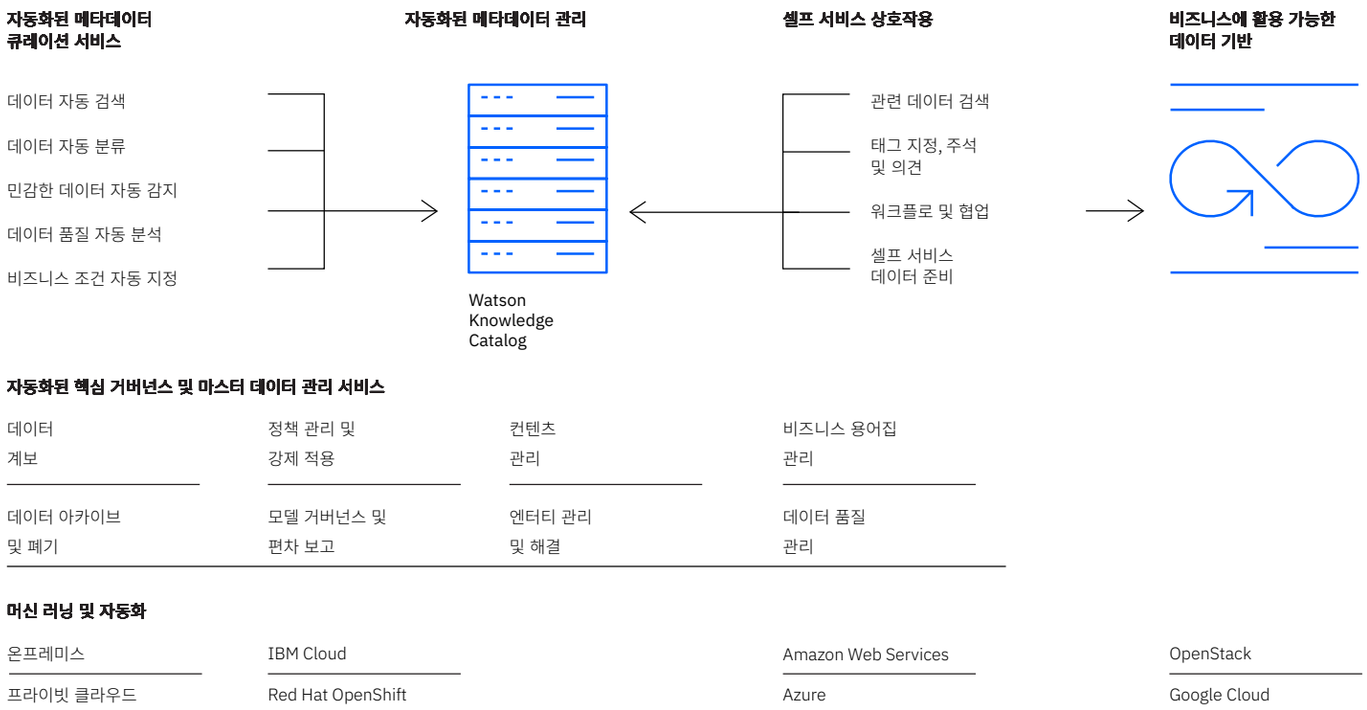


그림 3. IBM Watson Knowledge Catalog 지능형 메타데이터 색인 덕분에 구조화 및 비구조화 데이터 모두 원본 시스템에 상주하는 상태에서 사용자가 더욱 스마트한 분석을 위해 이를 신속히 검색할 수 있습니다.

IBM Watson Knowledge Catalog는 메타데이터에 최우선 순위를 부여하여 비즈니스가 액세스하는 모든 데이터 세트에 대한 단일 정보 소스 및 단일 액세스 포인트를 제공합니다.

지능형 데이터 검색 기능 기본 제공

검색 성능의 추가적인 향상을 위해 사용자가 카탈로그에서 데이터 세트 및 분석 자산에 대해 태그를 지정하고 의견을 추가함으로써 메타데이터가 더욱 풍성해지고 동료들이 필요한 데이터를 찾는 데 추가적인 맥락을 제공할 수 있습니다. 이 솔루션에는 또한 ML을 사용해 각 데이터 세트의 컨텐츠를 자동으로 분류하는 검색 알고리즘이 기본적으로 내장되어 있습니다. 이름, 주소, 우편번호, 주민등록번호와 같은 일반 필드 유형을 식별해 작성자가 데이터에 수동으로 주석을 달아야 할 필요를 줄여 줍니다. 자동화 및 ML을 적용해 데이터 큐레이션 및 메타데이터 관리를 자동화합니다. 기본 제공되는 데이터 품질 기능을 기반으로 솔루션에서 심층 데이터 프로파일링, 데이터 품질 및 검증 규칙을 활성화할 수 있습니다.

자동화된 데이터 운영은 데이터 품질과 거버넌스를 갖춘 큐레이션된 데이터 파이프라인을 제공하며 지속적으로 고품질의 통제된 데이터가 데이터 레이크로 유입되도록 보장합니다.

마찬가지로, 지능형 메타데이터 자산 모델을 추가하면 독보적인 방법으로 일반 데이터 보호 규정(GDPR) 및 캘리포니아 소비자 개인 정보 보호법(CCPA)과 같은 규정을 자동으로 적용할 수 있습니다.

IBM Watson Knowledge Catalog(IBM Cloud Pak for Data 기반)는 비즈니스에 활용이 가능하며 신뢰할 수 있는 고품질 데이터를 기본적으로 모든 데이터 사용자에게 제공합니다.

솔루션의 모든 구성요소는 단일 설계 원칙과 일반적인 접근법을 통해 비기능적 요구사항, 즉 확장성, 오류 관리, 보안, 로깅 등을 처리하도록 마이크로서비스로 설계되었습니다.

IBM Watson Knowledge Catalog는 대규모 AI 구현을 위한 ML 엔터프라이즈 거버넌스 플랫폼입니다.

단편적인 DIY 방식의 접근법으로 인한 혼란스러운 오류와 성능의 병목 현상이 없는 IBM Watson Knowledge Catalog는 ML 엔터프라이즈 거버넌스 플랫폼이기 때문에 AI를 대규모로 구현할 수 있습니다.

IBM Watson Knowledge Catalog 는 다음과 같은 3가지 유형으로 제공됩니다.

- IBM Cloud™의 SaaS 솔루션 형태
- IBM Cloud Pak for Data 에 포함되어 있는 형태
- IBM Watson Studio 와 통합된 형태

IBM Watson Knowledge Catalog와 같은 솔루션은 데이터 레이크 프로젝트가 당초 추구했던 가치를 실현할 수 있도록 해 줍니다. 지능형 카탈로그 지정 및 거버넌스 기능을 제공하는 Watson Knowledge Catalog는 AI를 위해 신뢰할 수 있고 통제된 데이터 레이크 구축을 지원합니다.

AI를 위한 통제된 데이터 레이크 구축에 따르는 4가지 장점

1. 품질 및 거버넌스를 통해 데이터에 대한 신뢰와 확신 구축

- 데이터 품질 기능은 데이터의 품질을 개선하고 데이터 레이크에서 고품질 데이터를 사용할 수 있도록 해 줍니다.
- 거버넌스 정책은 자동으로 설정 및 적용되기 때문에 데이터 세트를 검색할 때 사용자가 해당 데이터 사용 가능 여부 및 방법을 알 수 있습니다.
- 사용자가 추가하는 평가, 의견 및 기타 정보를 바탕으로 데이터를 큐레이션할 수 있으며 이를 통해 다른 사용자가 데이터 세트의 유용성을 판단하는 데 도움이 됩니다.

2. 데이터 사용자 지원

- 회사의 기간 업무(LOB) 팀은 데이터가 적절히 통제되고 오용으로부터 보호되고 있다는 확신이 있기 때문에 데이터를 적극적으로 공유합니다.
- 동적인 데이터 정책과 적용을 통해 협업을 수행하고 데이터를 신뢰할 수 있는 기업 자산으로 바꿀 수 있습니다.
- 시간이 지날수록 사용자가 관련 태그와 메타데이터를 추가하면서 데이터 검색이 더욱 용이해지고 재사용이 가능해져 다른 사용자가 해당 데이터로부터 가치를 추출할 수 있습니다.
- 단일 인터페이스를 통해 데이터가 어디에 저장되어 있더라도 조직이 소유한 모든 데이터 세트에 액세스할 수 있습니다.

3. 시간 절감

- 자동 데이터 검색 기능으로 새로운 데이터 세트에 메타데이터를 추가하는 데 투입해야 하는 시간과 노력을 줄일 수 있습니다.
- 자동 데이터 큐레이션 및 메타데이터 관리는 메타데이터를 찾고 조건을 지정하는 시간을 줄여 주며 비즈니스 용어집 생성 시간도 단축됩니다.
- 단순하고 직관적인 셀프 서비스 데이터 준비 도구를 사용하면 데이터 사용자가 적은 시간에 데이터를 준비해 통찰을 확보하는 데 집중할 수 있습니다.
- 데이터 과학자와 비즈니스 분석가가 짧은 기간에 더욱 효과적인 분석을 제공할 수 있도록 지원합니다.
- 다른 팀에서 필요한 데이터를 제공해 줄 때까지 몇 주를 기다리는 대신 단 몇 초 만에 필요한 데이터를 찾을 수 있도록 AI 기반의 지능형 검색 도구가 제공됩니다.

4. 데이터 및 비용 증가 관리

- 데이터 레이크로 가치가 떨어지는 데이터 세트를 수집하는 비용이 발생하지 않아 스토리지 비용을 최적화할 수 있습니다.
- 또한 조직이 구독하는 모든 외부 데이터 세트를 파악할 수 있기 때문에 필요 이상의 구독 비용을 지불할 위험이 감소합니다.
- 사용자의 데이터 수요를 바탕으로 데이터 레이크에 새로운 데이터 소스 수집의 우선 순위를 결정할 수 있어 가장 중요한 소스를 먼저 통합할 수 있습니다.

데이터 가치 극대화

여러분이 CDO의 사무실 소속이든, IT 부서 소속이든, LOB 데이터 과학자거나 분석가든 관계없이 여러분과 동료들은 공통의 목표를 보유하고 있습니다. 당초의 가치가 그대로 실현되는 데이터 레이크를 구현할 수 있다면 업무를 더욱 쉽고 생산적으로 수행할 수 있을 뿐만 아니라 기업이 경쟁자를 제치고 앞서 나가는 데 핵심 역할을 수행할 수 있을 것입니다.

경쟁사가 여전히 늪에서 헤매는 동안 데이터 레이크의 수질을 청정하게 유지할 수 있다면 그 동안은 꿈에 지나지 않았던 가능성이 활짝 열릴 것입니다. 기존에 가려졌던 데이터의 가치를 가장 먼저 알아보는 사람만이 가장 먼저 도전하는 사람에게 주어지는 혜택을 누릴 수 있습니다.

결론

모든 데이터의 상주 위치, 데이터의 사용자, 분석으로 얻어 낼 비즈니스의 가치를 알고 있어야 합니다.

DataOps 추진 과제의 핵심은 데이터 거버넌스, 품질 및 활성 정책 관리의 통합으로 자동화된 개방형 메타데이터 관리를 제공하도록 지원할 수 있는 데이터 카탈로그입니다.

지능형 카탈로그 지정 및 거버넌스 기능을 갖춘 IBM Watson Knowledge Catalog는 AI를 위해 신뢰할 수 있고 통제된 데이터 레이크 구축을 지원합니다. 카탈로그는 데이터 레이크 환경에서 데이터 통합, 데이터 품질 및 거버넌스를 기본 제공하기 때문에 DataOps에 필요한 비즈니스용 데이터와 단일 정보 소스를 제공할 수 있습니다.

추가 정보

자세한 내용 확인:
ibm.com/cloud/watson-knowledge-catalog

150-945
서울특별시 영등포구 국제금융로 10
서울국제금융센터 (Three IFC) 한국 아이비엠(주)

미국에서 제작, 2019 년 10월 IBM, IBM 로고, **ibm.com**, IBM Cloud, IBM Cloud Pak, IBM Watson 및 InfoSphere는 전 세계에 등록되어 있는 International Business Machines Corp.의 상표입니다.

Red Hat 및 OpenShift는 미국 및 기타 국가에서 Red Hat, Inc. 또는 그 자회사의 상표 또는 등록상표입니다. 기타 제품 및 서비스 이름은 IBM 또는 타사의 상표일 수 있습니다. 최신 IBM 상표 목록은 웹 "저작권 및 상표 정보(www.ibm.com/legal/copytrade.shtml)"에 있습니다.

1.Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders–Gartner, 2019년 9월
2. Forrester Wave: Machine Learning Data Catalogs, 2018년 2분기

ASW12449-KRKO-03