

PART 02

빅데이터 분석 지형을 바꾸다  
원스톱 플랫폼 ‘펜타호’가 남다른 까닭

SNS, 인터넷, 이메일 등 정형 데이터뿐만 아니라 비정형을 포함한 새로운 구조의 데이터를 처리하는 빅데이터 분석 플랫폼에 대한 요구 조건이 까다로워지고 있다. 단절되거나 산재되어 있는 여러 시스템에서 진행되는 작업으로는 빅데이터를 통해 효율적으로 비즈니스 가치를 창출하기 어렵다는 판단에서다.

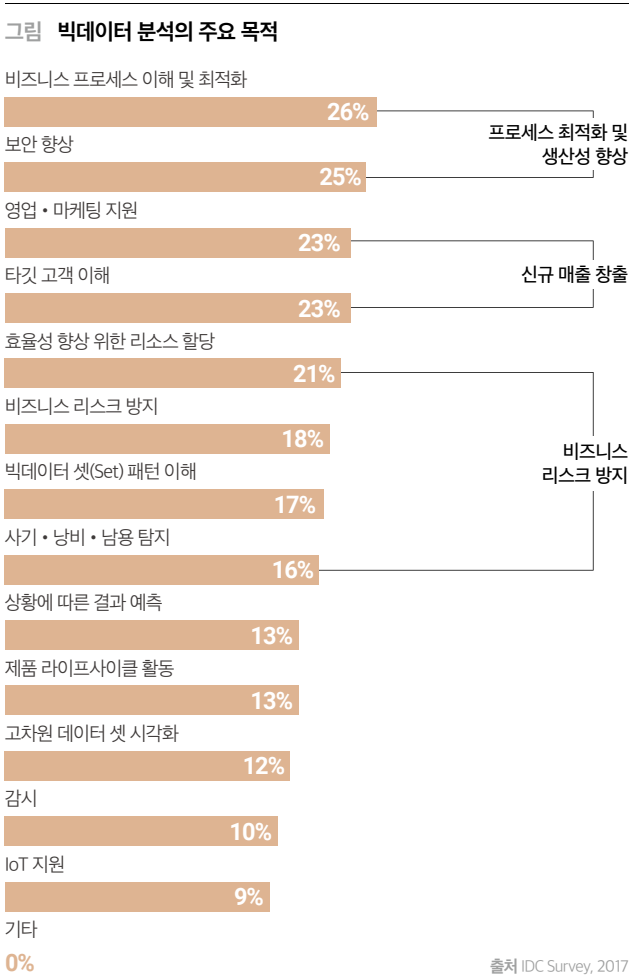
빅데이터에서 가치를 이끌어낼 수 있는, 제대로 된 빅데이터 분석 플랫폼 ‘펜타호(Pentaho)’를 알아보자.

시장조사업체 IDC에 따르면, 전 세계 데이터 규모는 2025년에 163 제타바이트(Zetabytes)에 이를 것이라고 한다. 2016년 데이터 발생량을 기준으로 10배에 달하는 예상치다. 또한 이 데이터 중 4분의 1 이상이 실시간으로 발생하고, 그 중 95% 이상이 IoT(사물인터넷) 데이터가 차지할 것으로 전망되고 있다.

이처럼 데이터는 끊임없이 증가한다. 관계형 데이터부터 비정형 텍스트, 센서 데이터, 머신 데이터까지 그 종류도 무척이나 다양하다. 기업은 더 나은 성과를 얻고 새로운 비즈니스 성과를 내기 위해 이러한 데이터를 충분히 이해하고 활용할 수 있어야 한다. 디지털 트랜스포메이션을 추진하는 이유도 이 때문이다.

디지털 트랜스포메이션의 핵심은 데이터를 통해 기업을 혁신시키고 비즈니스 성과를 높여 매출 확대를 이루는 일이다. 이에 데이터의 소스, 형태, 규모 등의 조건에 제약 받지 않고, 기업이 보유하고 있는 모든 데이터를 통합하여 분석하며, 이를 비즈니스 확장을 위한 인사이트로 활용할 수 있다면 좋을 것이다. 하지만, 데이터의 수집, 정제, 분석 및 활용까지 단계별로 각각의 솔루션을 도입하기엔 기업의 입장에서 IT 투자비용에 대한 부담감이 클 수밖에 없다.

데이터의 수집, 정제, 분석 및 활용까지 한 번에 해결할 수 있는 펜타호라면 이들 요건을 충족시킬 수 있다.



펜타호는 전 세계 1,800여 고객을 확보한 빅데이터 분석 플랫폼으로, 정형 및 비정형 데이터에 대해 수집부터 변환, 적재, 분석, 시각화, 그리고 머신러닝까지 일련의 과정을 모두 제공하는 원스톱 플랫폼이다.

다양한 소스의 데이터에 쉽게 접근해 읽어오고, 이를 즉각적으로 가공할 수 있다. 더 나아가 이를 바탕으로 최신의 머신러닝 및 인공지능(AI) 알고리즘을 접목시킴으로써, 최선의 결과를 예측하고 시뮬레이션이 가능하다. 물론 이 모든 작업은 하나의 워크플로우로 통합되어, 조직의 일부가 아닌 전체가 효과를 얻고 추가적인 수익을 창출할 수 있다.

펜타호는 다양한 데이터 소스들을 추출, 변환, 적재 할 수 있는 Pentaho Data Integration(DI)과 고급 분석 및 시각화 기능을 제공하는 Pentaho Business Analytics(BA)로 구성되어 있다. 일반적인 빅데이터 솔루션과는 달리 오픈 소스를 기반으로 하여, 비용 효율적이며 우수한 커스터마이징을 제공한다. 펜타호는 아래와 같이 네 가지 특징이 있다.

GUI 환경 기반의 직관적이고 편리한 인터페이스

강력한 빅데이터 통합 및 처리

데이터 블렌딩과 머신러닝의 유기적 결합

다양하고 신속한 리포팅 및 시각화

그림 펜타호 구현 원리



## Q&A

**시중에 펜타호와 비교되는 플랫폼이 있는가.**

펜타호는 데이터 블렌딩<sup>2)</sup>부터 시각화까지 전 과정을 제공하는 업계 유일의 엔드-투-엔드 빅데이터 분석 플랫폼이다.

**배치 분석이 아닌 실시간 분석에도 적용할 수 있나.**

당연하다. 최근의 트렌드는 배치 데이터 처리 및 분석뿐 아니라 실시간으로 데이터를 수집하고 분석한 결과를 통해 인사이트를 확보하는 것이다. 펜타호는 실시간 데이터 분석을 위한 다양한 기능을 제공한다.

**기존에 다양한 데이터 취합 및 저장 시스템이 구축되어 있는 상황에서 펜타호를 적용하고자 할 때 어려움은 없나.**  
기존의 데이터 취합 프로세스를 유지하면서 펜타호를 도입할 수도 있고, 혹은 펜타호로 이를 대체할 수 있다.

**여러 종류의 DB를 통합해 분석할 수 있나.**

**DB 종류에 따라서 추가 비용은 없는지.**

펜타호는 시장에 소개된 거의 모든 데이터 소스 및 DBMS를 지원한다. 다양한 DB에서 추출된 데이터를 통합해 분석이 가능한데, DB 종류의 추가에 따른 비용은 별도로 없다.

**미세먼지 실시간 측정과 같이 실시간으로 제공되는 데이터의 경우, 완전히 가공되지 않은 파일은 어느 정도까지 지원되나.**

가공되지 않은 파일에 대해 파싱 및 정제한 후, 분석 가능한 데이터로 처리하는 과정을 거친다.

**중소기업에서 펜타호를 도입하기엔 비용 부담이 크지 않나.**

펜타호는 오픈소스인 커뮤니티 에디션과 상용 엔터프라이즈 에디션, 두 가지가 있다. 중소기업에서는 주로 커뮤니티 에디션을 사용해 업무에 활용할 수 있을 것이다. 다만 커뮤니티 에디션을 사용할 경우 별도의 기술지원을 받기는 어려우며, 보통 자체 인력으로 커스터마이징 및 관리를 해야 한다.

**펜타호를 데모로 사용해 볼 수 있는 방법은.**

무료 체험이 가능하다. 현재 히타치 벤틀라 홈페이지에서 30일 무료 체험 버전을 다운로드할 수 있으며, 링크는 [www.hitachivantara.com/go/pentaho.html](http://www.hitachivantara.com/go/pentaho.html)이다.

2) 데이터 블렌딩(Data Blending) : 데이터 블렌딩은 여러 소스의 데이터를 결합해 분석 데이터 셋을 만드는 프로세스를 말한다.

## Case Study

사례 1 제조업

### HYOSUNG

효성그룹의 베트남 소재 공장에서는 ‘펜타호’를 통해 제품의 생산 공정에서 발생하는 문제점의 원인을 분석하고 효율적인 운전 조건을 식별해 작업 환경을 개선해냈다.

또한 공정의 각종 센서와 실측 관리 데이터를 수집 및 통합, 분석함으로써, 데이터 간의 상관성을 분석해내어, 제품 품질을 유지할 수 있는 체계를 마련할 수 있었다.

사례 2 금융업



미국 금융감독기관인 FINRA(Financial INdustry Regulatory Authority)는 사기나 내부자 거래 및 규정 준수 문제를 탐지하기 위해 펜타호를 이용하고 있다.

애널리스트들은 IT 부서의 지원 없이 검색 알고리즘을 직접 작성해 불법 행위의 검색 속도를 10~100배 향상시켰으며, 매일 최대 5PB의 트랜잭션 데이터를 생성하며 약 4,250개의 중개 회사를 효율적으로 감독하고 있다. 실제 FINRA는 2015년 월스트리트 브로커들의 9,620만 달러 상당의 불법 행위를 발견, 시정 명령을 내렸다. 이는 2014년 대비 3배에 가까운 금액이다.

3) KPI(Key Performance Indicator) : 핵심성과지표. 목표를 성공적으로 달성하기 위해 핵심적으로 관리해야 하는 요소들에 대한 성과지표  
4) UBI(Usage-Based Insurance) : 운전자의 운전 습관을 바탕으로 보험료를 결정하는 자동차 보험 상품

사례 3 보험업



IMS는 펜타호를 통해 실시간 모니터링, 시계열 KPI(Key Performance Indicator)<sup>3)</sup>, 사기 방지 및 예측을 수행하고 있다. 폭증하는 데이터에도 불구하고 펜타호 도입 후 2년 동안 약 5조의 데이터 포인트를 분석해낼 수 있었다. IMS 고객은 자신의 차에 센서를 부착, 펜타호 기반 분석에 의거 자동차 보험료를 절감하는 것은 물론 사고 예방 효과까지 얻게 되었다.

IMS의 프로젝트 개발 및 관리 부문 시니어 디렉터인 크리스토퍼 델은 “펜타호의 데이터 통합 및 분석 기능을 DriveSync 플랫폼에 임베디드함으로써 UBI(Usage-based insurance)<sup>4)</sup> 솔루션을 보완할 수 있었다.”고 전한다.

사례 4 산업기계 및 중공업



Caterpillar Marine Asset은 8척의 선박을 소유한 고객의 예인선 엔진 고장을 차단하여 수명을 연장시킴으로써 매년 낭비되는 연료 비용 200만 달러를 절감하고 있다.

자체 대시보드 기능 개발에 필요한 상당한 시간과 비용을 절감하는 것은 물론, ETL 기능으로 데이터 품질 및 무결성 보장을 위한 보안 및 업무 감독 개선 효과를 얻을 수 있었다. 자동화된 데이터 통합 프로세스를 도입하고 머신러닝 기반 예측정비(Predictive Maintenance)인프라를 구축했다는 점도 커다란 성과다.