데이터 수집부터 통찰력 확보까지, 펜타호가 그리는 데이터 통합 분석 로드맵

권동수 / 효성인포메이션시스템 전문위원

정형·비정형·반정형 데이터 등으로 데이터의 종류와 형태가 다양해지고 데이터 양이 급증했지만, 기업에는 빅데이터를 처리할 이렇다 할 플랫폼이 없어 문제가 된 시기가 있었다. '빅데이터'라는 용어가 회자된 2010년대 초반의 얘기다.

몇 년이 지나서야 기업들 사이에서 데이터 수집을 목적으로 하는 빅데이터 인프라 플랫폼 구축이 시작되었다. 데이터는 넘쳐나는데 정작 분석에 활용할 수 있는 데이터는 많지 않았기 때문이다. 그런데 빅데이터 플랫폼을 구축하고 보니 또 다른 문제가 발생했다. 분석에 활용되지 못한 채 방치되는 '다크 데이터(Dark data)'가 많아진 것이다.

이는 빅데이터 인프라 구축의 목적을 '데이터의 품질'이 아닌 '데이터의 수집'에만 두었기 때문에 발생한 문제다. 다크 데이터는 현재도 심각한 문제로 지적되고 있다. IDC는 2025년까지 전 세계의 다크 데이터가 175제타바이트에 달할 것으로 전망했다. 게다가 급증하는 다양한 형태의 데이터 중 비정형 데이터의 약 90%가 분석에 활용되지 않고 있으며, 이들 데이터의 90%가 불과 2년 사이에 생성되었다는 사실도 주목해야 할 점이다.

다크 데이터 양산의 핵심 원인, '데이터 분석 도구의 부재'

이처럼 다크 데이터가 급증하게 된 원인은 어디에서 찾을 수 있을까?

근본적인 원인들 중 첫 번째는 데이터 분석 도구의 부재다. 많은 기업이 정형 데이터를 기준으로 저장, 조회, 분석 등을 할 수 있는 시스템(도구)이 구축되어 있다. 하지만 IoT 센서 데이터, 음성 데이터, 비디오 및 이미지 데이터와 같은 비정형 데이터는 마땅한 분석 툴이 없었다.

데이터의 양이 지나치게 많은 것도 문제다. 불과 몇 년 전 까지만 해도 테라바이트 수준이던 데이터가 페타바이트, 엑사바이트를 넘어 제타바이트 수준으로 급증했다. 그러는 사이 분석에 바로 활용하기 어려운 불완전한 데이터도 증가했다. 불완전한 데이터는 분석에 활용하기 어렵다. 비정형 데이터를 처리할 수 있는 스킬(도구)이 부족하니 분석을 위한 요구사항 자체가 만들어지지 못한 것이다.

물론 빅데이터 플랫폼 구축 이후 수년 동안 기술이 발전되며 데이터 처리와 관련한 문제는 어느 정도 해결되고 있다. 그러나 기업 곳곳에 있는 다크 데이터를 비즈니스 데이터로 전환해야 한다는 문제는 여전히 남아있다.

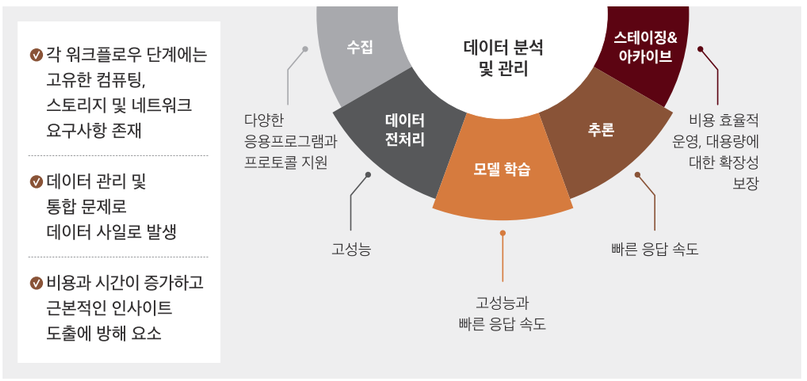
다양하고 복잡해진 데이터 분석 요구

데이터 분석 및 관리 프로세스는 데이터 수집, 데이터 전처리, 모델 학습, 추론, 스테이징[1] 및 아카이빙 등 다섯 가지 과정을 거친다.

정형 데이터만 수집하던 과거와 달리 현재는 데이터를 처리하는 다양한 알고리즘과 애플리케이션을 통해 데이터가 수집되고, 수집된 데이터는 데이터 정제라는 전처리 과정을 거친다. 실제로 프로젝트가 진행되면, 전체 프로젝트의 약 70~80%가 데이터 정제 작업에 할애된다. 상당한 시간과 노력이 소요될 뿐만 아니라 그에 따른 비용도 만만치 않다. 빅데이터 플랫폼 구축을 통해 수집되는 데이터의 양이 많아, 데이터 전처리부터 모델을 생성할 때까지 고성능 CPU, GPU 서버 등 고가의 장비가 필요하기 때문이다.

이 과정을 거친 데이터는 현업 담당자가 사용할 수 있는 단계인 '추론시스템'으로 전달된다. 현업 담당자의 경우 실무 관련 지식은 풍부하지만, 데이터 분석 결과만으로는 직관적으로 내용을 파악하기 어렵기 때문에 이해를 돕는 시각화 도구를 함께 제공해야 한다.

↓ 다양하고 복잡한 데이터 분석 요구사항



1) 스테이징(staging): 대용량 보조 기억 장치에 있는 데이터를 직접적으로 액세스할 수 있는 영역으로 이동시켜 컴퓨터 시스템에서 접근할 수 있도록 하는 기능

그렇다면 추론시스템 구축으로 모든 과정이 완료된 것일까? 그렇지 않다. 데이터 수집부터 추론시스템까지 모든 과정이 주기적으로 반복하기 때문이다. 기존에는 이 주기가 1년 또는 분기, 반기 정도로 길었지만, 지금처럼 데이터가 급증하는 상황에서는 이 주기가 더 빨라져야 한다. 그뿐만 아니라 생성된 모델을 관리할 수 있는 저장소도 필요하고, 마지막으로 아카이빙과 스테이징이 가능한 장비도 연계해야 한다.

이처럼 데이터 분석 과정이 복잡하고, 데이터의 형태가 다양하고 많아지면서 빅데이터를 분석할 수 있는 데이터 과학자를 필요로 하게 되었다. 기존에는 통계 관련 지식이 있는 데이터 분석가가 데이터 엔지니어로부터 데이터를 전달받아 분석한 후, 그 결과를 토대로 현업 의사결정자가 이해할 수 있도록 비즈니스 분석가가 시각화 했다. 그러나 데이터가 급증한 현재는 정보계 시스템, 데이터 레이크, 레이크 하우스 등 여러 저장소에 존재하는 대량의 데이터를 대상으로 분석 작업을 진행해야 한다. 대량의 데이터를 빅데이터 시스템에서 분석할 수 있는 데이터 과학자가 필요해진 이유다.

데이터 과학자들은 상용 솔루션, R, 파이썬 등 자신에게 익숙한 툴을 이용해 분석을 진행한 후 최종적으로 MLOps[2] 환경에서 자동화된 분석 결과를 제시한다. MLOps 환경은 대용량의 다양한 사용자를 연결할 수 있는 컨테이너 기반, 단일 플랫폼 등 현업 환경에 따라 다양하게 구현할 수 있다.

이처럼 다양한 솔루션이 존재하기 때문에 분석한 내용을 전체적으로 아우르고, 자동화와 스케줄링까지 해줄 수 있는 전문 플랫폼이 필요하다. 분석 도구와 빅데이터 등 관련 시장이 지속해서 성장하는 이유이기도 하다. IDC 자료에 따르면 국내 빅데이터와 분석 도구 시장은 향후 5년간 연평균 성장률 6.9%를 기록하고, 2026년까지 3조 2,485억 원 규모에 이를 것으로 전망되고 있다.

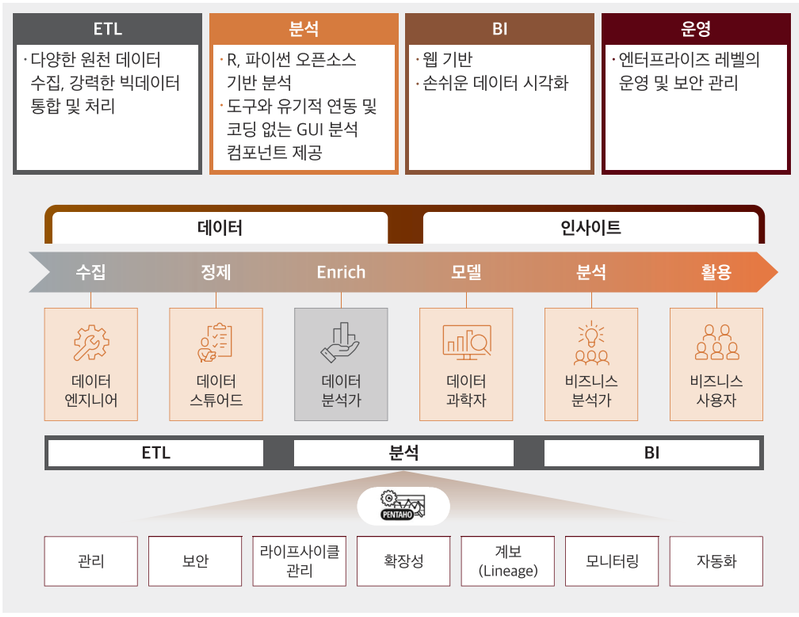
엔드-투-엔드 빅데이터 통합 플랫폼 '루마다 펜타호'

효성인포메이션시스템이 제공하는 루마다 펜타호는 20여 년 전, 오픈소스로 선보인 솔루션이다. 당시만 해도 데이터 통합 솔루션, BI 솔루션 등 여러 가지 솔루션이 하나의 프로젝트에 사용되곤 했다. 이 때문에 통합 관리에 대한 필요성이 제기되었고, 펜타호 솔루션의 등장은 당시로선 획기적인 일이었다. 펜타호를 한마디로 정의하면 'ETL(데이터 추출, 변환, 적재) 및 고급 데이터 분석(R/Python)과 BI 시각화가 가능한 엔드-투-엔드 빅데이터 통합 플랫폼'이라고 할 수 있다.

2) MLOps(ML옵스): ML Dev(머신러닝 모델 개발)과 Ops(머신러닝 운영)의 합성어로, 머신러닝을 운영하는 데 기반이 되는 소프트웨어, 인프라, 배포, 개발 방법론 등의 전반적인 것을 아우르는, 머신러닝 엔지니어링의 핵심 기능

데이터가 수집되어 인사이트로 도출되기까지는 데이터 엔지니어, 데이터 관리자, 데이터 분석가, 데이터 과학자, 비즈니스 분석가의 손을 거치게 된다. 데이터 엔지니어가 데이터를 수집하면 현업 비즈니스에 익숙한 데이터 관리자가 정제 작업을 진행하고, 이 데이터를 이용해 데이터 분석가, 데이터 과학자가 모델링을 수행한 후 비즈니스 분석가의 분석을 거치면 최종 사용자가 이용할 수 있는, 시각화된 최종 결과물이 생성된다.

↓ 펜타호 기반의 데이터 활용 프로세스



이러한 프로세스를 가능하게 해주는 펜타호의 특징은 크게 다섯 가지로 꼽을 수 있다.

01 | 구조화, 분석, 시각화, 예측까지 가능한 통합 플랫폼

과거의 데이터 통합 프로젝트는 '오라클 또는 MSSQL 지원' 여부가 중요했다. 현재는 '퍼블릭 클라우드의 데이터를 가져올 수 있는지, 제조 공장의 IoT 데이터를 가져올 수 있는지' 등이 통합의 핵심이다. 펜타호에는 IoT 데이터를 포함한 빅데이터 계열, 온프레미스, 퍼블릭 및 프라이빗 클라우드 데이터 등 위치에 상관없이 여러 종류의 데이터를 가져와 처리할 수 있는 엔진이 탑재되어 있다. 특히 다양한 프로젝트가 추진될 경우, 순환 로직이 필요한 상황이 많아지는데, 조건식에 의해 DB의 조건식이 바뀌면 그에 맞춰 여러 가지 형태로 변환할 수 있다.

02 | 다양한 메시징 프로토콜 지원으로 실시간 처리

기존엔 실시간 처리를 위해 고가의 솔루션을 구축해야 했다. 그러나 분산 메시징 시스템인 아파치 카프카(Apache kafka)가 등장하면서 시장 판도가 완전히 바뀌었다. 따라서 최근에는 실시간 처리와 관련해 '카프카 연동' 여부가 핵심이 되었다. 펜타호는 IBM MQ, AMOP, 액티브MQ, 카프카 등 다양한 메시징 프로토콜을 지원한다.

03 | 상세 스케줄링 가능, 모니터링도 지원

R 또는 파이썬을 통해 모델을 생성하고 재학습을 진행하는 경우, 스케줄링이 필요하면 해당 솔루션을 별도로 운영해야 한다. 펜타호는 상세 스케줄링 기능도 솔루션 내에서 제공한다. 데이터 통합이 가능한 솔루션이므로 신규 데이터를 가져와 변환하고 재학습 한 후, 그 결과 모델을 바로 추론시스템으로 보내주는 기능을 제공한다. 물론 모니터링도 가능하다.

04 | BI 솔루션으로 간편한 시각화

펜타호는 데이터 마트 생성부터 셀프-대시보드, 보고서 작성 등을 지원한다. 그리고 사용 중인 다양한 써드파티 BI솔루션과 쉽게 연계해 간편하게 시각화를 지원한다는 장점이 있다.

05 | 간편한 보안 관리

보안은 무엇보다 중요한 기능이다. 펜타호는 LDAP나 AD와 같은 기본적인 인증 가입 방식을 모두 제공하며, 자체 인증 관리 기능도 갖고 있기 때문에 보안 관리가 수월하다. 또한 업무 자동화를 통해 간단하게 스케줄링, 모니터링을 할 수 있도록 전체적인 시스템 구성을 지원한다.

↓ 펜타호 엔터프라이즈 에디션의 구성 및 특징

국내 다양한 산업군에서 활약 중인 펜타호

현재 펜타호는 국내에서 게임회사, 금융기관, 공공기관, 제조기업 등 다양한 산업 영역에 구축되어 맹활약하고 있다.

게임업체 A사는 ETL 업무 프로세스의 GUI 기반 자동화, 실시간 대용량 스트리밍 데이터 처리, 데이터 업무 프로세스에 대한 실시간 모니터링을 통한 생산성 증대라는 세 가지 과제를 해결해야 하는 상황이었다. 이를 위해 펜타호를 도입, 빅데이터 통합 기능으로 대용량 스트리밍 데이터를 실시간으로 처리할 수 있게 되었으며, ETL 작업도 코딩 없이 GUI로 구현했다. 현재 A사는 시간당 약 2,000만 건의 데이터를 카프카를 통해 실시간으로 Hive(하둡용 데이터 웨어하우스 시스템)에 저장해 빅데이터 시스템을 ODS(Operating Data Store)[3]로 활용하고 있으며, 실시간 데이터를 활용해 집계 정보 확인 시간도 1시간 이상에서 5분으로 단축했다.

금융보험사인 B사는 ETL 업무 프로세스의 GUI 기반 자동화와 엔터프라이즈급 금융권 보안 감사 요건을 충족시키는 솔루션이 필요했다. 이 회사는 펜타호를 도입함으로써 중앙집중식 리파지터리(repository)를 통한 워크플로우 단일화, 클라이언트 툴 제공을 통한 개발 환경 제공, 대용량 데이터 처리 프로세스의 스케줄링을 통한 자동화를 구현했다.

제조기업인 D사는 공장의 중요 설비에 이상이 발생하거나 갑작스러운 고장 때문에 재료 손실이나 불량 제품 생성과 같은 큰 비용 손실이 발생했다. 분석을 진행하려 해도 각 공장의 IoT 데이터를 수집하기 어려웠기 때문에 분석 모형을 통해 사전에 이상 현상을 예측하고 조치해야 했다. D사는 펜타호 솔루션을 도입한 후 예지정비 모델을 생성하고, IoT 데이터를 실시간으로 수집해 저장, 처리할 수 있게 되었다. 또한 펜타호는 예지정비 모델의 장애 예측 일자도 알람으로 제공하고 있다.

E 공공기관의 경우는 공공 클라우드에 데이터를 수집할 수 있는 빅데이터 플랫폼을 이미 구축한 상태로, 각 기관에서 데이터를 취합해야 했다. 그러나 시스템, 스크립트 등이 개별적으로 개발 운영되고 있어 중앙집중식 관리 저장소가 필요한 상황이었다. E 기관은 펜타호 라이트 버전을 구축해 중앙집중식 저장소를 구현하고, 기관별로 다양한 RDBMS의 데이터와 비정형 이미지 데이터 수집 등에 펜타호를 효율적으로 활용하고 있다.

3) ODS(Operating Data Story): ODS는 EDW로 데이터를 저장하기 전에 임시로 운영계 데이터를 보관하는 장소로, 운영계 시스템의 이력성 데이터를 보관한다.