IBM DataOps

IBM DataStage

IBM Cloud Pak for Data DataStage를 활용하여
AI를 위한 비즈니스에 활용 가능한 데이터 실시간 전달

IBM

데이터 통합을 통해 비즈니스에 활용 가능한 데이터 전달

오늘날 디지털 기업은 이전과는 완전히 다른 데이터를 생성하고 사용합니다. 여기에는 여러 시스템 및 리포지토리에 저장되는 고객, 거래 및 직원에 대한 데이터가 포함됩니다. 이러한 데이터 저장소는 다양한 멀티클라우드, 하이브리드 클라우드 환경 및 데이터 레이크 간에 분산되어 있어 조직에서는 이러한 이질적인 소스 및 환경을 한데 모을 수 있는 방법을 모색하고 있습니다. 그러면 AI를 사용해 인사이트를 더욱 빠르게 얻어 고객에게 차별화 및 맞춤화된 경험을 제공할 수 있습니다. Forrester의 조사에 따르면 데이터 과학자는 AI 이니셔티브를 위한 데이터 준비 및 관리에 시간의 80%를 사용한다고 합니다. 91%의 조직이 데이터를 효율적으로 사용하지 못한다고 답변한 IBM 조사와 연결 지어 보면, 이러한 결과는 기업이 데이터 사일로에서 가치를 이끌어 내는 데 어려움을 겪고 있음을 의미합니다. 대규모 데이터에서 실시간 데이터 접근성을 얻고, 비즈니스에 활용 가능한 데이터를 제공하는 데 사용되는 아키텍처 기술, 사례 및 도구를 데이터 통합이라고 합니다. 유연하고 확장 가능한 데이터 통합 기술을 사용해 기업에서는 최선의 제안을 위한 분석, 고객 이탈 예측 및 분석, 공급망 예측을 수행하고, 여러 데이터 소스에서 데이터 추출, 변환 및 로드(ETL)를 통해 즉각적인 사기 감지를 실행할 수 있습니다.

여러 클라우드 또는 데이터 레이크 간에 데이터를 관리하는 데 어려움을 겪고 있으며 AI 모델 및 애플리케이션을 구축 및 업데이트하는 데 걸리는 시간을 단축하고자 하는 엔터프라이즈 설계자 또는 운영 리더인 CXO를 위해, ETL을 벗어나 비즈니스에 활용 가능한 신뢰할 수 있는 데이터 기능을 제공하는 시장 선도 데이터 통합 솔루션인 IBM® InfoSphere™ DataStage는 확장 가능한 멀티클라우드 데이터 통합 및 전달 솔루션을 제공하여 비즈니스에 활용 가능한 신뢰할 수 있는 정보를 실시간으로 사용할 수 있도록 합니다. DataStage의 주요 기능으로는 자동 워크로드 분산 및 대기 시간이 짧은 병렬 엔진을 사용하여 워크로드를 확장할 수 있는 동시에 설계를 한 번 사용하여 모든 클라우드에서 실행할 수 있는 멀티클라우드 런타임 지원 기능이 있습니다. 이외에도 내장 복제 기능을 사용한 실시간 데이터 전달, CI/CD(지속적 통합 및 지속적 전달) 지원으로 DevOps에 필요한 시간 및 비용 절감, 자율적인 통합 설계를 통해 AI 모델을 구축하는 데 필요한 시간 단축 그리고 인라인 데이터 품질을 사용하여 데이터 문제를 자동으로 감지 및 해결하기 위한 검증 규칙이 있습니다.

DataStage는 지속적인 고품질 데이터를 운용하도록 하는 IBM DataOps 기능의 일부로, AI를 지원하고 모든 데이터 소스에서 적시에 적임자에게 자동화된 셀프 서비스 데이터 파이프라인을 제공합니다. IBM InfoSphere DataStage는 온프레미스, IBM Cloud 및 하이퍼 컨버지드 플랫폼(예: 어디에서나 배포할 수 있는 IBM® Cloud Pak™ for Data)에서 사용할 수 있습니다. IBM® Cloud Pak™ for Data는 비즈니스에 맞춰 확장 가능한 DataStage의 클라우드 네이티브 아키텍처를 제공하는 Red Hat® OpenShift® 기반 AI 플랫폼이자 완전 통합형 데이터입니다. 또한 조직에 데이터 통합, 데이터 복제 및 데이터 가상화를 비롯한 여러 데이터 전달 스타일을 지원하는 플랫폼을 제공하고, 동시에 CDC는 발생 시 로그 기반 변경사항을 파악하고 Kafka 기반 메시지 큐를 사용하여 클라우드 및 데이터 레이크에 있는 대상 데이터베이스에 정보를 제공합니다.

한 번의 설계로 모든 클라우드에서 실행

IDC 연구에 따르면 기업 고객의 90%가 여러 클라우드를 사용하고 있다고 합니다. 멀티클라우드 데이터 통합 기능이 있으면 사용자는 런타임에서 설계를 분리할 수 있습니다. 즉, ETL 작업을 한 번 설계한 다음 모든 클라우드 환경에서 컨테이너를 통해 런타임 구성요소를 배포하여 대용량 데이터 처리로 인한 지연 시간을 줄일 수 있습니다. 온프레미스에서 작업을 생성 및 테스트한 다음 클라우드 환경(예: Microsoft Azure 인스턴스)에서 실행하여 온클라우드 Azure 데이터 레이크를 사용할 수 있습니다. 작업 매개변수 및 그 값이 Kafka 메시지를 경유하여 DataStage의 원격 인스턴스로 전달됩니다.

멀티클라우드 데이터 통합이 제공하는 이점:

- 온프레미스 및 클라우드 환경 간에 데이터를 통합하는 기능
- 설계 프로세스를 간소화하는 자동화된 작업 설계 환경
- 데이터의 외부 이동에 필요한 송신 비용을 최소화하기 위한 원격 작업 실행
- 지정학적 요구사항 이행
- 데이터를 이동할 필요 없이 원래 위치에 그대로 유지하면 되므로 대용량 데이터 세트 처리에 수반되는 대기 시간 단축

자동 워크로드 분산 및 병렬 처리

완전한 클라우드 네이티브 아키텍처를 통해 DataStage용 로컬 컨테이너 또는 공유 컨테이너를 사용하여 워크로드를 동적으로 확장하고 업계 최고의 병렬 엔진(PX)을 사용하여 대용량 데이터 세트에 맞춰 최적화할 수 있습니다. 사용자는 IBM DataStage Flow Designer에서 병렬, 순차 또는 Apache Spark 작업을 생성할지 선택할 수 있습니다.

DataStage Flow Designer 작업은 다음 두 가지 런타임 엔진에서 실행할 수 있습니다.

- 작업 유형이 병렬 또는 순차인 작업은 병렬 엔진에서만 실행할 수 있습니다. 일반적으로 리소스 집약적인 작업은 병렬 엔진에서 실행되기 때문에 병렬 처리를 사용하여 복잡한 작업을 완료하는 데 걸리는 평균 시간은 2분입니다.
- 작업 유형이 Spark인 작업은 Spark 엔진에서만 실행할 수 있습니다.

실시간 데이터 전달

컨테이너로 배포된 실시간 캡처를 위한 변경 데이터 캡처(CDC) 기술이 내장된 DataStage는 최고의 데이터 통합 및 데이터 복제 기능을 제공할 수 있습니다. DataStage는 대용량 데이터 세트가 포함된 복잡한 변환 작업을 허용하고, CDC는 로그 기반 변경사항을 발생 시 캡처하여 복잡한 변환 작업을 사용해 변환한 후 Kafka 기반 메시지 큐를 사용해 클라우드 및 데이터 레이크에 있는 대상 데이터베이스에 전달합니다. 또한 DataStage는 배치 기반 및 대량 데이터 변환 작업을 데이터 웨어하우스에 제공하도록 허용합니다.

CI/CD 지원을 통해 DevOps에 필요한 시간 및 비용 절감

여러 운영 체제에서 여러 컨테이너식 애플리케이션을 관리하는 문제를 해결하기 위해 조직은 Cloud Pak for Data에서 사용할 수 있는 Red Hat OpenShift 등과 같은 강력한 오픈 소스 도구가 필요합니다. Cloud Pak for Data 플랫폼은 이러한 조직에서 컨테이너를 확장 및 프로비저닝해 마이크로서비스 및 클라우드 마이그레이션 전략 등과 같은 주요 IT 이니셔티브를 지원합니다. DataStage 컨테이너는 GitHub 등과 같은 소스 제어 도구를 지원하여 개발에서부터 테스트 및 프로덕션까지 여러 작업에 대한 지속적 통합/지속적 전달(CI/CD)의 생성 및 자동화를 허용해 작업을 빈번하게 게시하고 프로덕션으로 릴리스할 수 있습니다.

AI를 촉진하기 위한 자율적인 통합 설계

자산을 자동으로 검색 및 분류하고, 내장된 맞춤 변환 및 품질 규칙을 기반으로 통합 흐름을 생성하고, 중요한 정보를 검색 및 보호하여 AI를 위한 데이터 수집 및 통합을 대규모로 가속화합니다.

자동화된 작업 설계로 가치 창출 시간 단축

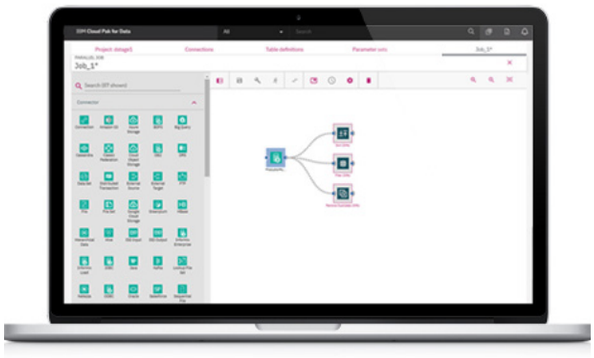

그림 1. 자동화된 설계 기능을 갖춘 DataStage Flow Designer

IBM DataStage Flow Designer는 DataStage용 웹 기반 UI로, 사용자를 지원하기 위해 기계 학습(ML) 기능을 갖추고 있어 기술적인 지식이 없는 사용자라도 업무 범위 내에서 흐름 및 스테이지를 구축할 수 있습니다.

DataStage Flow Designer가 제공하는 이점:

- **이전 버전과의 호환성.** 작업을 마이그레이션할 필요가 없습니다. 다수의 기업에서는 단일 프로젝트에 수천 가지 작업이 포함되어 있으며 기업은 이러한 작업을 통해 연중무휴 24시간 운영됩니다. 이러한 기업에 오류 및 정전이 발생할 가능성과 마찬가지로 마이그레이션은 선택의 여지가 없는 문제입니다. 이러한 기업은 기존 DataStage 작업을 가져와 IBM DataStage Flow Designer에서 제공할 수 있으므로 이러한 작업을 새로운 위치로 마이그레이션할 필요가 없습니다.
- **개발자 생산성 증대.** IBM DataStage Flow Designer에는 내장 검색, 기업에서 바로 시작할 수 있도록 돕는 빠른 둘러보기, 자동 메타데이터 전파, 스마트 팔레트, 제안된 스테이지 및 모든 컴파일 오류의 동시 강조 등과 같은 기능이 있습니다. 개발자는 이러한 기능을 사용하여 작업 설계 시 생산성을 더욱 높일 수 있는데, 기존의 수동 코딩 작업보다 생산성이 9배나 더 빨라집니다.
- **광범위한 연산자 및 연결성.** 설계 및 개발 기능 이외에 DataStage는 사전 작성되어 있고, 바로 사용할 수 있는 연산자 수백 가지를 제공합니다. 따라서 개발자는 분석 작업을 위해 데이터를 준비하는 데 걸리는 시간을 크게 줄일 수 있습니다. 몇 주마다 새로운 연산자가 추가되어 시간이 흐름에 따라 개발자의 생산성이 점점 더 개선됩니다.

신뢰할 수 있는 데이터 전달을 위한 전송 중 데이터 품질 및 보안

DataStage는 권한이 없는 사용자에게 중요한 데이터에 대한 액세스 권한을 제공하는 잠재적인 보안 문제 및 품질 문제를 피할 수 있도록 대상 환경(예: 데이터 레이크)으로 데이터 전달 시 데이터 검증, 표준화 및 일치하는 규칙 실행을 위해 DataStage Flow Designer를 사용하여 데이터 통합을 위한 단일 사용자 경험을 제공합니다. 데이터 품질의 이러한 개념은 데이터 웨어하우스(DWH) 전체에서 포괄적인 데이터 거버넌스를 지원하도록 확장할 수 있습니다.

요약

DataStage가 제공하는 이점:

- 한 번만 설계하면 내장된 자동 워크로드 분산, 병렬 처리 및 확장성을 통해 어디에서나 실행 가능
- 실시간 또는 배치 기반 전달 스타일을 통해 업데이트 파악
- 내장된 복원력, 간단한 작업 및 CI/CD
- AI를 위해 최적화된 데이터 통합
- ML 기능을 사용하여 자동화된 작업 설계
- 신뢰할 수 있는 데이터 전달을 위한 전송 중 데이터 품질 및 데이터 보안

IBM은 하이브리드 멀티클라우드 환경, 온프레미스 또는 하이퍼 컨버지드 시스템(예: IBM Cloud Pak for Data) 또는 선택한 모든 클라우드 플랫폼에서 광범위한 데이터 통합 기능을 제공합니다. 이러한 여러 기능은 유연하고 확장 가능한 데이터 통합 솔루션을 제공하여 선택한 배포 모델에서 AI를 위한 고품질 데이터에 신속하게 액세스할 수 있도록 합니다.

무료로 제공되는 안내식 데모를 보고 IBM InfoSphere DataStage에 대해 자세히 알아보십시오.

왜 IBM인가?
IBM DataOps 기능은 AI 지원 자동화, 주입된 거버넌스 및 강력한 지식 카탈로그와 함께 작동하는 시장 최고의 기술을 제공하여 비즈니스에 활용 가능한 분석 기반을 생성하도록 도와 기업 전반에서 지속적인 고품질 데이터를 운용할 수 있도록 합니다. 데이터 품질을 높여 모든 소스에서 적시에 적임자에게 효율적인 셀프 서비스 데이터 파이프라인을 제공합니다.

DataOps에 대해 자세히 알아보려면 다음 웹페이지를 방문하십시오.
ibm.com/dataops

IBM InfoSphere DataStage에 대해 자세히 알아보려면 다음 웹페이지를 방문하십시오.
ibm.com/products/infosphere-datastage

다음 빅데이터 및 분석 허브를 방문해 보십시오.
ibmbigdatahub.com

IBM Corporation
New Orchard Road, Armonk, NY 10504
Produced in the United States of America
2020년 4월