Informatica

백서

데이터 기반 지능형 엔터프라이즈를 위한 인공 지능

CLAIRE의 기계 학습 기반 혁신이 데이터 관리 분야에서 새로운 발전을 이끌어 내는 방법

Informatica 정보

디지털 변환은 우리의 기대치를 변화시켰습니다. 이제 더 적은 비용으로 더 나은 서비스를 더 빠르게 제공할 수 있어야 합니다. 이러한 상황에 부응하기 위해서 기업이 변화해야 하며, 데이터가 이에 대한 해답을 가지고 있습니다.

세계적인 선도 기업인 Informatica는 엔터프라이즈 클라우드 데이터 관리 분야의 모든 부문, 카테고리, 틈새시장에서 지능적인 방식으로 고객을 지원할 준비가 되어 있습니다. Informatica는 귀하의 조직이 더욱 유연하고 기민하게 운영되고 새로운 성장 기회를 발굴하며 나아가 놀라운 혁신을 이룰 수 있도록 통찰력을 제공합니다. 또한 당사는 모든 종류의 데이터에 100% 초점을 맞추어 귀하의 성공에 필요한 다양한 서비스를 제공하고 있습니다.

Informatica가 제공하는 모든 서비스와 솔루션에 관해 알아보십시오. 그리고, 데이터의 힘을 활용하여 미래의 차세대 지능형 혁신을 주도하시기 바랍니다.

목차

"데이터 및 분석 리더는 데이터 환경의 복잡성에 직면하고 있습니다. 당사의 데이터 관리 솔루션 예측 성능은 클라우드 기능, 연결된 데이터 아키텍처, 메타데이터 및 AI 애플리케이션을 통한 일상적/비일상적 작업의 자동화에 핵심이 되는 개발 작업과 증가하는 수요를 인지하고 있습니다."[1]

— Gartner

AI의 중요성

인공 지능(AI)과 기계 학습(ML)은 전 세계 모든 업계에서 발생하는 디지털 변혁을 주도하고 있습니다. AI는 비즈니스를 혁신하기 위한 전략으로 이사회 임원들이 가장 중요하게 생각하는 것입니다. 또한 AI는 우리가 감상하는 영화부터 우리가 운전하는 자동차에 이르기까지, 우리의 일상생활을 향상시키는 데 스며들었습니다. AI/ML은 생명 과학 분야에서 새로운 치료법을 검색하고, 금융 서비스에서 사기/위험을 줄이며, 진정으로 개인화된 고객 경험을 제공하는 데 중요합니다.

비즈니스 리더에게는 AI/ML이 마법처럼 보일 수 있습니다. 잠재적인 영향력은 분명하지만, AI/ML 또는 이러한 강력한 혁신을 활용할 최선의 방법을 전적으로 이해하지 못할 수 있습니다. AI/ML은 차선책, 고객 만족도 추적, 효율적인 운영, 혁신적인 제품 등 새로운 여러 비즈니스 솔루션의 기반 기술입니다. 일반적으로 머신 러닝, 특히 딥 러닝은 데이터가 부족합니다. 요구되는 정확성을 확보하려면 ML에 방대한 양의 교육 데이터가 필요합니다. 이 데이터는 현재 비즈니스 상태를 정확하게 반영해야 합니다. 불안정하거나 제한된 데이터로 교육받은 AI는 비즈니스 이니셔티브에 부정적인 영향을 주고 원하는 결과에 역효과를 줄 수 있습니다.

올바른 기능을 사용하고 교육받은 효과적인 AI를 이용하려면 기업 내/외부의 광범위한 데이터를 활용해야 합니다. 이 데이터는 ML 모델을 구축하고 교육할 수 있는 방식으로 함께 가져와야 합니다. 그러려면 데이터 관리가 필요합니다. 이는 규모와 복잡성을 처리하는 문제일 뿐만 아니라 신뢰에 관한 문제이기도 합니다. 모델을 교육하는 데 사용되는 데이터가 올바른 시스템에서 제공되고 있습니까? 개인 식별 정보(PII)를 제거하고 모든 규정을 준수했습니까? 투명하며, 모델이 사용하는 데이터의 계보(Lineage)를 입증할 수 있습니까? 데이터에 편견이 없음을 문서화하여 규제 기관이나 조사관에게 보여 줄 수 있습니까? 이 모든 일에는 적절한 제어와 데이터 관리 기반이 필요합니다. 견고한 데이터 관리 기반이 없으면 AI는 이해할 수도, 신뢰할 수도 없습니다. 즉, 데이터 관리 기능이 없으면 AI가 의도하지 않은 결과를 초래하는 블랙박스가 될 수 있습니다.

AI에 필요한 데이터

AI의 성공 여부는 데이터 과학자가 교육하고 확장할 수 있도록 설계된 모델의 효과에 달려 있습니다. 그리고 이러한 모델의 성공 여부는 적절한시기에 신뢰할 수 있는 데이터의 사용 가능성에 따라 좌우됩니다.

AI/ML 모델을 구축하는 데이터 과학자에게 고품질 데이터가 필요한 이유는 무엇입니까? 예를 들어, 소비자의 행동을 예측하는 예측 모델을 생각해 보십시오. 이러한 모델의 유용한 기능은 우편 번호로 표시되는 소비자 위치일 수 있습니다. 그러나 우편 번호 데이터가 누락되거나, 불완전하거나, 부정확하다면 어떻겠습니까? 모델의 동작이 교육과 구축에 모두 악영향을 끼쳐 예측을 잘못하게 되고 전체 노력의 가치가 감소할 수 있습니다. 정확하고 완전하며 유효성이 검사된 우편 번호는 개인의 시장 세분화, 소득 등급, 연령, 기대 수명 등을 예측하는 데에도 도움이 될 수 있습니다. 그러므로 더욱 올바르게 이해해야 합니다. 우리는 모두 '설명 가능한 AI'가 단순한 옵션이 아닌 규제된 명령이 될 것으로 예상해야 합니다. 메타데이터 기반의 계보(Lineage)와 추적 가능성 없이는 AI 기반 애플리케이션과 통찰력을 프로덕션에 구축할 수 없습니다.

[1] Gartner, 2020년 예측: 데이터 관리 솔루션, Rick Greenwald, Donald Feinberg, Mark Beyer, Adam Ronthal, Melody Chien, 2019년 12월 5일.

AI는 모델의 모든 기능을 빠르게 찾고, AI 모델의 요구 사항(기능 확장, 표준화 등)을 충족하기 위해 데이터를 자동으로 변환할 뿐만 아니라 데이터 중복을 제거하고 고객과 환자, 파트너 및 제품에 대해 신뢰할 수 있는 마스터 데이터를 제공하며, 모델 및 해당 운영 범위 내 정보를 포함하여 데이터의 엔드 투 엔드 계보(Lineage)를 제공하는 데 지능형 데이터 관리 기능을 사용합니다. AI의 성공 여부는 데이터 과학자가 교육하고 확장할 수 있도록 설계된 모델의 효과에 달려 있습니다. 그리고 이러한 모델의 성공 여부는 적시에 신뢰할 수 있는 사용 가능성에 따라 좌우됩니다.

데이터에 필요한 AI

AI/ML은 데이터 관리 관행을 확장하는 데에도 중요한 역할을 담당합니다. 디지털 변혁에 필요한 데이터의 양이 방대한 까닭에 기업에서 관련성, 가치, 보안 성능을 보증하고 투명성을 보장하려면 가장 관련성 높은 데이터와 메타데이터를 검색하고 카탈로그화해야 합니다. 기업은 이 데이터를 정제하고 마스터링해야 합니다. 그리고 이 데이터를 효과적으로 관리하고 보호해야 합니다. 데이터가 효과적으로 관리 및 확장되지 않으면 AI/ML 모델은 지난 30년간 기존의 모든 데이터 웨어하우징 이니셔티브와 마찬가지로 품질이 낮은 데이터를 사용하고 신뢰할 수 없는 통찰력을 제공하는 일을 겪게 될 것입니다.

최근 연구에 따르면, 데이터 센터 트래픽의 전체 볼륨이 2021년에 20.6제타바이트에 이를 것으로 예상되며 연결된 장치와 연결 개수가 2022년까지 250억 개 이상에 달할 것으로 추정됩니다[2]. 이 모든 데이터는 거버넌스 정책을 준수하면서 처리되며 사용 가능하고 신뢰할 수 있어야 합니다. 이 모든 것에 더해 비즈니스 전략과 프로세스의 변화에 신속하게 조치를 취하고 대응해야 합니다. 디지털 변혁 이니셔티브에 사용할 데이터를 준비하는 일과 관련된 활동에서는 데이터의 양이 증가함에 따라 복잡성도 늘어났습니다. LinkedIn에 따르면, 데이터 과학자는 미국에서 가장 유망한 직종 중 하나로 꼽힙니다.[3] 게다가 기업이 찾는 데이터 엔지니어의 수는 최근 해마다 96%의 변화를 보였습니다.[4] 그러나 채용만으로 데이터 양의 증가 문제를 관리하기에는 부족합니다.

급증하는 문제에 선형적으로 접근하지 않기

단순히 문제에 더 많은 엔지니어와 개발자를 투입해서 이러한 문제를 해결할 수는 없습니다. 이는 사람의 기준에서 선형적으로 해결할 수 있는 문제가 아닙니다. 전통적인 접근 방식은 비효율성으로 점철되어 있습니다. 프로젝트는 엔드 투 엔드 메타데이터를 거의 볼 수 없고 자동화 기능이 제한된 사일로에서 구현됩니다. 학습 과정이 없고, 처리 비용이 많이 들며, 거버넌스 및 프라이버시 단계가 여러 차례 반복됩니다. 그렇다면 기업은 어떻게 비즈니스 속도에 맞춰 움직이고, 셀프서비스를 활성화하며, 고객에게 더 나은 서비스를 제공하고, 운영 효율성을 높이며, 빠르게 혁신할 수 있습니까?

[2] Cisco, Global Cloud Index Forecast and Complete Visual Networking Index Forecast(글로벌 클라우드 지수 예측 및 전체 비주얼 네트워킹 지수 예측).
[3] LinkedIn, 'LinkedIn's Most Promising Jobs of 2019(LinkedIn이 꼽은 2019년 최고 유망 직종).'
[4] Datanmi, 'Data Engineering Continues to Move the Employment Needle(고용 바늘을 계속 움직이게 하는 데이터 엔지니어링)'

바로 여기에서 AI가 빛을 발합니다. AI는 데이터 검색, 통합, 정제, 거버넌스 및 마스터링 전반에 걸쳐 데이터 관리와 관련된 작업을 자동화하고 단순화할 수 있습니다. 기계 학습 방법은 일상적인 반복 작업을 학습하고 인계받아 개발자와 사용자가 자유롭게 고부가가치의 혁신적인 프로젝트를 수행할 수 있도록 합니다. AI는 데이터 이해 능력을 개선하고 데이터 프라이버시 및 품질 이상 현상을 식별합니다. AI는 개발자, 분석가, 스튜어드 및 현업 부서 사용자에게 완벽한 파트너로, 추천 사항과 차선책을 통해 자동화하고 보강함으로써 작업 속도를 높여 줍니다.

AI가 전체 데이터 환경에서 엔드 투 엔드 프로세스를 가속화하는 데 어떻게 도움을 줄 수 있는지 생각해 보면, AI가 가장 효과적입니다. 그렇기 때문에 데이터 관리에 AI는 필수적이라고 생각합니다. Informatica®가 메타데이터 기반 AI 기능인 CLAIRE® 엔진에 막대한 혁신 투자를 집중해 온 이유가 바로 여기에 있습니다. CLAIRE는 엔터프라이즈 통합 메타데이터를 모두 활용하여 일상적인 데이터 관리 작업을 자동화하고 확장합니다.

데이터 관리 부문에서 AI가 선사하는 네 가지 주요 이점

일반적으로 AI는 데이터 전문가의 생산성 향상, 운영 효율성 향상, 더욱 지능적으로 안내되는 데이터 경험 및 심도 있는 지식 제공, 데이터 거버넌스 프로세스 속도 향상 등 네 가지 주된 방식으로 데이터 관리팀에 도움이 됩니다. 아래에는 오늘날 도움을 줄 수 있는 몇 가지 항목의 예가 나와 있습니다.

생산성: 데이터 통합용 추천 시스템은 데이터 엔지니어가 데이터를 추출, 변환하고 전달하기 위한 매핑을 신속하게 구축하는 데 도움이 됩니다. 추천 시스템은 기존 매핑을 기반으로 학습하고, 데이터베이스 및 파일 시스템의 비즈니스 콘텐츠를 이해하며, 대상 시스템 및 데이터 소비자에게 전달하기 전에 데이터를 표준화하고 정제할 수 있는 적절한 변환 작업을 제안합니다.

효율성: 일반 엔터프라이즈에서는 매일 수천 개의 데이터 통합 프로세스를 실행하고 있습니다. 이러한 프로세스의 모니터링은 주로 수동적이며, 소요된 시간과 소모된 CPU 및 메모리만 기록하는 관리 툴을 사용합니다. AI는 로그 및 모니터링 파일에 있는 시계열 데이터의 기록값을 통해 학습할 수 있으며, 이상 수치 값을 사전에 플래그하고 사전에 처리하지 않으면 발생할 수 있는 문제를 예측할 수 있습니다.

데이터 경험: 실제 엔터티(예: 환자 기록 또는 판매 주문)가 데이터베이스 또는 파일 집합에 저장되면 해당 데이터가 여러 테이블이나 파일로 분할되고 분산되어 스토리지 및 성능에 맞게 최적화됩니다. AI는 데이터 간의 관계를 탐지하고 원래 엔터티를 신속하게 재구성할 수 있습니다. 사용자가 기본 키/외래 키 관계에 관한 오래된 문서를 기억하거나 검색할 필요가 없고, 다양한 데이터 세트를 수동으로 결합할 필요도 없습니다. 또한 AI는 유사한 데이터 세트를 식별하고 사용 패턴, 데이터 품질 및 크라우드 소싱 협업을 기반으로 추천 사항을 만들 수 있습니다.

데이터 거버넌스: 데이터 거버넌스의 공통적이면서도 지루한 단계는 물리적 데이터 요소에 비즈니스 용어를 연결하여 데이터 요소에 대한 비즈니스 컨텍스트 및 관련성을 설정하고 사용자가 데이터를 이해할 수 있도록 하는 과정입니다. 대부분의 경우 AI는 자연어 처리(NLP) 기술과 비즈니스 유형 식별의 조합을 사용하여 물리적 데이터에 비즈니스 용어를 자동으로 연결할 수 있습니다. 이렇게 하면 오류가 발생하기 쉬운 이 작업의 단순 업무를 대폭 줄일 수 있습니다. 클라우드 시대에 이러한 접근 방식이 SaaS 애플리케이션에도 효과적이라는 점을 인지하는 것은 중요합니다. 메타데이터는 Salesforce 및 Workday와 같은 SaaS 애플리케이션에서 수집되어 엔터프라이즈 카탈로그에 추가될 수 있습니다.

AI 기반 데이터 관리: 뱅킹 사례

AI에 데이터 관리가 필요한 이유와 데이터에 AI가 필요한 이유를 설명하기 위해 뱅킹 서비스를 예로 들어 보겠습니다.

점점 더 많은 고급, 예측 및 실시간 분석용 데이터에 AI를 적용하여 은행이 얻을 수 있는 이점은 다음과 같습니다.

- 고객 유지율을 높여 주는 보다 개인화된 서비스 제공
- POS에서 발생하는 사기오류 거래 감소
- 소비자 투자자의 성과는 높이는 동시에 자산 고문의 비용은 낮춤
- 프로젝트 관련 규정 준수 비용 절감

데이터 관리 관점에서 AI는 ERP, CRM, 클라우드 및 웹 앱, 머신 및 로그 파일, 타사 데이터 등과 같은 모든 유형의 관련 데이터를 자동으로 검색하고 카탈로그화할 수 있습니다. 이를 통해 데이터 과학자는 소비자 행동, 사기 활동, 위험에 대한 소비자 성향과 일치하는 투자 기회 등과 관련된 통찰력을 보여 주는 패턴을 찾기 위해 수백 번 실험해야 하는 모든 데이터에 빠르게 액세스할 수 있습니다.

AI는 데이터 관리와 관련해 고객 데이터 간의 관계를 확인하고 특정 인물에 대한 통찰력을 일치시킴으로써 고객 및 관심 대상자(POI)에 대한 360도 뷰를 자동으로 보완할 수 있습니다. 이를 통해 기업은 관련성이 더 높은 제안으로 고객과 더 잘 소통하고 온라인, 모바일 또는 휴대폰 등의 다양한 채널에서 원활하게 경험을 제공할 수 있습니다. POI에 대한 360도 뷰를 통해 은행은 사기 행위의 패턴과 네트워크를 훨씬 더 빠르게 검색하여 잠재적으로 수백만 달러를 절약할 수 있습니다.

또한 AI는 데이터 통합및 데이터 품질 작업을 자동화하고 안내하여 수백 개의 데이터 소스에서 데이터를 결합하고 정제함으로써 분석 모델 및 알고리즘의 예측력을 높일 수 있습니다. AI/ML 및 고급 분석과 결합된 우수한 많은 데이터를 통해 차선책을 개선하고 사기를 식별하는 등의 중요한 결과를 산출하는 것으로 입증되었습니다.

또한 AI는 정책이 문서화될 뿐만 아니라 실제로 시행되도록 보장하는 데이터 거버넌스를 지원합니다. 이를 통해 정보 보안 전문가는 GDPR(일반 데이터 보호 규정), SOX(Sarbanes-Oxley Act), Basel II 및 Basel III 등과 같은 데이터 프라이버시 규정을 준수할 수 있습니다.

Informatica CLAIRE: Intelligent Data Platform의 '인텔리전스'

기계 학습을 통해 데이터 관리 생산성을 이끌어 내기 위한 Informatica의 접근 방식은 다음과 같습니다.

1. Intelligent Data Management Cloud™: Informatica는 생산성을 극대화하기 위해 통합된 엔드 투 엔드 데이터 관리 플랫폼을 제공했습니다. 이 통합 플랫폼은 통합 연결성, 메타데이터 및 운영 관리를 제공함으로써 새로운 데이터 관리 프로젝트의 개발 및 구축 작업을 가속화합니다. 이 플랫폼은 기업 내, 클라우드, 멀티 클라우드 및 멀티 하이브리드 소스 전반에 걸쳐 데이터를 관리하기 위해 강력하고 일관된 기능 세트를 제공합니다. 이 통합 데이터 관리 플랫폼을 Intelligent Data Management Cloud라고 합니다.

이 플랫폼은 모듈식입니다. 단일 툴에서 시작하여 고객의 속도에 맞춰 확장합니다.

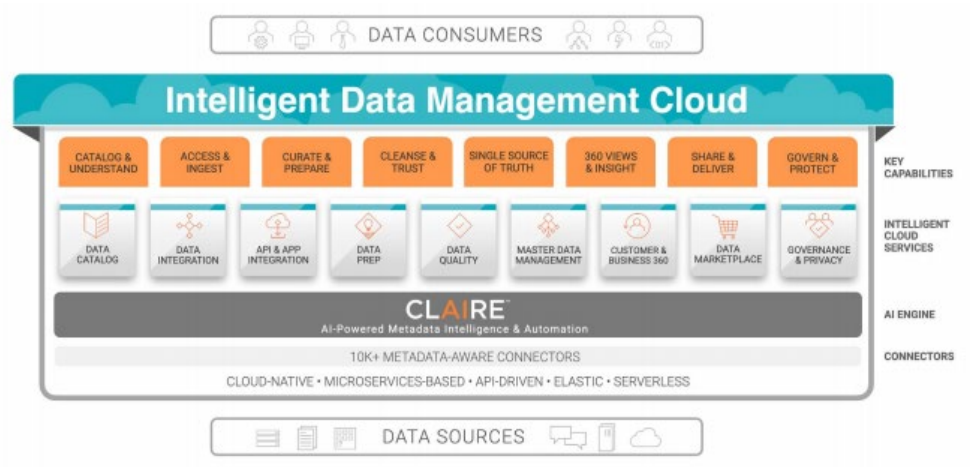

그림 1: Intelligent Data Management Cloud는 데이터 관리 기능을 공유된 연결성, 작업 통찰력, 데이터 및 메타데이터 인텔리전스와 통합합니다.

2. 메타데이터: Informatica는 오랫동안 기술 및 비즈니스 메타데이터 관리 부문의 선두 주자로 알려져 왔습니다. Informatica는 기업 전반에서 보다 광범위한 메타데이터를 수집함으로써 이 부문에서 역량을 확장했습니다. 여기에는 다음이 포함됩니다.
   - 데이터베이스 테이블, 열 정보, 데이터 프로필 통계, 스크립트 및 데이터 계보(Lineage) 등의 기술 메타데이터
   - 의미, 관련성 및 다양한 비즈니스 프로세스와 기능에 대한 중요도 등 데이터 관련 컨텍스트를 캡처하는 비즈니스 메타데이터
   - 다음과 같은 질문에 답할 수 있는 시스템 및 프로세스 실행에 관한 운영 메타데이터: 데이터가 마지막으로 업데이트된 시기가 언제인가요? 로드 프로세스를 마지막으로 실행한 시기는 언제인가요? 가장 많이 액세스한 데이터에 대한 정보는 무엇인가요?
   - 액세스한 데이터세트, 클릭한 검색 결과, 제공된 평가 또는 의견 등 사용자 활동에 대한 사용량 메타데이터

이렇게 광범위한 메타데이터를 수집하는 일이 기계 학습의 핵심입니다. 이는 기계 학습 알고리즘을 '교육'하는 데 사용되는 데이터세트를 제공하며 더 나은 결과를 얻고 조정할 수 있도록 합니다.

3. 인텔리전스: Informatica는 메타데이터 및 AI/기계 학습을 CLAIRE와 결합하여 통합된 기능을 제공하고 있습니다.

   Intelligent Data Management Cloud에서 수집한 메타데이터는 CLAIRE의 알고리즘에서 엔터프라이즈 데이터 환경에 대해 학습하는 데 사용할 수 있는 방대한 양의 정보를 제공합니다. 이 지식은 CLAIRE에서 지능형 추천 정보를 제공하고 데이터 관리 프로젝트의 개발/모니터링을 자동화하며 엔터프라이즈 내외부의 변화에 적응하는 데 도움이 됩니다. CLAIRE는 Intelligent Data Management Cloud에서 모든 데이터 관리 기능의 인텔리전스를 주도합니다.

CLAIRE는 다음과 같이 광범위한 사용자를 지원합니다.

- 데이터 엔지니어는 많은 구현 작업을 부분적으로 또는 완전히 자동화된 상태에서 찾을 수 있습니다.
- 데이터 분석가는 필요로 하는 데이터의 위치를 훨씬 쉽게 찾아 데이터를 준비할 수 있습니다.
- 현업 부서 사용자는 규정된 데이터 거버넌스와 규정 준수 통제의 영향을 받는 데이터를 신속하게 식별합니다.
- 데이터 과학자는 데이터를 보다 신속하게 이해합니다.
- 데이터 관리자는 고품질의 데이터를 보다 쉽게 시각화합니다.
- 데이터 보안 및 프라이버시 전문가는 보다 간단하게 데이터 잘못된 사용 상황을 감지하고 민감한 데이터를 보호하며 적절하게 통제된 상태를 유지하는지 입증할 수 있습니다.
- 관리자와 작업자는 데이터 관리 프로세스의 성능 최적화 및 예측 유지보수 기능을 활용할 수 있습니다.

다음은 CLAIRE를 통해 제공된 인텔리전스가 오늘날 어떻게 활용되고 있는지 보여 주는 몇 가지 예시입니다.

데이터 카탈로그화에 사용되는 CLAIRE

보유한 데이터를 검색하고 파악하는 일은 모든 데이터 기반 이니셔티브에서 실시할 첫 번째 단계입니다. CLAIRE는 기계 학습 기반의 검색 엔진을 제공하여 엔터프라이즈 전체에서 데이터 자산을 스캔하고 카탈로그화합니다. CLAIRE가 제공하는 지능형 데이터 카탈로그는 데이터 과학자, 분석가 및 데이터 엔지니어가 필요한 데이터를 찾고 추천하는 데 도움이 되어 데이터 검색 및 준비에 소요되는 시간을 크게 줄일 수 있습니다.

고급 관계 검색

주요 데이터 카탈로그화 및 데이터 모델링 작업 중 하나는 데이터세트 간의 관계를 문서화하는 것입니다. CLAIRE는 기계 학습 기술을 사용해 구조화된 데이터세트에서 기본 키, 고유 키 및 조인을 자동으로 식별합니다. 이렇게 하면 문서화하는 데 소요되는 수개월의 노력이 몇 분으로 줄어듭니다. CLAIRE는 데이터 큐레이션 프로세스에 사람을 포함하여 관계를 식별하는 능력을 지속적으로 향상시킵니다. 예를 들어, 사용자는 추론된 관계를 수락하거나 거부할 수 있으며 CLAIRE는 이러한 작업을 통해 학습합니다.

예컨대 어떤 고객이 마케팅 캠페인에 반응할 가능성이 가장 높은지에 대한 보고서를 작성하는 은행의 데이터 분석가는 모든 고객에 대한 기존 상품/대출 정보를 검색할 수 있어야 합니다. 그러나 엔터프라이즈 전반에 걸쳐 데이터가 고립되어 있기 때문에 부서 및 데이터 저장소에서 이러한 데이터세트를 찾기가 어렵습니다. CLAIRE는 데이터베이스에 문서화된 결합, BI 및 ETL 같은 다른 툴에서 수행된 결합, 데이터 값에서 파생된 통계를 사용하여 데이터 분석가에게 조인을 추론하고 추천합니다. 이는 사용자의 분석을 확장하고 사용 가능한 모든 정보를 사용하여 캠페인에 적합한 대상 고객을 찾는 데 도움이 됩니다.

CLAIRE는 키 및 조인 검색 시 여러 기술을 결합합니다. 키의 경우 고유성, null 개수, 열 메타데이터(예 : 'ID'가 포함된 열 이름) 등의 프로파일링 통계와 기타 정보를 결합해 기본 키와 고유 키를 검색합니다. 조인 및 조인 키 추론은 열 서명 분석과 같은 기계 학습 기술을 결합해 사용함으로써 수많은 잠재적 데이터세트에서 대규모 조인을 검색합니다.

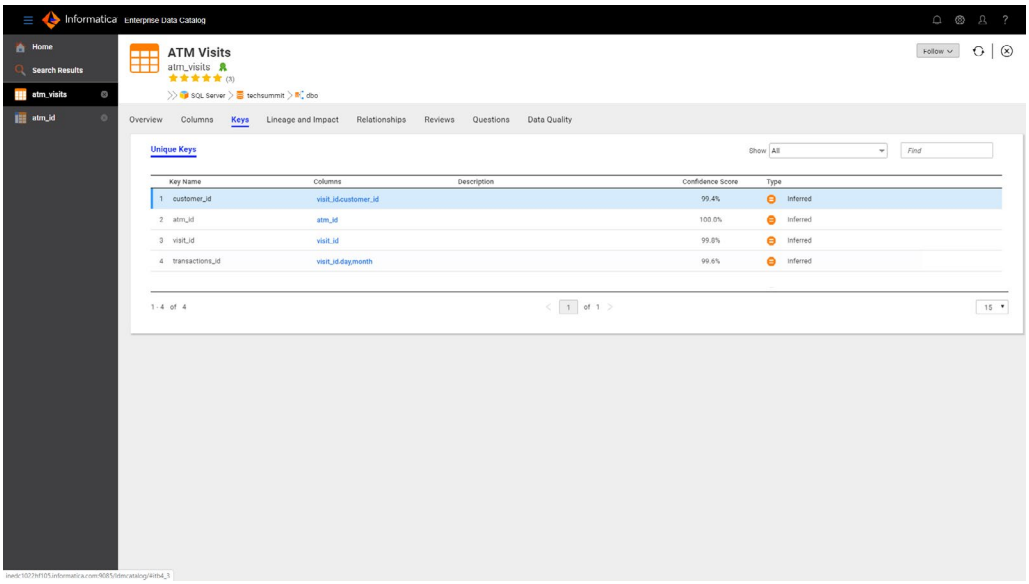

그림 2: 기계 학습 기술을 사용한 추론을 통해 고유 키 검색

지능형 데이터 유사성

CLAIRE는 클러스터링 같은 기계 학습 기술을 사용하여 수천 개의 데이터베이스 및 파일 세트 전체에서 유사한 데이터를 감지합니다. 지능형 데이터 유사성은 데이터 식별, 중복 항목 감지, 비즈니스 엔터티에 개별 데이터 필드 결합, 데이터세트 전체에서 태그 전달, 사용자에게 데이터세트 추천 등 다양한 목적으로 사용되는 핵심 기능 중 하나입니다.

데이터 유사성은 두 개 열에 있는 어떤 데이터가 어느 정도 동일한지에 대해 컴퓨팅을 수행합니다. 엔터프라이즈 설정(열 전체가 1억 개라고 가정)에서 모든 2개의 열 쌍을 시도하고 비교하는 무차별 대입 공격과 같은 접근 방식은 계산에 무리가 따릅니다. 대신, 데이터 유사성은 기계 학습 기술을 활용하여 유사한 열을 클러스터링하고 유사한 일치 항목을 식별합니다.

이 프로세스는 여러 단계에서 수행됩니다. 먼저 열 특성을 기준으로 열을 클러스터링합니다. 그런 다음 클러스터 각각에서 데이터 중복을 계산해 고유한 값을 산출합니다. 마지막으로 Bray-Curtis 및 Jaccard 계수를 사용하여 데이터 유사성을 계산하기 위해 가장 유력한 쌍을 선택합니다.

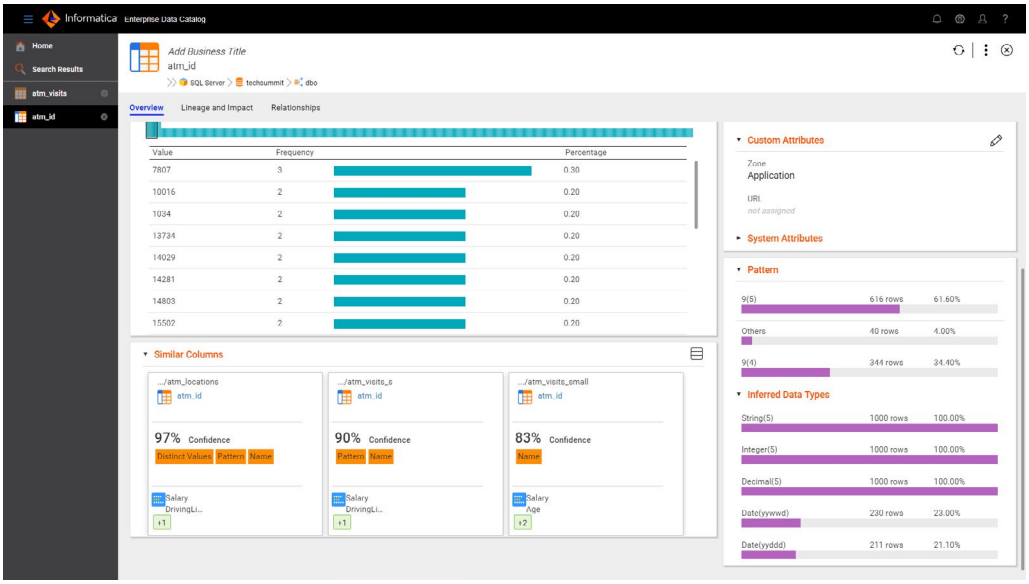

그림 3: 클러스터링과 Bray-Curtis 및 Jaccard 계수를 사용하여 유사한 열 식별

태그를 이용한 지능형 도메인 검색

CLAIRE는 각 열에 의미 라벨을 적용하여 데이터 필드를 분류할 수 있습니다. 이러한 의미 라벨은 데이터 도메인이라고 불립니다.

그림 4: CLAIRE에서 데이터 필드를 자동으로 분류하고 태그라는 의미 라벨을 적용합니다.

일반적으로 의미 라벨은 정규식, 참조 테이블 또는 기타 복잡한 핸드 코딩 로직에 기반한 규칙을 평가함으로써 적용됩니다. 이렇게 수천 개의 규칙을 정의하고 유지하는 것은 지루한 작업입니다.

대신 CLAIRE는 태그 개념을 사용하여 데이터 필드를 검색하고 라벨 지정하는 프로세스를 매우 간소화합니다. 아직 분류되지 않은 열의 경우, 사용자는 열 내용을 표시하는 간단한 태그(예를 들어, "보험금 날짜")를 제공하기만 하면 됩니다. 시스템은 연계를 통해 학습한 다음 이 태그를 모든 유사한 열에 자동으로 전달합니다. 데이터 기술의 '안면 인식'은 Facebook 사진에서 사람들에게 태그를 지정하는 것과 유사하며, 이는 수백만 개의 다른 사진에서 동일한 사람에게 태그를 지정하는 그물망 효과가 있습니다.

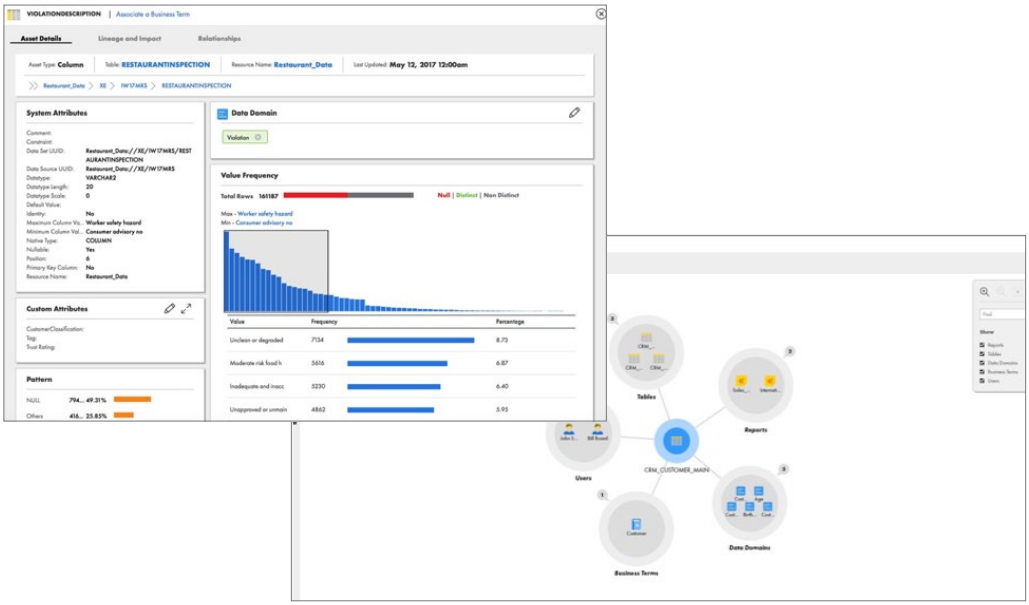

그림 5: 자동 데이터 분류.

지능형 엔터티 검색

일단 열의 도메인을 식별하면 CLAIRE는 이 개별 필드를 높은 수준의 비즈니스 항목으로 모을 수 있습니다. 아래 예시는 구매 주문이라고 하는 항목이 고객 및 제품으로 식별된 필드를 결합하여 생성되는 방식을 보여줍니다. 엔터티 검색은 사용자가 분석 또는 데이터 통합 프로세스에서 이종 데이터 필드를 모으는 방법을 통해 학습하며, 학습한 이 내용을 적용하여 엔터프라이즈 데이터 환경 전반에서 엔터티를 도출합니다.

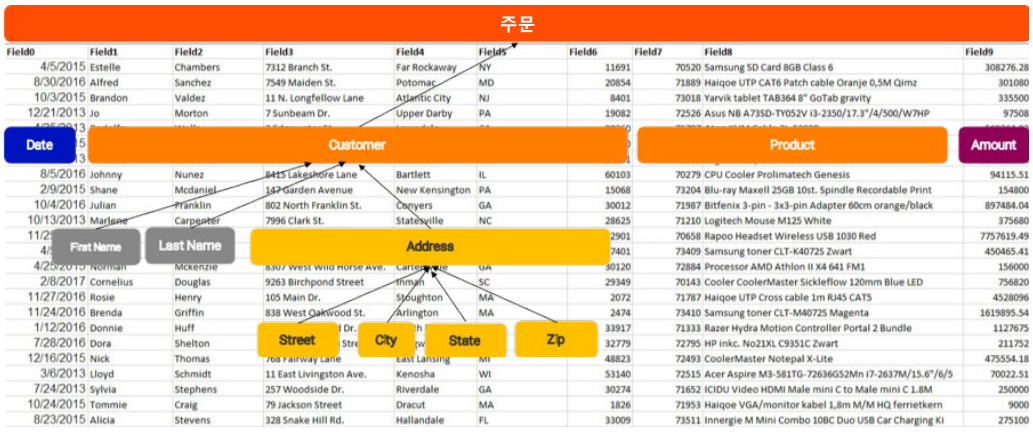

그림 6: 테이블과 파일에서 엔터티를 감지하기 위한 데이터 도메인 결합

분석에 사용되는 CLAIRE

CLAIRE 기반 자동화 및 인텔리전스는 분석 통찰력을 얻고 프로세스를 진행하는 속도를 크게 높이고, 데이터 감사 기능을 향상시키며, 분석에 필요한 데이터 준비 과정을 간소화합니다. CLAIRE는 데이터 파이프라인 추천 사항과 복잡하고 다양한 정형 데이터를 자동으로 파싱하는 기능을 통해 데이터 엔지니어링 생산성을 향상시킵니다.

변환 추천 사항

다음 변환 및 표현식을 예측하고 데이터 통합 매핑 생성 과정을 자동화하여 설계를 마무리 짓고 데이터 엔지니어의 생산성을 향상시킵니다. 기업에서 CLAIRE 기반 추천 사항을 받기로 선택하면, 기업의 데이터 파이프라인에서 익명의 메타데이터를 분석하고 AI/ML을 적용해 설계 추천 사항을 제공합니다. 이 메타데이터는 변환 및 표현식 추천 사항을 생성하는 데 사용됩니다. CLAIRE는 추천 사항을 수락하거나 거부할 때마다 더 좋아집니다. 이를 통해 개발을 가속화하고 반복 작업을 자동화하며 더 많은 유형의 사용자가 데이터를 빠르게 연결하고 통합할 수 있습니다.

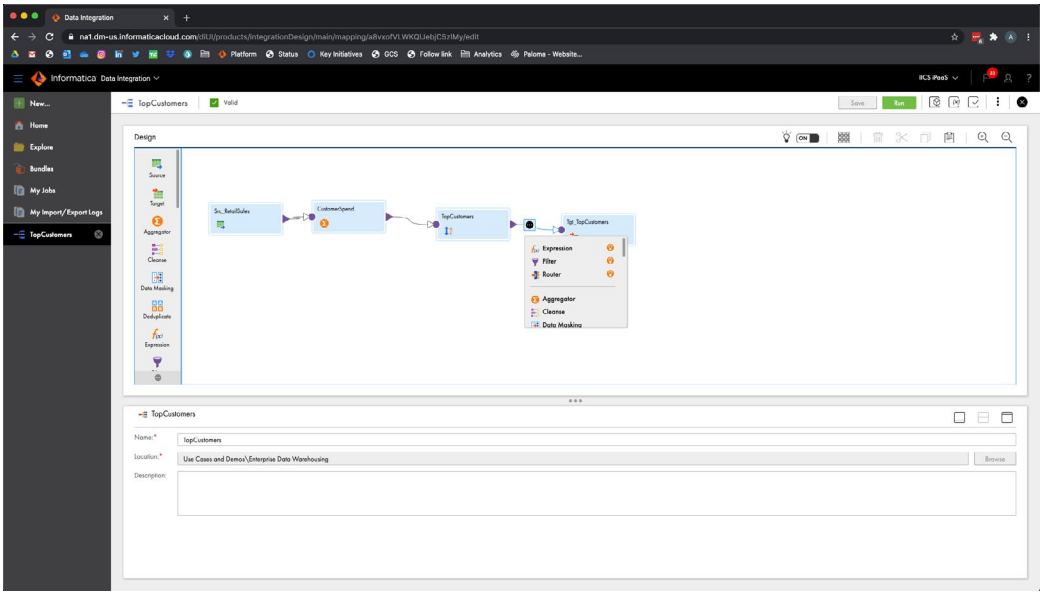

그림 7: CLAIRE는 데이터 파이프라인을 만들 때 차선의 변환 작업을 추천합니다.

대규모 프로세스 실행 최적화

CLAIRE는 다양한 최적화 방법을 사용하여 데이터 파이프라인 전체의 통합 성능을 향상시킵니다. 스마트 옵티마이저는 성능 특성에 따라 빅 데이터 워크로드를 실행하기 위한 최상의 처리 엔진을 결정합니다. 매핑 추천 사항은 과거 사용자 활동을 기반으로 데이터 엔지니어에게 제공되며, 비용 기반 최적화 프로그램과 경험적 지식을 통해 데이터 파이프라인의 조인 순서를 지능적으로 변경하여 성능을 최적화합니다. 이는 CLAIRE가 데이터 파이프라인을 최적화하는 방법의 몇 가지 예시에 불과합니다.

조인 열 추천 사항

CLAIRE는 사용자가 두 데이터세트를 결합하는 작업을 선택하면 Join-Column(예: 조인 키)을 자동으로 제안합니다. 그러면 데이터 분석가가 분석하기 위해 데이터세트를 복합 데이터세트로 병합하는 최고의 방법을 결정하기 위해 수작업에 수백 시간을 허비하지 않아도 됩니다. CLAIRE는 데이터 레이크로 가져온 데이터세트의 원래 소스 시스템(예: Oracle과 같은 관계형 데이터베이스)에 정의된 기본/외래 키 관계(즉, Pk-Fk)로 시작합니다. 동일한 데이터세트가 다른 프로젝트에 결합된 경우 추천 시 이 Join-Column 정보도 사용됩니다. CLAIRE에서 이 모든 정보를 처리하고 순위를 매겨 두 데이터세트 간에 가장 좋은 조인 열을 추천합니다. 또한 데이터세트 샘플링을 기반으로 제안된 열 간의 데이터 중복 비율도 표시됩니다.

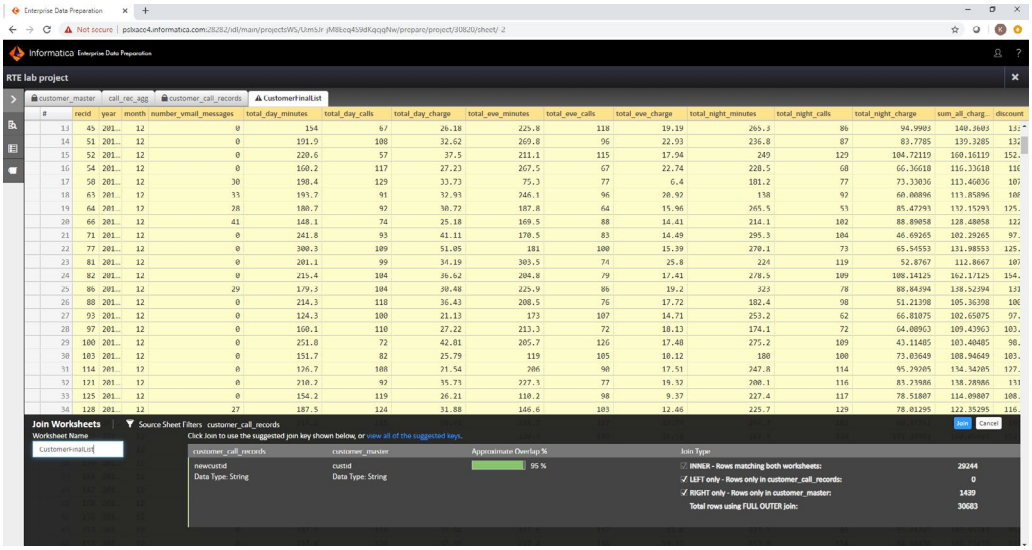

그림 8: 두 데이터세트를 결합할 때 자동 Join-Column 추천

Apache Zeppelin 시각화 추천 사항

Informatica Enterprise Data Preparation에서는 Apache Zeppelin을 사용하여 그래프와 차트가 포함된 노트북 형태로 워크시트를 확인할 수 있습니다. 사용자가 게시물의 노트북을 열면 CLAIRE 시각화 추천 사항을 볼 수 있습니다. 사용자가 게시 후 처음으로 노트북을 열면 파생된 숫자 열의 히스토그램 시각화를 볼 수 있습니다. 게시물에 파생된 숫자 열이 포함되어 있지 않은 경우, 사용자에게 노트북의 첫 번째 단락에 'SELECT * FROM' 테이블 쿼리가 표시됩니다.

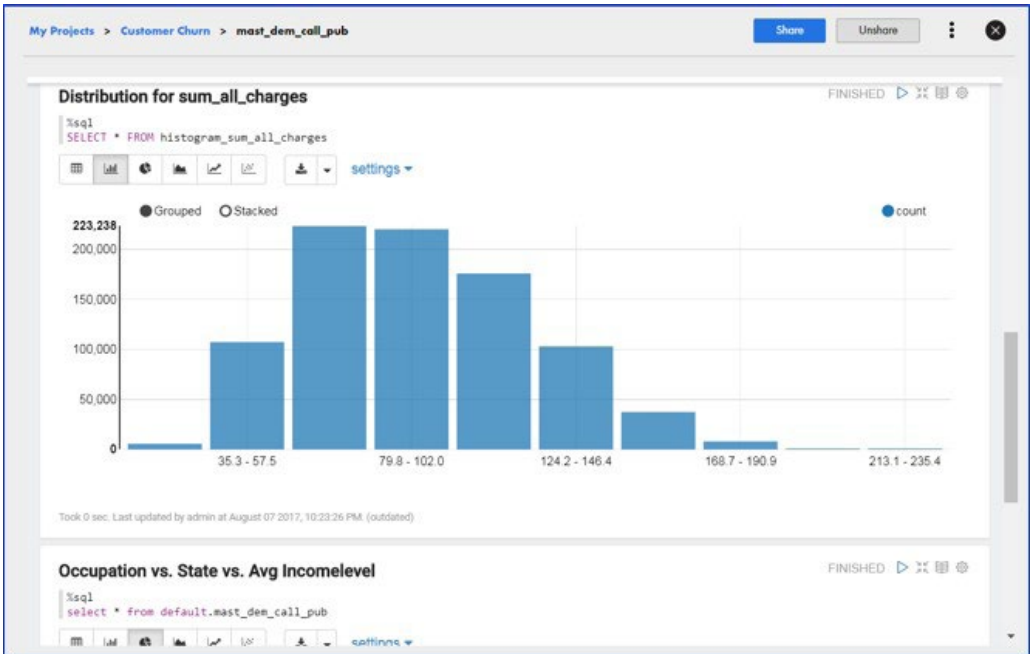

그림 9: Apache Zeppelin 노트북에 추천되는 시각화

지능형 데이터 추천 사항

CLAIRE는 데이터 분석가와 데이터 과학자에게 프로젝트에서 사용할 데이터세트를 제안합니다. 또한 사용자가 선택한 데이터세트를 관찰하고 더 우수한 평가를 받은 유사한 데이터세트 또는 사용 중인 데이터세트를 보완할 수 있는 추가 데이터세트를 제시합니다. 지능형 데이터 추천은 사용자가 다수의 동료가 이미 수행했던 동일한 작업을 반복하는 것을 방지하는 데 도움이 됩니다. 추천은 다음을 포함합니다.

- 동일한 데이터의 준비된 버전(대체 가능 데이터)
- 동일한 유형의 레코드를 포함하는 다른 테이블(통합 가능 데이터)
- 추가 특성을 사용하여 결합 시 데이터를 보완할 수 있는 테이블(결합 가능 데이터)

데이터 추천 시 콘텐츠 기반 필터링 기술을 사용하여 추가 데이터세트에 관한 사항을 제안합니다. 데이터세트에 사용된 특성(조건)에는 계보(Lineage) 정보, 사용자 순위 및 데이터 유사성이 포함됩니다. 여러 유사성 측정값을 사용해 다양한 데이터세트의 등가성 점수를 매깁니다. 이 점수를 토대로 속성이 유사한 데이터세트를 추천합니다. 다양한 사용자가 일반적으로 함께 사용되는 데이터세트를 찾기 위해 메타데이터 그래프를 쿼리하는 방식으로 보완 항목을 추천합니다.

지능형 구조 발견

점점 더 많은 양의 데이터가 비정형 또는 비관계형 형식으로 기기, 엔터프라이즈 및 애플리케이션에서 생성되고 수집됩니다. 이러한 데이터 유형의 특징으로는 대용량, 벨로시티, 다양성 및 변동성을 꼽을 수 있습니다. 현재 '데이터 드리프트'라는 용어는 이같이 새로운 데이터 유형에서 데이터의 형식, 속도 및 콘텐츠의 변동을 나타내는 데 일반적으로 사용됩니다.

CLAIRE로 구동되는 Informatica ISD(Intelligent Structure Discovery)는 파일 수집/온보딩 프로세스를 자동화하여 엔터프라이즈가 복잡한 파일을 검색 및 파싱할 수 있도록 설계되었습니다. ISD는 클릭스트림, IoT 로그, CSV, 텍스트 구분, XML, JSON, Excel, ORC, Parquet, Avro, PDF 양식 및 Word 테이블 파일 등 다양한 데이터 파일 형식을 기본적으로 지원합니다. CLAIRE는 이러한 파일에서 구조를 자동으로 파생하여 한층 수월하게 파악하고 작업할 수 있게 해 줍니다. 콘텐츠 기반 접근 방식을 사용하여 파일을 구문 분석하면 파일 처리에 영향을 미치지 않고 빈번하게 발생하는 파일 변경 사항에 적응할 수 있습니다.

ISD는 파일에서 패턴 인지를 자동화하기 위해 유전 알고리즘(genetic algorithm)을 사용합니다. 이 접근 방식은 결과를 개선하기 위해 '진화'라는 개념을 사용합니다. 각 후보 솔루션에는 더 적합한 솔루션을 제공하는지 여부를 판별하기 위해 자동으로 변경 및 테스트될 수 있는 속성 세트가 있습니다. 그러면 결과 구조에서 입력 범위 및 파생된 도메인과 같은 여러 요인을 기준으로 점수를 매깁니다. 최고 점수의 구조는 점수가 향상되는지 여부를 확인하기 위해 구조를 결합하는 등 여러 가지 변경 사항을 구조에 적용하는 '변형' 단계에 들어갑니다. 데이터 구조의 적합성 여부가 판단되면 프로세스를 종료합니다.

또한 ISD는 커스터마이징 ML 기반 NER(명명된 엔터티 인식) 및 NLU(자연어 이해) 메커니즘을 사용해 필드 및 필드 유형을 식별하므로, 통합 후 작업을 단순화하고 외부 애플리케이션에서 ISD를 기본 NER/NLU 엔진으로 사용할 수 있도록 합니다. 예를 들어, ISD는 수신/발신 API 페이로드에서 PII 정보를 감지하는 데 사용되며 규정을 준수하고 더 높은 엔터프라이즈 보안을 구현할 수 있게 해 줍니다. ISD는 데이터 검색, 수집 및 스트리밍 활용 사례에서도 사용됩니다.

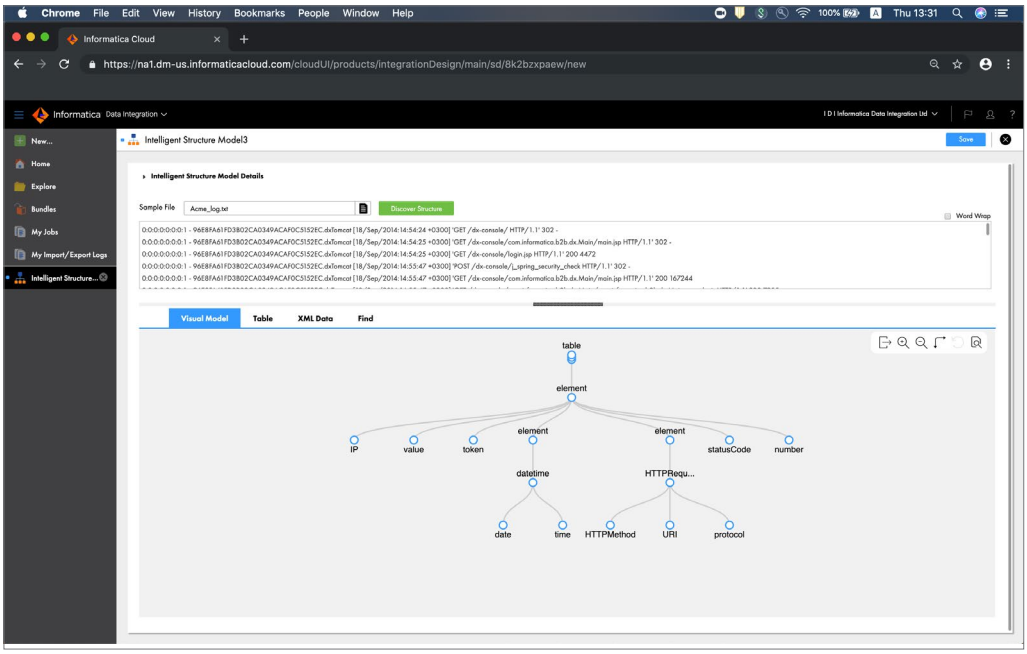

그림 10: 비정형 데이터 파일에서 지능적으로 구조 찾기

마스터 데이터 관리에 사용되는 CLAIRE

고급 AI 및 기계 학습을 사용하는 CLAIRE 기반 자동화 및 인텔리전스는 고객, 제품, 공급업체 및 기타 도메인에 대한 비즈니스 360 뷰의 정확성을 보완하고 개선합니다. 결정론적, 경험적, 확률적 알고리즘에서 상황별 합성 매칭 및 능동적 학습 엔터티 매칭에 이르는 다양한 AI/ML 혼합 기술을 사용해 빠르고 확장 가능하며 매우 정확한 레코드 매칭 및 마스터 데이터 보완 성능을 제공합니다.

합성 매칭

합성은 예컨대 잠재 고객을 고객과 일치시키고, 상호 작용 및 비정형 데이터를 고객과 일치시키며, 명확하지 않은 관계를 검색하는 차세대 매칭 기술입니다. 합성 매칭은 '상황별 속성', 기계 학습, NLP 및 선언적 규칙과 확률적 매칭의 조합을 사용해 관련 항목을 일치시킵니다.

인구 통계학적 속성(예: 이름, 주소, 전화번호), 상호 작용 속성(예: 이메일, 웹 채팅) 및 상황별 속성(예: 언제, 무엇을, 어디서, 누구)은 지정된 신뢰 수준으로 고객 관련 데이터를 일치시키는 데 강력한 효과를 발휘합니다. NLP는 비정형 텍스트에서 '상황별 속성'을 추출하여 매칭 프로세스에서 사용할 데이터 포인트를 더 많이 제공할 수 있습니다. ML 알고리즘은 데이터 스튜어드 및 주제 전문가가 적절하게 선택된 일치 쌍 세트를 일치 또는 불일치로 레이블링하는 지도 학습 방식을 사용하는 경우 매칭 시 큰 효과를 발휘할 수 있습니다. 이와 같이 레이블이 지정된 일치 쌍은 매칭 알고리즘을 생성하는 데 사용되는 교육 세트를 형성합니다.

합성은 인구 통계, 계정, 트랜잭션, 상호 작용 및 비정형 데이터로 구성된 완전한 360도 고객 뷰를 결합합니다. 기존의 매칭 알고리즘은 레코드를 병합하여 고객 싱글뷰를 형성하는 반면, 합성 매칭은 그래프의 모든 고객 데이터를 관리합니다. 데이터는 신뢰 수준과 함께 연결되어 고객에 대한 다양한 뷰 또는 '관점'을 제공할 수 있습니다.

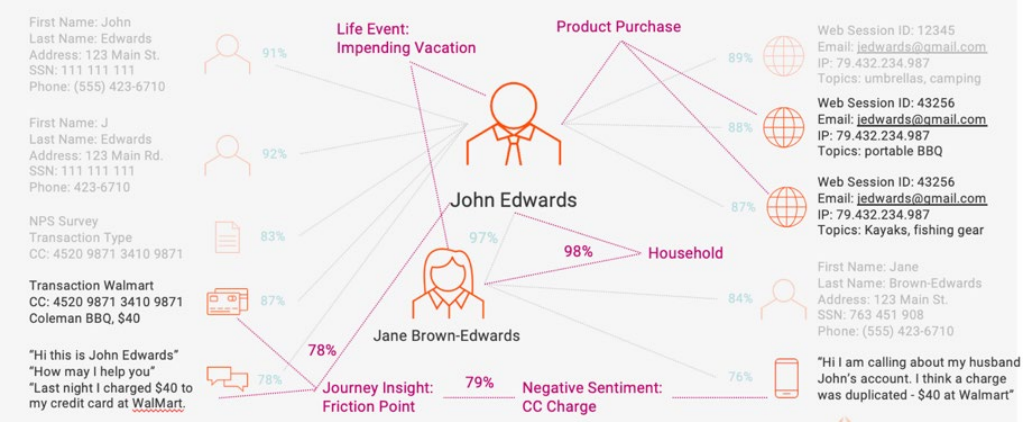

그림 11: 합성 매칭 및 추론은 인텔리전스를 추론한 후 Customer 360의 일부로 저장됩니다.

ID 매칭

CLAIRE의 NAME3 ID 매칭은 인덱싱 및 차단을 위한 스마트 키 생성, 의미론적 텍스트 안정화 및 당사자와 위치 데이터의 비교, 80명의 인구에 대한 목록 및 텍스트 안정화 규칙 편집, 다양한 목적의 기능 중요도에 대해 지능형 가중치 적용 같은 여러 기술을 사용해 30년 이상의 교육/조정 정보를 요약합니다. 이같이 강력한 기술은 여러 필드에 대한 인덱싱 및 차단, 요구 사항에 따른 클라이언트 정의 일치 및 일치 방지 규칙, 구현 정의 일치 및 일치 방지 규칙을 지원하여 다른 AI 규칙을 보완합니다.

엔터티 매칭

엔터티 매칭에서는 동일한 실제 엔터티(예: 고객, 제품 등)를 참조하는 데이터 레코드를 찾습니다. 데이터 레코드는 비정형(예: 웹 채팅에 숨겨진 고객 정보)일 수도 있고, 정형일 수도 있습니다. 일치 분류에서는 일치 쌍을 비교하고, 신뢰 수준과 함께 일치, 경우에 따라 일치 또는 불일치 항목이 있는지 여부를 확인합니다. 사람이 구성한 규칙(예: 선언적 규칙) 또는 AI 규칙(예: 기계 학습 구성)을 사용하는 기술이 있습니다. 이 두 기술을 함께 혼합하면 최상의 매칭 결과를 얻을 수 있습니다.

주제별 전문가가 만든 선언적 규칙은 CLAIRE에서 학습된 무작위 포리스트 분류기 형태로 사용하는 강력한 AI 규칙을 보완합니다. CLAIRE는 일치 쌍 10개 또는 20개의 마이크로 배치가 사용자에게 레이블링용으로 제공되는 AI 교육 프로세스를 가속화하기 위해 클라우드 소싱 또는 다중 사용자 학습이 아닌 지도 능동 학습을 사용합니다(예: 일치, 경우에 따라 일치, 불일치). 레이블이 지정되면 CLAIRE에서 무작위 포리스트 분류기를 다시 학습시키고, 이 반복적인 레이블 지정 프로세스에서 사용자에게 제공할 차선의 일치 쌍을 판별합니다. CLAIRE는 레이블이 지정된 쌍을 사용해 차단 규칙을 추론(즉 명백한 불일치 제거)하고, 차단을 수행하며, 모델을 교육하고, 엔터티 매칭을 수행합니다.

CLAIRE는 Jaccard 같은 문자열 비교/유사점, 데이터 프로파일링에서 파생된 선언적 규칙, 안정화된 데이터세트(인구 파일, 별명, 의미 비교 등) 및 예외를 처리하는 사용자 정의 규칙을 조합하여 사용합니다. 이러한 선언적 규칙은 격차와 예외 문제를 해결할 뿐만 아니라 능동 학습 교육 프로세스를 가속화(즉 학습에 필요한 일치 쌍 개수를 줄임)하고, AI 규칙 기능 구축 작업을 가속화하며, 일치 정확도를 높이는 데 도움이 됩니다. 예를 들어 이름, 생년월일 및 SSN은 비교될 때 규칙에서 이를 일치 상태로 분류합니다. 이처럼 선언적 규칙과 AI 규칙을 혼합하면 교육을 가속화하고 매칭 정확도를 개선할 수 있습니다.

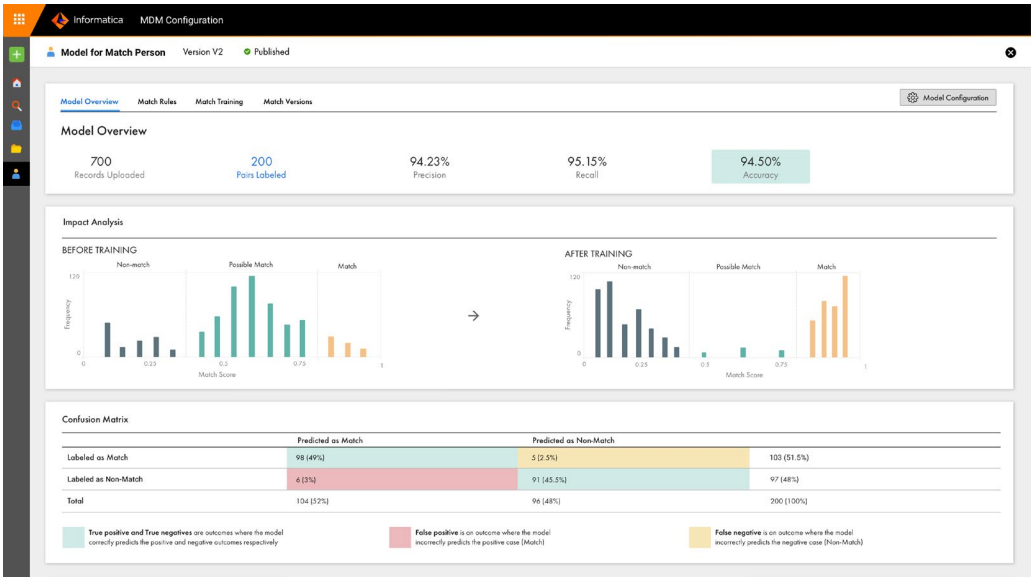

그림 12: 엔터티 매칭

데이터 거버넌스 및 규정 준수에 사용되는 CLAIRE

AI와 기계 학습은 오늘날 가장 까다로운 데이터 거버넌스 작업, 즉 데이터를 찾아 그 품질을 측정하고 이를 관리하는 데 도움이 되는 협업 지원 작업을 자동화하는 데 반드시 필요합니다. CLAIRE는 정책 규칙(예: 데이터 품질)을 자동으로 생성하고, 비즈니스 시맨틱을 기술 메타데이터에 연결하며, 비즈니스 요구 사항에 가장 관련성 높고 신뢰할 수 있는 데이터로 사용자를 안내하는 데 도움을 줍니다.

자동 데이터 품질 보완

CLAIRE는 Stanford NER에 기반을 둔 NLP 접근 방식을 사용해 비정형 텍스트에서 엔터티를 파싱 및 추출합니다. 일반적으로 문자열(예: 제품 코드)에서 엔터티를 추출하기 위해 사용자는 참조 테이블과 정규식을 이용해 파싱 규칙을 작성하게 됩니다. 데이터의 양과 복잡성 및 패턴은 계속해서 증가하고 있습니다. 가능한 모든 입력 사항이 일치하도록 규칙을 작성하는 것은 실용적이지 않고 확장 가능하지 않습니다. 대신 CLAIRE는 사전 교육받은 모델을 사용해 Stanford NER를 기반으로 엔터티를 식별하고 추출합니다.

CLAIRE는 기계 학습을 사용해 들어오는 텍스트를 분류합니다. 예로 언어, 제품 유형 및 기술 지원 문제를 들 수 있습니다. 사용된 기계 학습 방법론을 Naïve-Bayes 및 Max Entropy(다항 로지스틱 회귀)를 사용한 지도 학습이라고 합니다. 지도 학습은 모델을 교육하고 레이블을 할당하는 데 사용됩니다. 그런 다음 교육받은 모델을 데이터 처리 중에 구축하여 서로 다른 입력 클래스에 레이블을 지정하고 라우팅 및 처리할 수있습니다. 예를 들면, 의미가 유사한 '구성' 문제와 별도로 ' 엔진 문제'를 처리하고 의미가 다양한 단어의 용법을 구분할 수 있습니다. CLAIRE는 제품 분류 시 NLP/ML 모델을 활용하고 이미지 메타 태그를 추출하여 이미지 태깅 및 분류 작업을 자동화합니다.

한 글로벌 의료 서비스 대기업에서 기술 자산 21,000개를 비즈니스 용어 6,000개로 매핑하는 정규직 직원이 있었는데, 이 작업을 처리하는 데 2개월이 걸렸습니다. Axon Data Governance 및 Enterprise Data Catalog를 사용하는 CLAIRE는 8분 만에 99%의 정확도로 기술 자산 18,000개의 매핑 작업을 자동화했습니다.

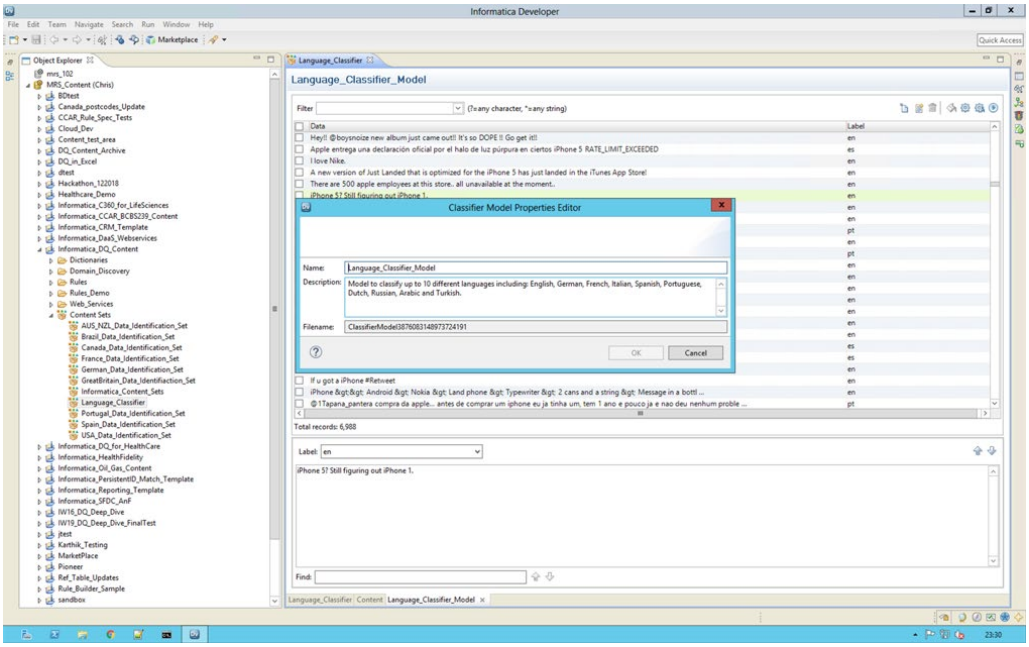

그림 13: 기계 학습 NLP는 텍스트를 분류하고 엔터티를 추출합니다.

비즈니스 용어를 물리적 데이터 세트와 자동으로 연결

데이터 거버넌스에는 비즈니스 아티팩트, 정의, 이해 관계자, 프로세스, 정책 등에 대한 문서가 필요합니다. 정확하게 정렬된 뷰를 사용할 수 있으려면 사용자가 데이터 자산에 기본 기술을 구현하는 데 정의와 비즈니스 뷰를 연결할 수 있어야 합니다. 일반적으로 이 작업은 느리고 힘들며 오류가 발생하기 쉽습니다. 핵심 인력이 커뮤니케이션하여 기술 표현을 하나씩 수동으로 정렬해야 합니다. 이 작업은 완료하는 데 며칠, 몇 주 또는 몇 달이 걸릴 수 있습니다.

Informatica Axon Data Governance는 Informatica Enterprise Data Catalog와 긴밀하게 통합함으로써 이 프로세스를 단축할 수 있습니다. CLAIRE는 메타데이터 스캔이 완료되면 연결되고, 관련성 있는 적절한 데이터 요소에 대한 사용자 추천 정보를 제공합니다. 이렇게 하면 데이터 요소를 검색하고 유효성을 검사하며 연결하는 작업이 줄어 들어 데이터 스튜어드와 데이터 거버넌스 사무소가 중요한 작업에 매진할 수 있습니다. 구현 작업이 진행됨에 따라 프로세스를 완전히 자동화할 수 있습니다.

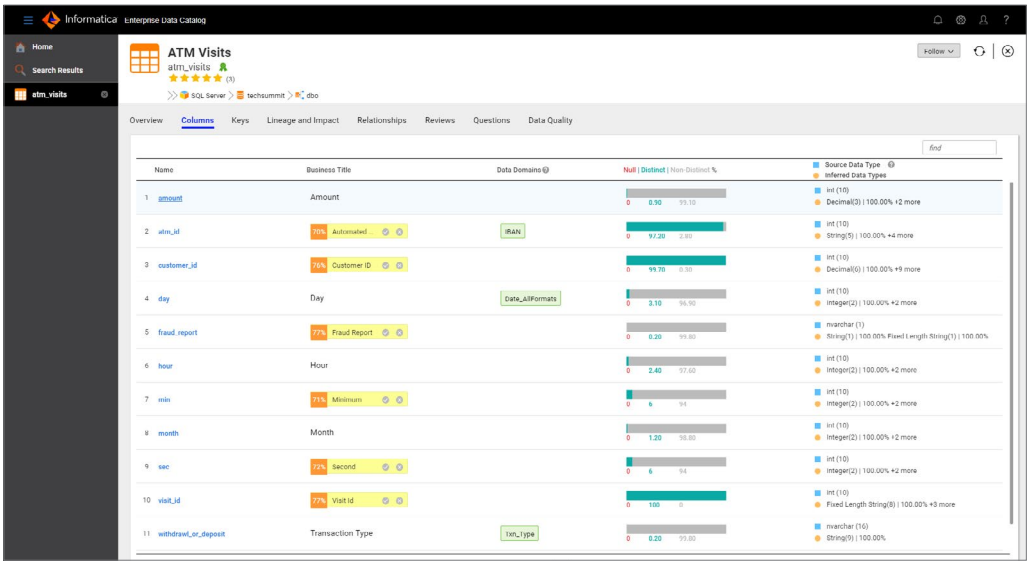

그림 14: 비즈니스 용어를 물리적 데이터세트와 자동으로 연결합니다.

데이터 품질 자동 평가

데이터 거버넌스의 중요 성과 지표(KPI)는 프로세스를 지원하고 정책을 뒷받침하는 등 시스템 전반에 걸친 데이터 품질입니다. 데이터 거버넌스 사무소는 데이터가 완전하고, 정확하고, 일관되고, 유효한지 등을 보장해야 합니다. 요컨대, 비즈니스 운영을 지원할 만큼 신뢰할 수 있고 우수해야 합니다. 데이터 거버넌스 구현 작업이 증가함에 따라 데이터베이스에서 데이터 레이크까지, 데이터 환경 전반에 걸쳐 점점 더 많은 시스템과 필드의 품질을 평가하는 데 소요되는 시간이 점차 증가하고 있습니다.

CLAIRE를 통해 Axon Data Governance는 Informatica Data Quality 및 Informatica Enterprise Data Catalog와 협력함으로써 엔터프라이즈 전체에서 데이터 품질 측정 적용 작업을 자동화하여 수천 시간에 달하는 작업 시간을 절약할 수 있습니다. 데이터 거버넌스 팀은 다양한 데이터 품질 차원에 대한 데이터 품질 규칙을 비즈니스 용어 및 중요한 데이터 요소에 연결합니다. 그러면 기본 시스템에서 다양한 시스템에 필요한 데이터 품질 검사를 생성하고 거버넌스 사무소에 메트릭을 다시 보고합니다.

이 자동화 기능은 다음의 세 가지 주요 정보를 결합하여 지원됩니다.

1. Axon에서 요구하는 중요한 비즈니스 요소 및 데이터 품질 규칙에 대한 지식
2. Informatica Data Quality에서 이식 가능하고 실행 가능한 데이터 품질 규칙과 유연한 실행 엔진
3. Enterprise Data Catalog의 물리적 데이터 자산을 통한 메타데이터 세부 정보

CLAIRE는 이 정보를 결합하여 Enterprise Data Catalog의 물리적 데이터 자산에 대해 Informatica Data Quality에서 데이터 품질 규칙 실행 작업을 생성합니다. 또한 CLAIRE는 Axon의 현업 부서 사용자 컨텍스트를 유지하여 결과가 올바른 대시보드와 집계된 뷰에 표시되어 거버넌스 사무소에서 사용할 수 있도록 합니다.

자동화를 통해 거버넌스 프로그램을 그 어느 때보다 빠르게 확장할 수 있으므로, 데이터 품질 평가를 만들고 거버넌스 컨텍스트에 하나씩 다시 연결하는 일과 관련해 수천 시간에 달하는 수작업이 필요하지 않습니다. 이뿐만 아니라 CLAIRE는 식별된 새로운 물리적 자산이 자동으로 품질 평가를 받도록 합니다. 게다가 데이터 품질 규칙에서 Named Entity Extraction 또는 Classifier를 사용하여 새 도메인을 검색합니다.

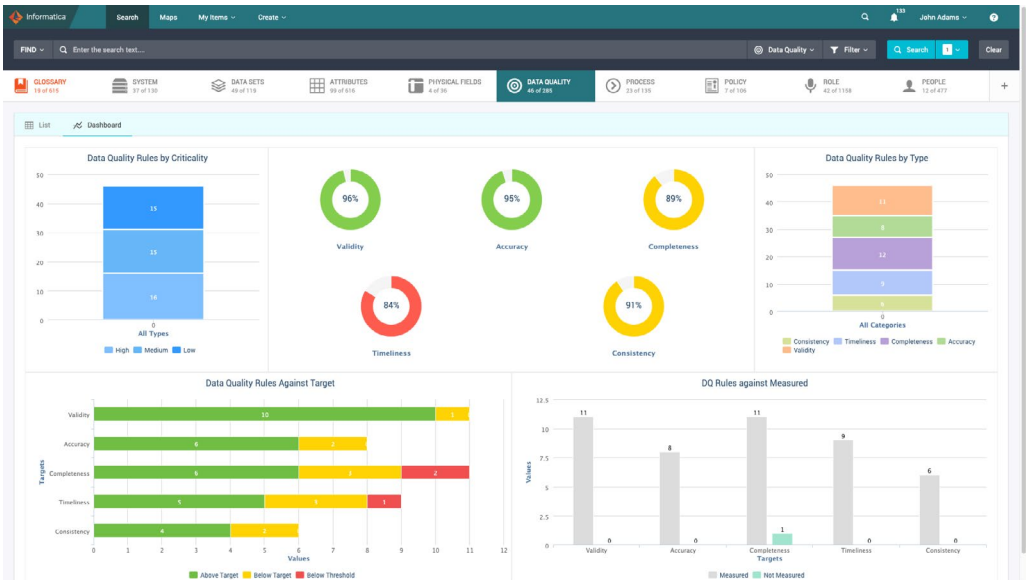

그림 15: 전체 데이터 자산에 대한 데이터 품질 평가를 자동화하면 수천 시간에 달하는 수작업 시간이 절약됩니다.

ML/NLP 지원 데이터 품질 규칙 및 식별

데이터 품질은 데이터 거버넌스 프로그램의 핵심이 되는 필수 사항으로, 대규모 구현 작업에는 많은 데이터 품질 규칙이 있을 수 있습니다. 데이터 스튜어드가 사용할 올바른 규칙을 식별할 수 있도록 CLAIRE는 규칙을 식별할 뿐만 아니라 누락된 규칙을 생성할 수도 있습니다.

Axon Data Governance 사용자는 규칙 요구 사항을 일반 텍스트로 지정(예: '고객 식별자는 8자이고 C로 시작해야 함')하고 CLAIRE를 호출하여 도움을 받을 수 있습니다. ML/NLP 기술을 통해 CLAIRE는 사용자 요구 사항을 분석하고, 이를 기술 표현으로 변환합니다. 이 표현과 관련 메타데이터(예: 글로서리 이름)를 기반으로 CLAIRE는 Informatica Data Quality 규칙을 검색하고 잠재적 후보를 식별합니다. 그러면 사용자가 일치하는 기존 규칙에서 선택하거나, 적용 가능한 규칙이 없는 경우 CLAIRE에 요청하여 새 데이터 품질 규칙을 생성할 수 있습니다.

적용 가능한 규칙을 찾을 수 없는 경우, CLAIRE는 Informatica Data Quality 저장소의 요구 사항을 충족하는 새 데이터 품질 규칙을 자동으로 생성하고 Axon Data Governance 컨텍스트에 다시 연결합니다. 또한 CLAIRE는 Microsoft CDM(Common Data Model) 및 Salesforce 소스를 기반으로 데이터 품질 규칙을 클라우드 프로필에 자동으로 연결합니다. 사용자가 이러한 소스 중 하나에서 핵심 개체에 대한 새 프로필을 만들면 CLAIRE가 측정에 적용해야 하는 베스트 프랙티스 데이터 품질 규칙을 자동으로 제안합니다.

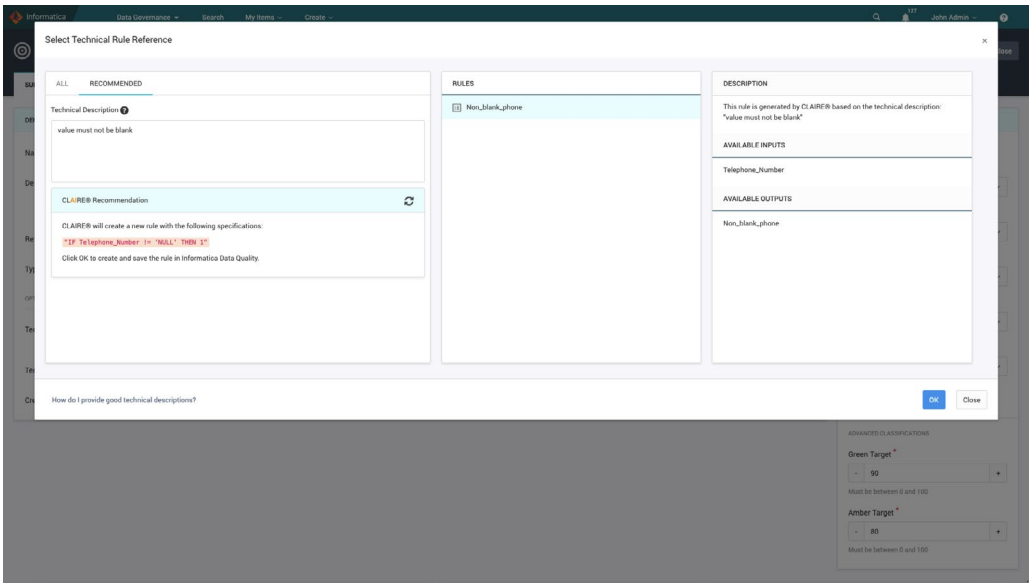

그림 16: NLP를 사용한 자동 데이터 품질 규칙 식별

데이터 프라이버시 및 보호에 사용되는 CLAIRE

CLAIRE가 제공하는 지능형 데이터 프라이버시 솔루션을 통해 기업은 데이터 자산 내에서 개인 식별 정보(PII)를 엔터프라이즈 전반에 걸쳐 확인 및 분석할 수 있습니다. AI 기반 자동화 기능을 사용하면 개인 데이터와 민감한 데이터를 검색하고, 데이터 이동을 이해하며, ID를 연결하고, 위험을 분석할 뿐만 아니라 문제를 해결할 수 있습니다.

주체 레지스트리 ID 매핑

CLAIRE는 프라이버시 규정 준수 및 데이터 주체 액세스 보고용 데이터 매핑을 제공하는 민감한 데이터에 대한 ID 상관관계를 판별합니다. 또한 조합하여 데이터 주체를 식별할 수 있는 데이터를 평가하고 점수를 매깁니다. 정확한 매칭뿐만 아니라, NER(명명된 엔터티 인식) 등의 다양한 고급 기술을 사용해 서로 다른 소스에서 데이터를 결합할 때 일반적으로 발견되는 결과를 개선합니다.

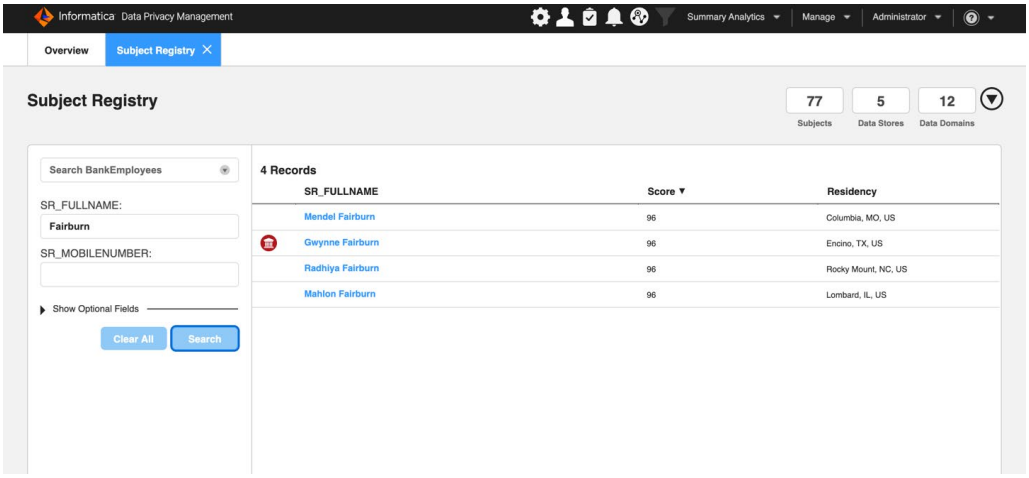

그림 17: 프라이버시 규정 준수 및 데이터 주체 액세스 보고용 주체 레지스트리 ID 매핑

민감한 데이터 매핑 및 이동

CLAIRE는 위에서 언급한 계보(Lineage) 기능을 활용하고 확장함으로써 보안 및 프라이버시 규정 준수 요구 사항을 지원하기 위해 리포지토리 전체에서 민감한 데이터가 어떻게 확산되는지도 확인할 수 있습니다. CLAIRE는 업스트림/다운스트림 이동뿐만 아니라 데이터의 특정 유형, 프로세스, 보호 상태 및 데이터의 위치 등 관련 메타데이터를 판별하여 위반이 발생했는지 여부를 평가할 수 있습니다. 예컨대 개인 데이터가 지리적 경계를 넘어 소스에서 대상으로 이동하거나, 청구 프로세스에 온보딩된 데이터가 현재 프라이버시 규정을 위반할 수 있는 마케팅 프로세스를 진행하기 위해 다른 부서 또는 위치로 확산되는 경우 위반이 발생할 수 있습니다. 그러면 정책 또는 프로세스 이해 관계자가 오류 수정 알림을 받을 수 있습니다.

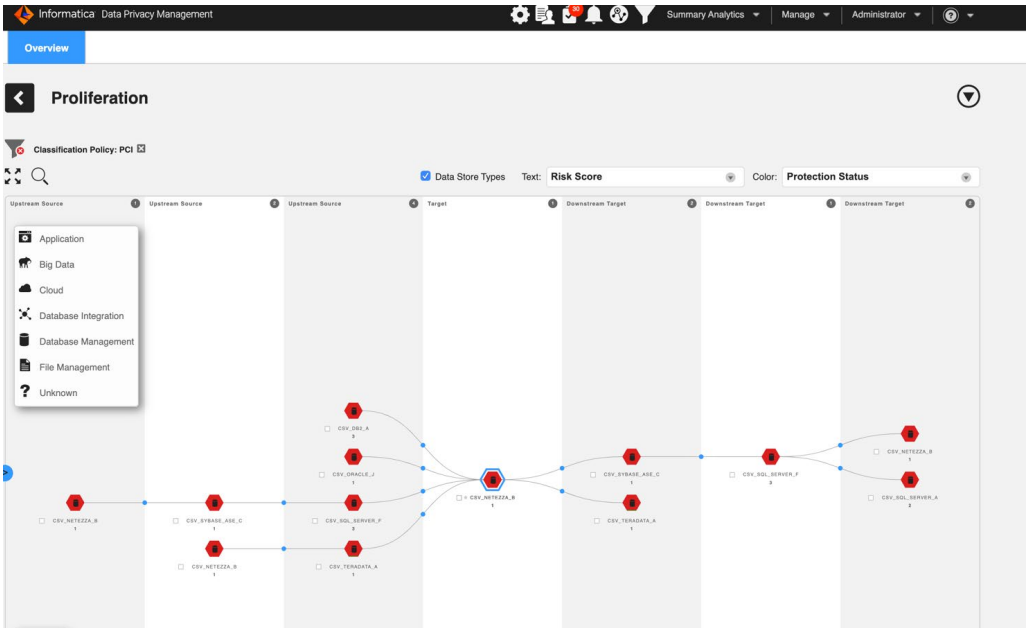

그림 18: 저장소 간의 민감한 데이터가 이동하는 과정을 식별하고 추적합니다.

위험 시뮬레이션 계획

프라이버시 규정에 따라 기업에서 데이터 보호 계획을 마련해야 할 필요성이 점차 증가하고 있습니다. CLAIRE는 기업이 이러한 보호 계획의 영향을 시뮬레이션하여 투자 수익을 높이고 예산 프로세스를 원활하게 수행하는 데 도움을 줄 수 있습니다. CLAIRE는 하나 이상의 데이터 도메인에 적용된 보호 기술을 평가한 후 위험 평가 점수의 변화, 민감한 데이터에 대한 노출, 선택한 각 데이터 저장소의 잔여 위험 비용, 예상 유틸리티 모델을 사용하는 기업에 대해 집계된 영향력을 계산합니다.

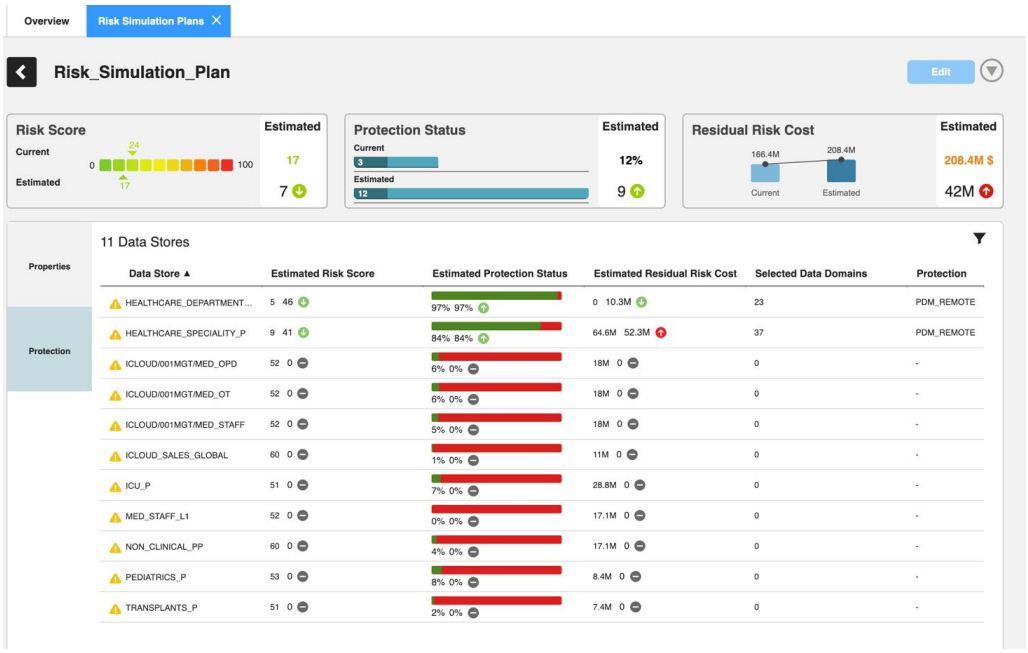

그림 19: CLAIRE는 데이터 도메인에 적용된 보호 기술을 평가하여 위험을 판단합니다.

지능형 이상 감지

CLAIRE는 통계 및 기계 학습 접근 방식을 활용하여 데이터 이상 수치와 이상을 감지합니다. 사용자 행동 분석(UBA) 기능은 위험성이 있거나 조직이 데이터를 오용하게 만들 수 있는 사용자 행동의 패턴을 감지합니다. UBA는 위장, 자격 증명 하이재킹 및 권한 에스컬레이션 공격을 감지할 수 있습니다.

UBA는 감시를 받지 않는 기계 학습을 사용자 활동(사용자가 액세스하는 데이터 저장소의 수, 요청 수 및 다양한 시스템에서 영향을 받는 레코드 수 포함)의 다차원 모델에 적용합니다. 차원수 감소를 위해 기본적인 구성 요소 분석이 이 모델에 적용됩니다. BIRCH 기술은 지정된 기간 동안 행동이 달랐던 사용자를 찾기 위해 감시를 받지 않는 계층형 클러스터링에 적용됩니다. 이상 행동의 유효성을 검사하기 위해 거리 및 밀도 기반의 이상 수치 감지 방법이 채택되었으며, 처음 두 가지 방법이 가리키는 대상이 클러스터 시스템에서 실제 이상 수치인지 확인하기 위해 이상 수치에 대한 통계적인 Grubbs 테스트가 수행되었습니다.

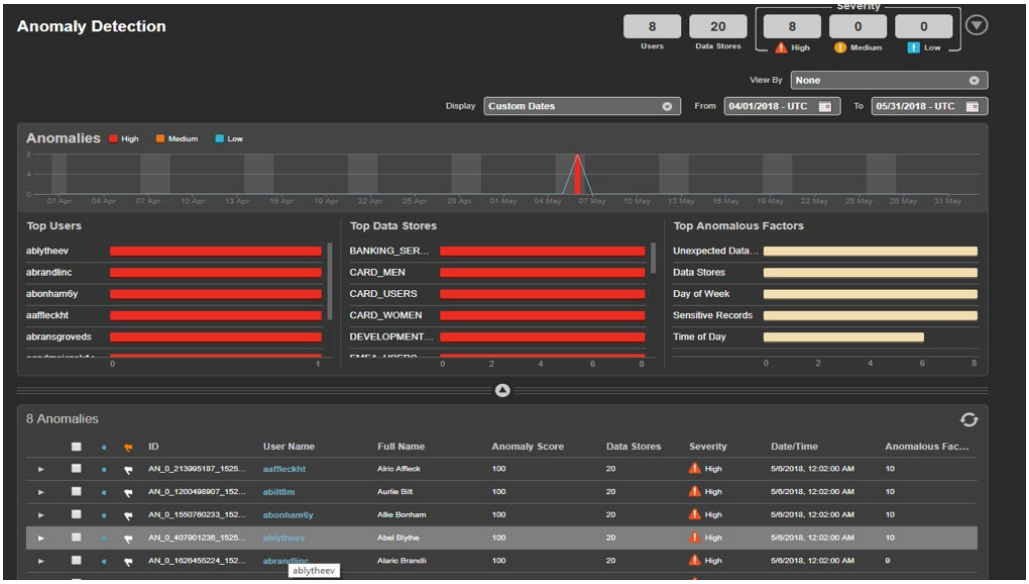

그림 20: 데이터 오용을 나타낼 수 있는 사용자 이상 징후를 자동으로 감지하는 사용자 행동 분석

실시간 API 데이터 보호

API에서 개인 데이터 유출을 확인하고 데이터를 차단 및 마스킹하여 민감한 데이터(예: PII)를 실시간으로 보호합니다. Informatica API Management는 데이터 보호 라이브러리를 통합해 수신/발신 API 호출에서 민감한 데이터를 차단함으로써 민감한 데이터가 노출될 위험을 최소화합니다.

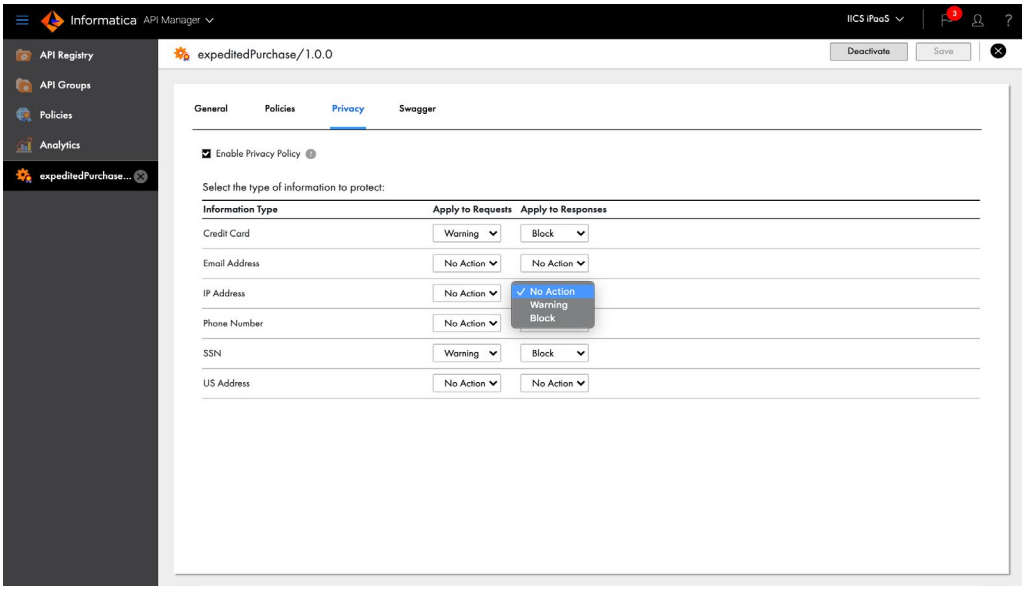

그림 21: 수신/발신 API 호출에서 민감한 데이터에 대한 액세스를 차단합니다.

DataOps에 사용되는 CLAIRE

CLAIRE를 사용하는 기업은 데이터 처리 파이프라인을 가속화하여 DataOps와 관련된 CI(지속적 통합) 및 CD(지속적 제공)를 위한 데이터 관리의 여러 측면을 자동화할 수 있습니다.

데이터 관리 환경을 위한 통찰력 있고 예측 가능한 분석

운영 분석은 기존 프로젝트 및 리소스의 현재 사용량을 파악하고 향후 용량을 계획하는 데 도움이 됩니다. 단일 데이터 관리 플랫폼에서 여러 LOB를 지원하면서 차지백(charge-back) 모델을 구축하기 위한 매개 변수를 제공합니다. 리소스 활용 트렌드에 대한 지속적인 관찰을 기반으로, 용량 계획에 도움이 되는 데이터 볼륨 처리 예측 기능이 제공됩니다. CLAIRE는 데이터 관리 런타임 리소스를 자동 확장하여 다음 단계로 넘어갑니다.

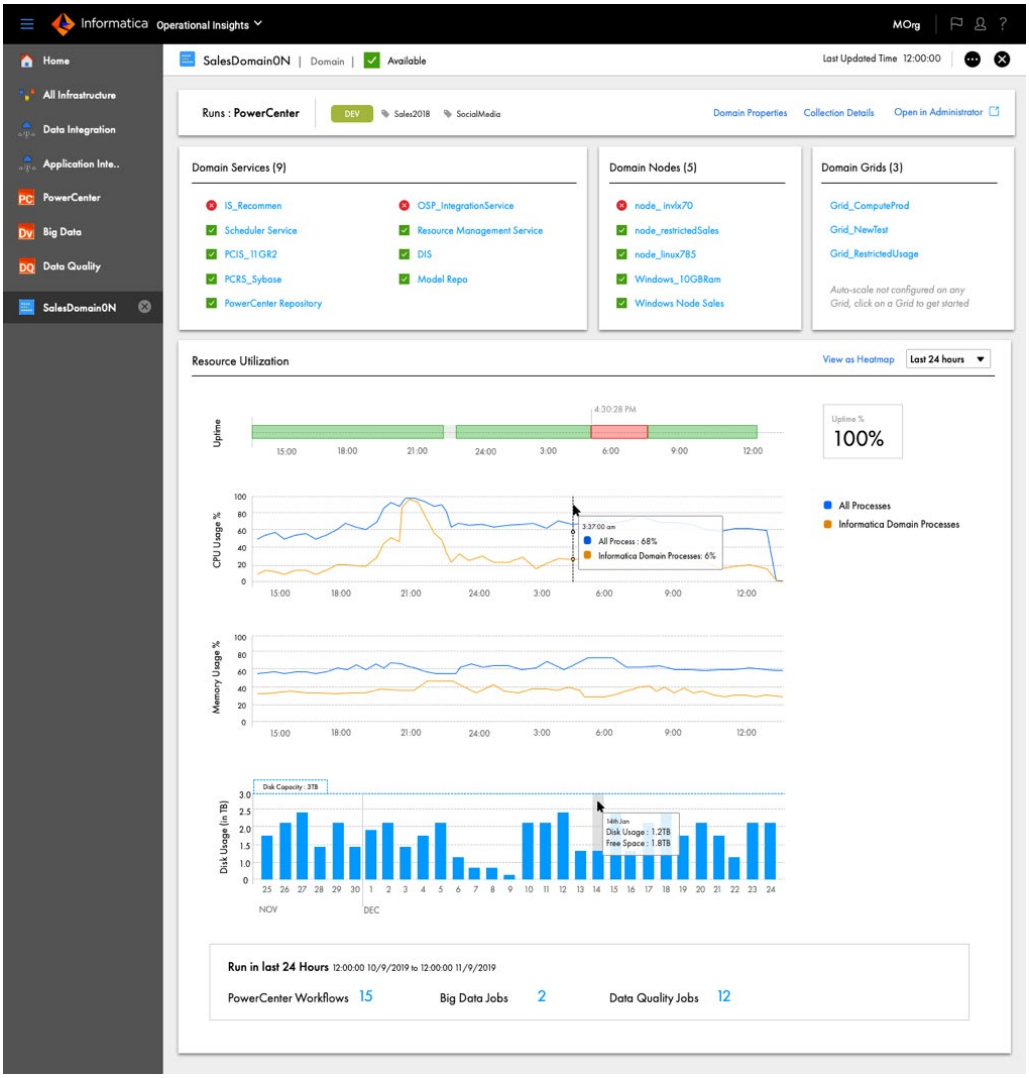

그림 22: Informatica 도메인 프로세스에 대한 운영 통찰력 리소스 활용

작업 실행 시 이상 감지

CLAIRE는 작업 실행 시간, 처리된 데이터, 로드된 데이터, 사용된 리소스, 처리량 등과 관련된 이상 징후를 자동으로 감지합니다. 이러한 이상 징후를 자동으로 감지하면 IT 부서에서 다운스트림 비즈니스 프로세스에 영향을 주기 전에 데이터 통합작업과 관련된 문제를 사전에 해결하는 데 도움이 됩니다. 계절별 하이브리드 ESD 알고리즘은 작업 실행 동작의 이상 징후를 감지하는 데 사용됩니다. 이 알고리즘은 계절성(월말 최대 부하, 휴가철 등)을 고려하여 비즈니스 주기로 인해 예상되는 이상징후가 나타나는 작업을 제거합니다.

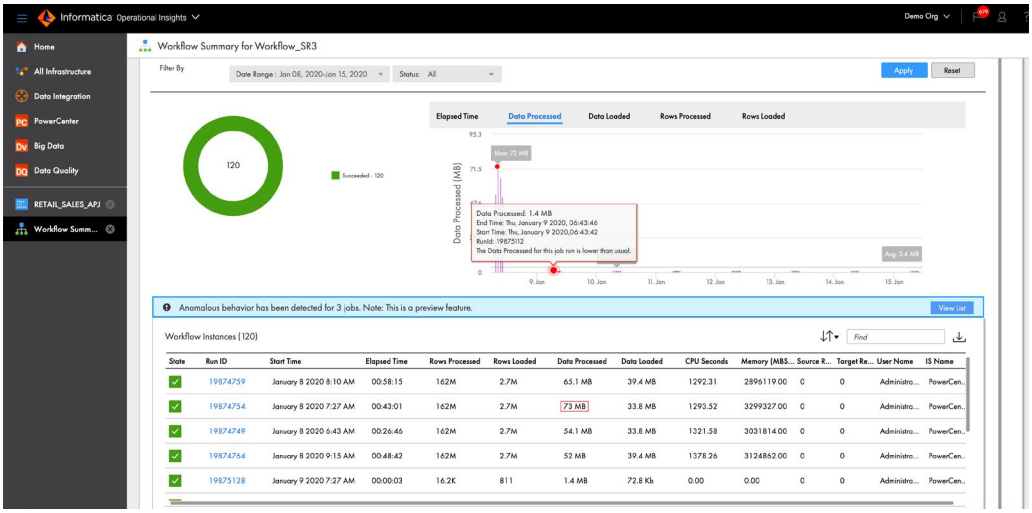

그림 23: CLAIRE는 Informatica 작업 및 데이터 처리와 관련된 이상 징후를 자동으로 감지합니다.

미래의 CLAIRE

CLAIRE가 개발되면서 생산성과 효율성이 계속 증가함에 따라 데이터 리더는 지능형 자동화 기능을 활용해 더 빠르게 우수한 통찰력을 얻고 더 효과적으로 데이터를 관리할 수 있습니다. 제공될 미래의 기능은 다음과 같습니다.

1. **자체 통합**: 새로 도착한 데이터를 데이터 통합 프로세스에 자동으로 통합합니다. 수백만 개의 기존 매핑 및 사용자 작업으로부터 학습하여 데이터를 식별하며, 유사한 데이터를 처리하는 통합 패턴의 위치를 찾고, 데이터를 자동으로 변환 및 이동합니다.
2. **개발 지원**: 다음을 포함하여 개발 프로세스 동안 사용자에게 추천사항을 제시하고 차선의 조치를 제안합니다.
   - 변환 자동 완성
   - 템플릿 추천
   - 민감한 데이터를 위한 마스킹-유형 제안
   - 정제 및 표준화를 위한 데이터 품질 제안
   - 자동 성능 최적화
3. **자동 매핑**: 엔터프라이즈 전반에서 마스터 데이터 엔터티를 감지하고, 필수 변환 및 품질 규칙을 적용하는 마스터 데이터 모델에 자동으로 매핑합니다.
4. **자체 치유**: 메모리 부족 또는 컴퓨팅 전력과 같은 외부 시스템 문제를 적절하게 처리합니다. 예를 들어, 데이터 급증을 처리하기 위해 추가 컴퓨팅 능력('클라우드로 버스팅')을 더합니다.
5. **자체 조정**: 성능 기준을 충족하도록 기록 정보, 현재 데이터 볼륨 및 사용 가능한 시스템 리소스에 따라 일정을 예측하고 조정하거나 리소스를 컴퓨팅합니다.
6. **자체 보호**: 민감한 데이터를 자동으로 감지하고 안전 영역에 보관하기 전에 마스킹합니다.

결론

오늘날의 데이터 중심적인 비즈니스 전략은 데이터라는 토대 위에 수립됩니다. 데이터의 힘을 성공적으로 실현하기 위해서는 데이터 관리 역량을 구축해야 합니다. 일반적인 상황에서 데이터 관리가 야기하는 모든 당면 과제를 기존의 접근 방식인 확장하는 방식으로 처리하면 미래의 요구 사항뿐만 아니라 오늘날의 요구 사항도 충족할 수 없습니다. 혼란을 돌파하기 위해 데이터를 활용하는 한 가지 방법은 플랫폼 사용자의 기술, 운영, 비즈니스 그리고 특히 비즈니스 셀프서비스 측면에서 생산성을 향상시키기 위해 데이터, 메타데이터 및 기계 학습/AI의 힘을 활용하는 엔드 투 엔드 데이터 관리 플랫폼으로 표준화하는 것입니다.

CLAIRE와 Intelligent Data Management Cloud를 사용하여 데이터의 힘을 활용할 수 있는 방법에 대해 자세히 알아보려면 당사에 문의해 주십시오.

Informatica

한국 인포매티카 06611 서울시 서초구 서초동 강남대로 465 교보타워 B동 13층 대표 전화: +82 2 6293 5019

IN09_0521_03328