새로운 빅데이터 소스의 제어 및 활용

단순한 **Hadoop** 이상**!**

세계가 상호 연결되면서
데이터의 양이 폭발적으로 증가

빅데이터는 단순한 **Hadoop** 이상

빅데이터란 무엇입니까**?**

Hadoop에 대한 모든 것을 알고 싶습니다.

Service Oriented Finance CMO

빅데이터는 단순한 **Hadoop** 이상입니다**!**

경쟁업체는 이 점을 이해하지 못하고 있으며, 빅데이터 유스 케이스의 전체 요소에 대한 가치를 전달하지 못하고 있습니다**.**

IBM

IBM 빅데이터 플랫폼은 빅데이터 과제를 해결할 수 있는 완전한 에코시스템

두 가지 주요 빅데이터 유형

움직이는 데이터(Data in motion)

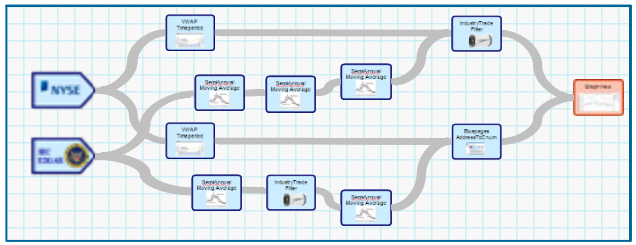
- 일반적으로 데이터가 저장되지 않음
- 매우 빠른 속도
- 여러 데이터 소스
- 엄청난 규모의 비정형 데이터
- 매우 빠른 처리 시간을 필요로 함

움직이지 않는 데이터(Data at rest)

- 데이터가 디스크에 저장됨
- 엄청난 규모의 비정형 데이터
- 사전 정의된 스키마가 없음
- 규모가 너무 커서 기존 도구로는 제시간에 처리할 수 없음

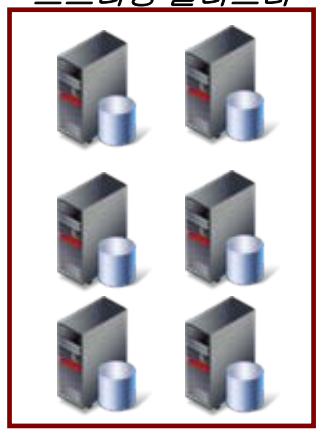
새로운 프로그래밍 모델과 저렴한 비용의 하드웨어가 빅데이터 문제를 해결

스트리밍 애플리케이션

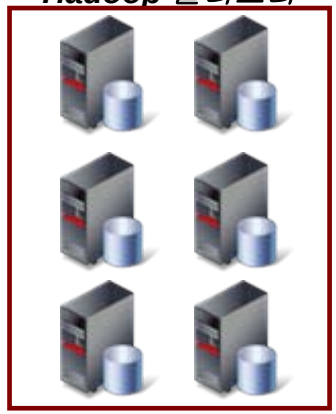


- 스트리밍 데이터 및 Apache Hadoop 애플리케이션
  - 대량 데이터 처리에 검증된 프레임워크
  - 움직이는 데이터에는 스트리밍, 움직이지 않는 데이터에는 Hadoop
  - 어플리케이션에게는 투명하게 대규모 노드 클러스터에서 병렬로 작업이 수행됨

스트리밍 클러스터



***Hadoop** 클러스터*



비용이 저렴한 System x 서버 클러스터는 Hadoop 및 스트리밍 애플리케이션에 이상적임

움직이는 데이터에서 가치 얻기

데이터 소스	분석	비즈니스 가치
의료 장비 	*다양한 의료 기기를 실시간으로 모니터링하여 추세 및 이상 식별*	*생명에 위협이 되는 상황을 미리 감지하여 개입*
주식 거래 	*수신 데이터에 대해 매우 빠르게, 대기 시간이 거의 없이 복잡한 계산 수행*	*정확하고 시기 적절한 정보를 시장 관리자에게 제공*
POS 데이터 	*POS(Point of Sale) 데이터와 고객의 관계 데이터를 실시간으로 결합*	*현재 구매 상황에서 제품에 대한 상향판매 기회 최대화*

빅데이터를 통해 경쟁 우위를 확보하려 하는 **Service Oriented Finance**

우리의 마켓 관리자는 이 애플리케이션으로 실질적인 혜택을 누릴 수 있습니다.

Service Oriented Finance 시장 관리자

Service Oriented Finance는 다음 요구사항을 충족하는 주식 거래 애플리케이션을 배포하고자 함

- 초당 수백만 건의 거래 처리
  - 애플리케이션이 확장 가능해야 함
- 일관된 입력 데이터 플로우
- 마이크로초 단위의 지연 시간
- 비정형 거래 데이터 입력
- 정교한 분석 논리

InfoSphere Streams는 움직이는 빅데이터를 위한 실시간 분석 플랫폼

시기 적절한 의사결정

강력한
분석

초당 수백만
건의 이벤트

마이크로초
대기 시간

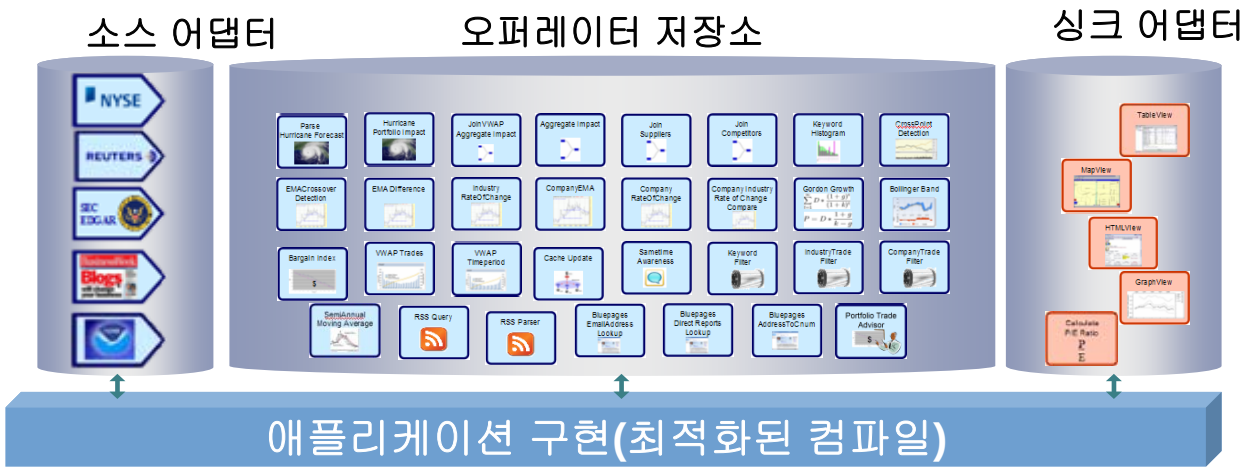
센서, 비디오, 오디오, 텍스트 및
관계 데이터 소스

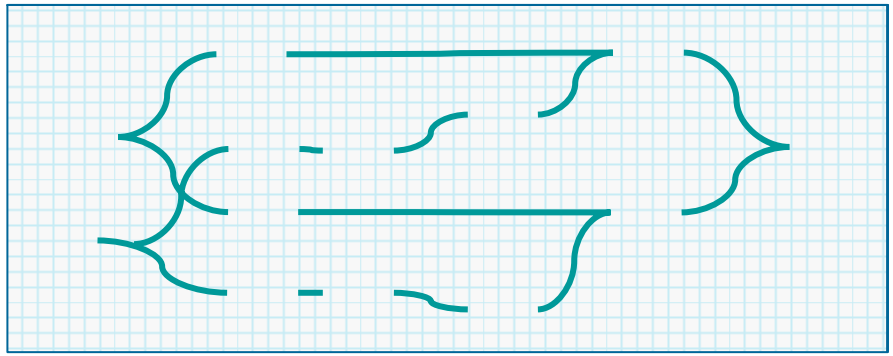
InfoSphere Streams는 이러한 요구사항을 충족시킬 수 있습니다.

Streams는 빅데이터를 위한 실시간 분석 플랫폼입니다.

경쟁업체에는 이러한 기능이 없습니다.

Drag and Drop 방식을 통한 간편한 Streams 프로그래밍

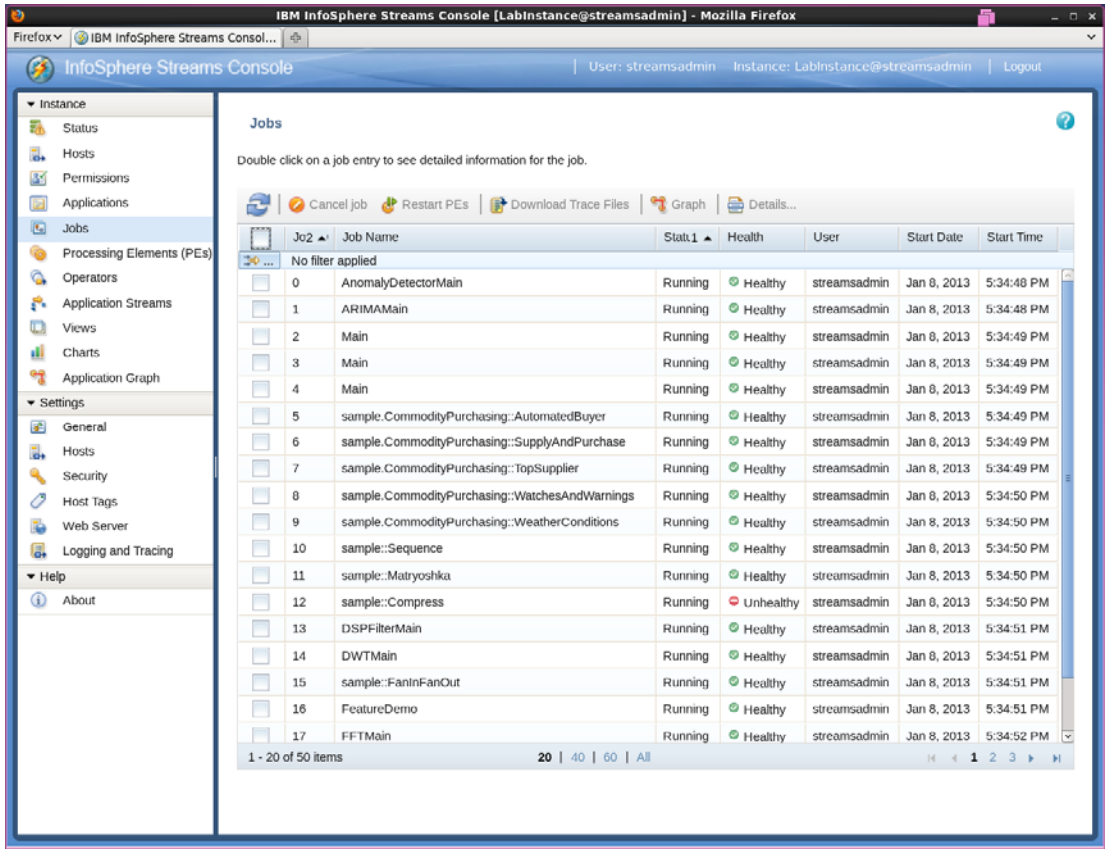




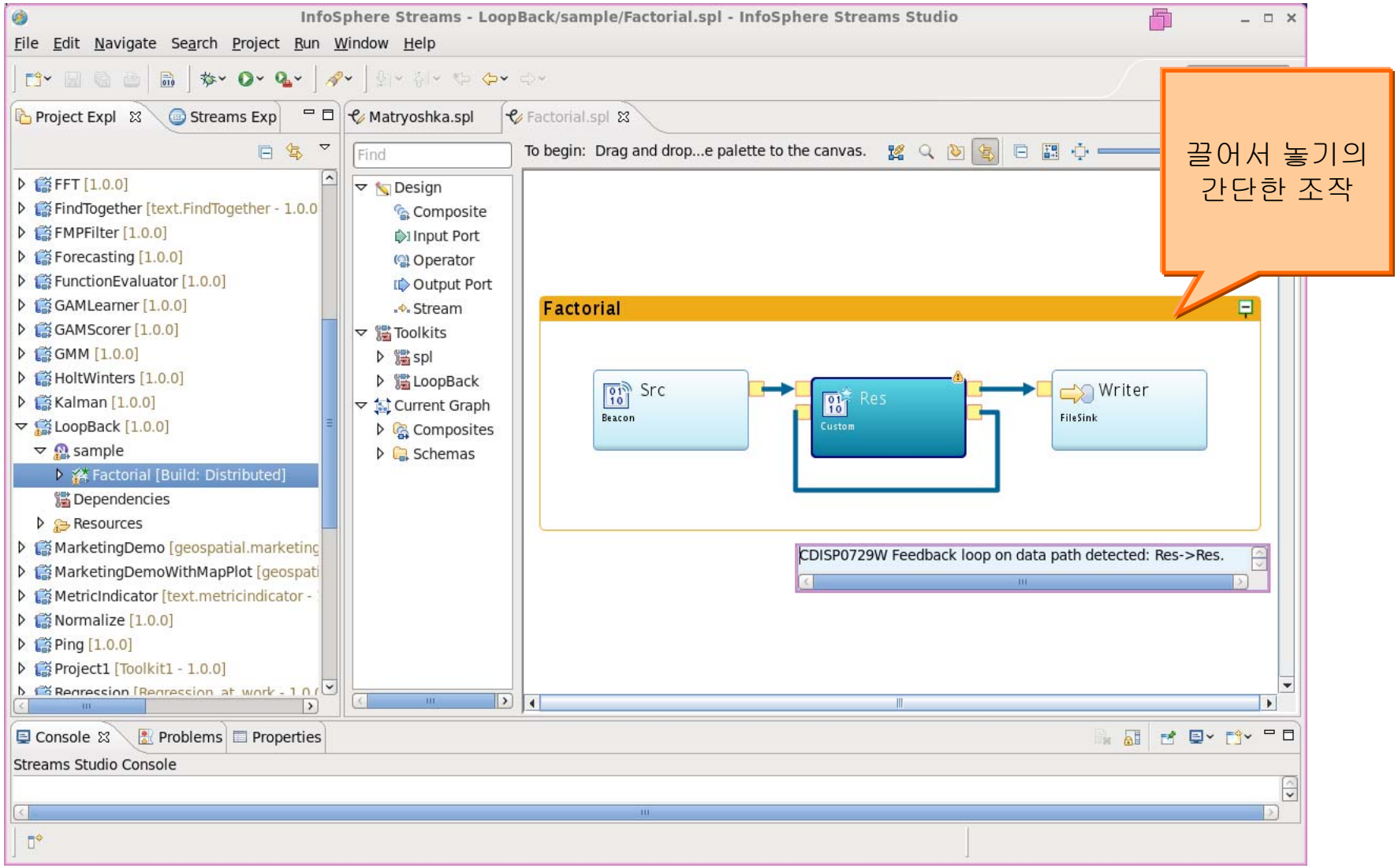
움직이는 데이터에 대한 프로그래밍을 손쉽게 수행

- 개발자 역할
  - Eclipse 기반 도구
  - 비주얼 어플리케이션 모니터링
  - 내장 액셀러레이터
- 관리자 역할
  - 시각화된 어플리케이션 관리
  - 스트림 데이터 시각화
  - 작업 시작/중지
- 비즈니스 사용자 역할
  - 어플리케이션 모니터링 시각화
  - 스트림 데이터 시각화

InfoSphere Streams Console

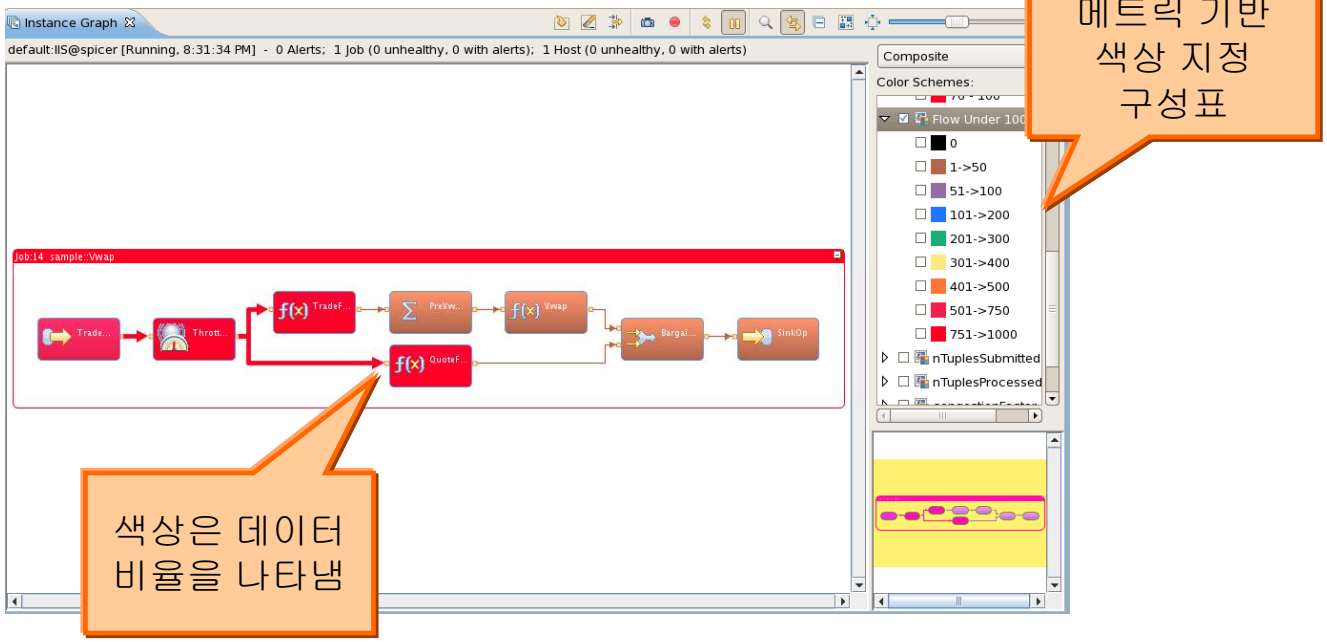


다양한 **Eclipse** 기반 도구 세트를 제공하는 **Streams Studio**

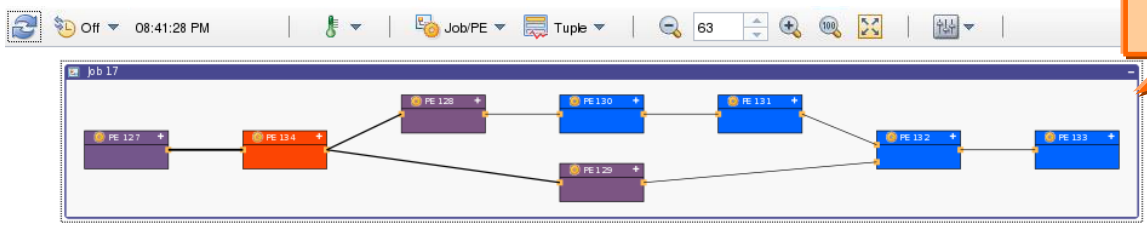


비주얼 애플리케이션 모니터링은 실행 중인 애플리케이션에 대한 명확한 뷰를 제공

Streams Studio의 개발 시간 모니터링



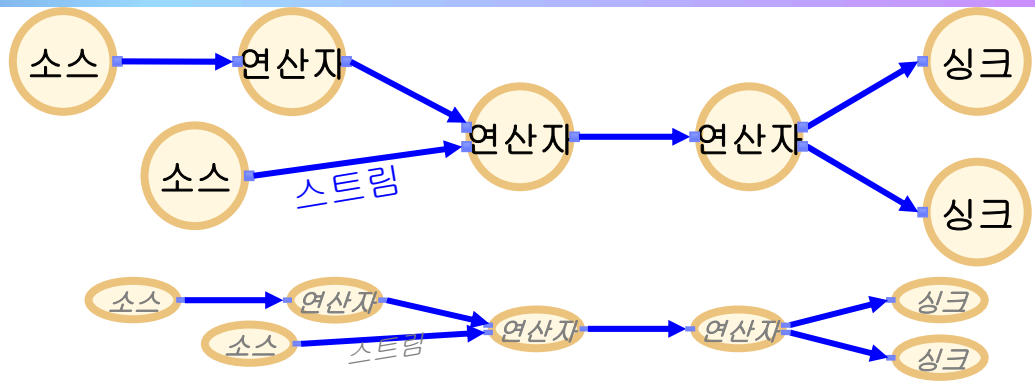
Streams Console의 프로덕션 모니터링



시각적으로 애플리케이션 상태 모니터링

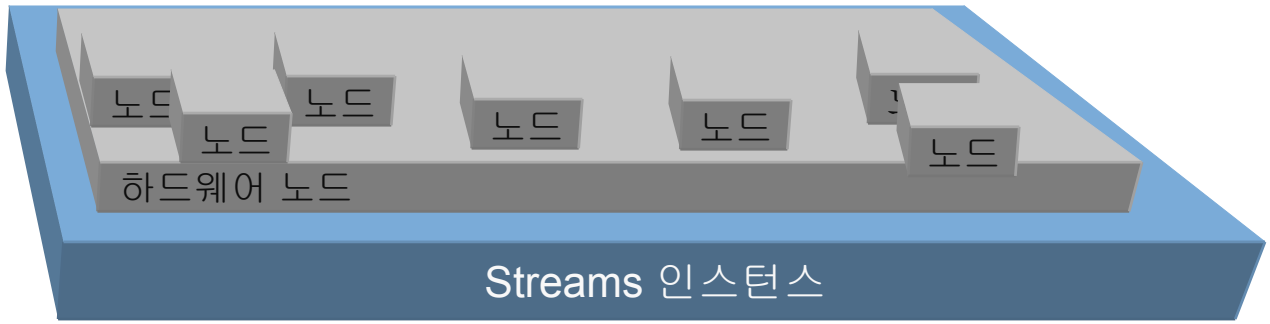
단일 노드 또는 노드 클러스터에 **Streams** 작업 배치

- Streams 작업
  - 연산자 집합
  - 스트림을 통해 연결됨



- Streams 인스턴스(또는 간단히 인스턴스)로 알려진 Streams 런타임 환경에 작업 배치
- 인스턴스는 단일 처리 노드를 포함(하드웨어)

  또는 여러 처리 노드를 포함

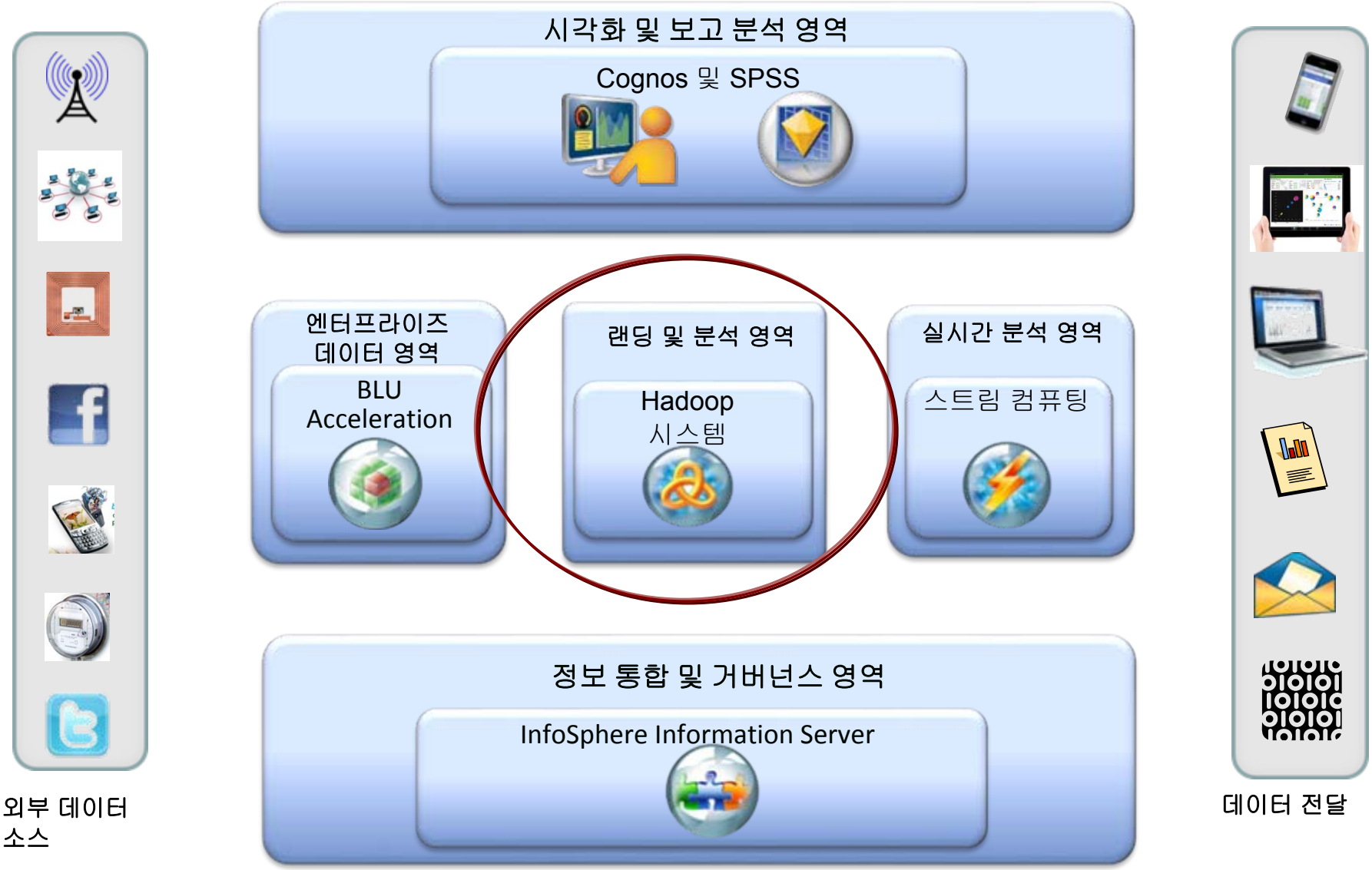


스마트한 병원의 빅데이터 분석

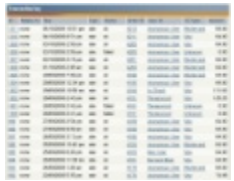
***University of Ontario** 의료진은 빅데이터를 통해 신생아 모니터링을 적용함으로써 **24**시간 전에 **ICU**의 감염을 예측합니다.*

IBM Data Baby
youtube.com

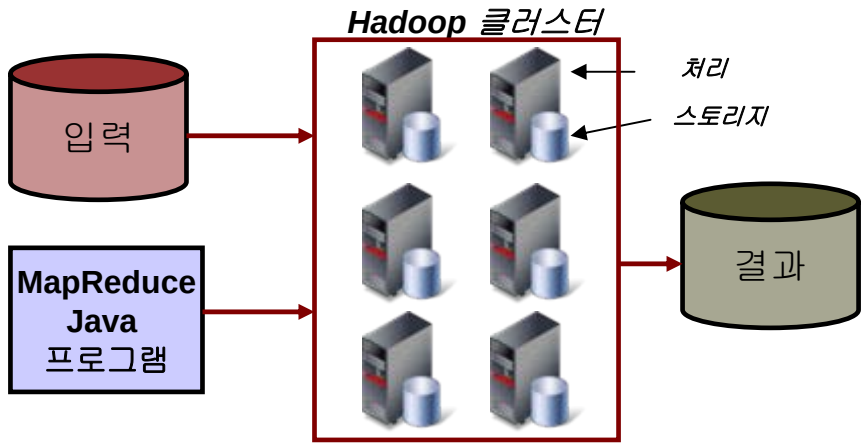
IBM 빅데이터 플랫폼은 빅데이터 과제를 해결할 수 있는 완전한 에코시스템



움직이지 않는 데이터에서 가치 얻기

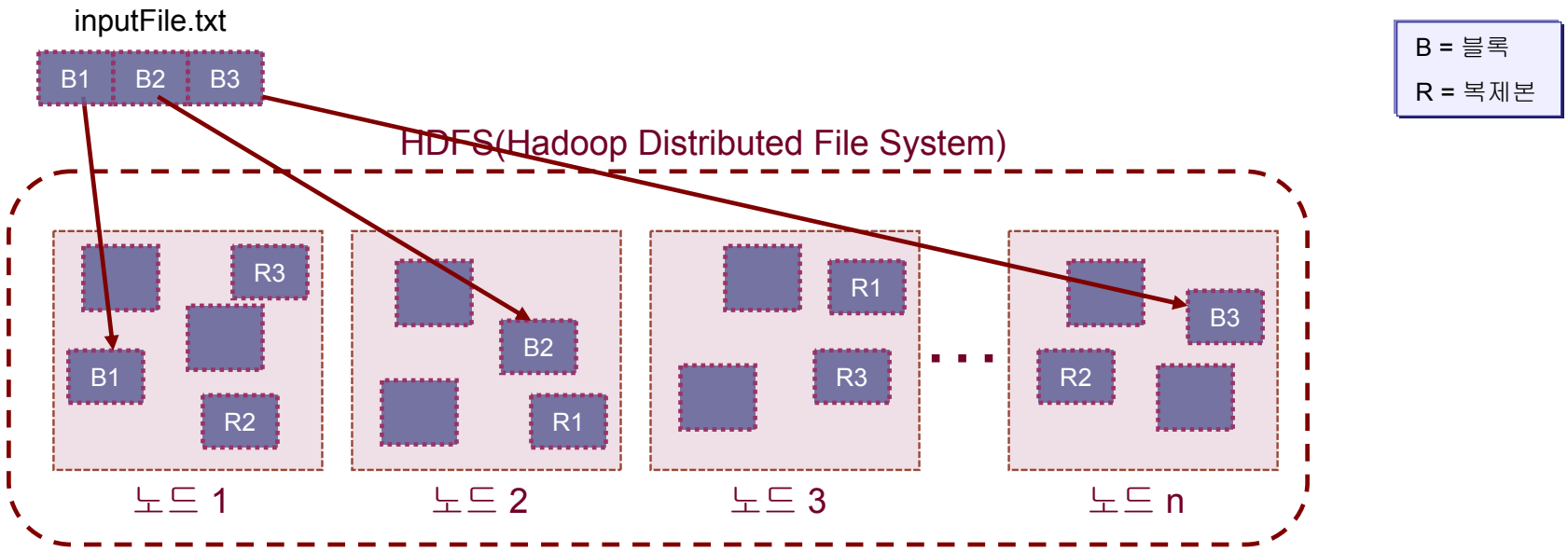
데이터 소스	분석	비즈니스 가치
웹 로그 	*e-commerce 사이트에서 온라인 쇼핑객 동작 분석*	*소매 웹 사이트 매출 최대화*
소셜 미디어 	*고객 감정 및 경험 분석*	*고객 유치 및 유지*
기상 데이터 	*방대한 양의 기상 데이터 내역 분석*	*최적의 풍력 발전용 터빈 배치 결정*

InfoSphere BigInsights는 Apache Hadoop을 활용합니다



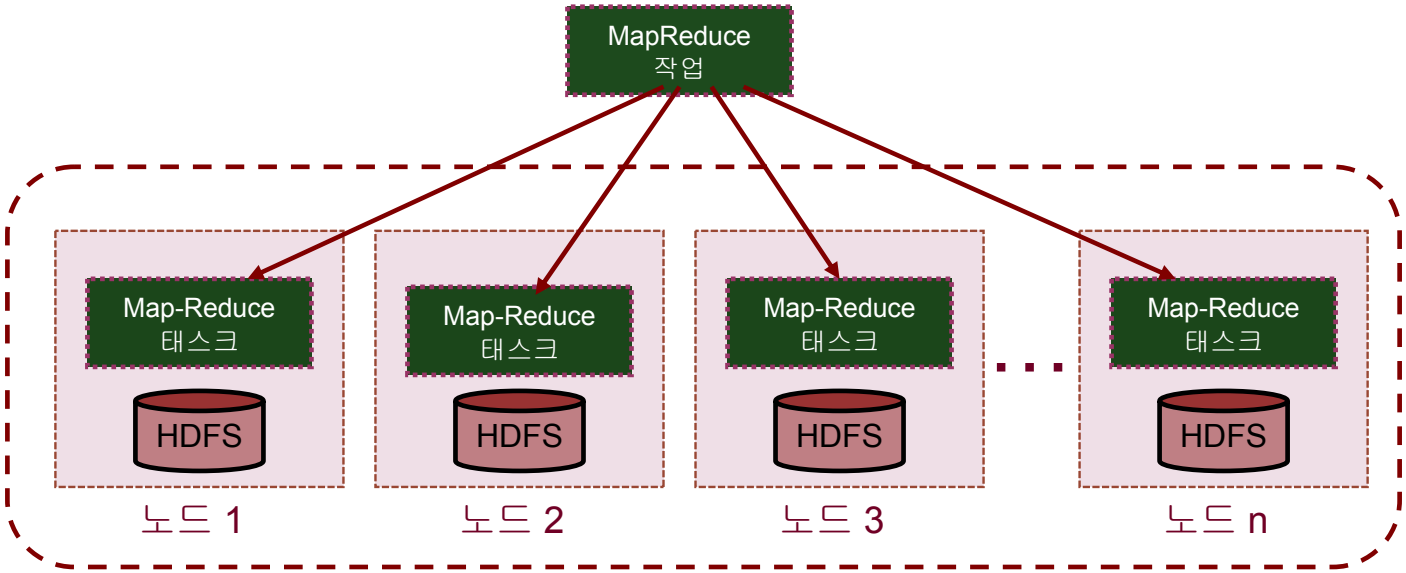
- 저렴한 하드웨어 클러스터로 구성
  - 노드에 프로세서, 메모리 및 디스크가 있음
- 특별한 파일 시스템 – HDFS(Hadoop Distributed File System)
- 특별한 프로그래밍 모델 – MapReduce

HDFS(Hadoop Distributed File System)는 Hadoop 클러스터에 데이터를 분산시킴



- Hadoop 클러스터의 모든 노드에 펼쳐져 있는 분산 파일 시스템
- 로드 시 파일이 블록으로 자동 분할되어 여러 데이터 노드로 펼쳐짐
- 시스템은 노드가 실패할 것으로 가정
  - 여러 노드에서 데이터를 복제하여 신뢰성 확보
- 유연하게 확장 가능

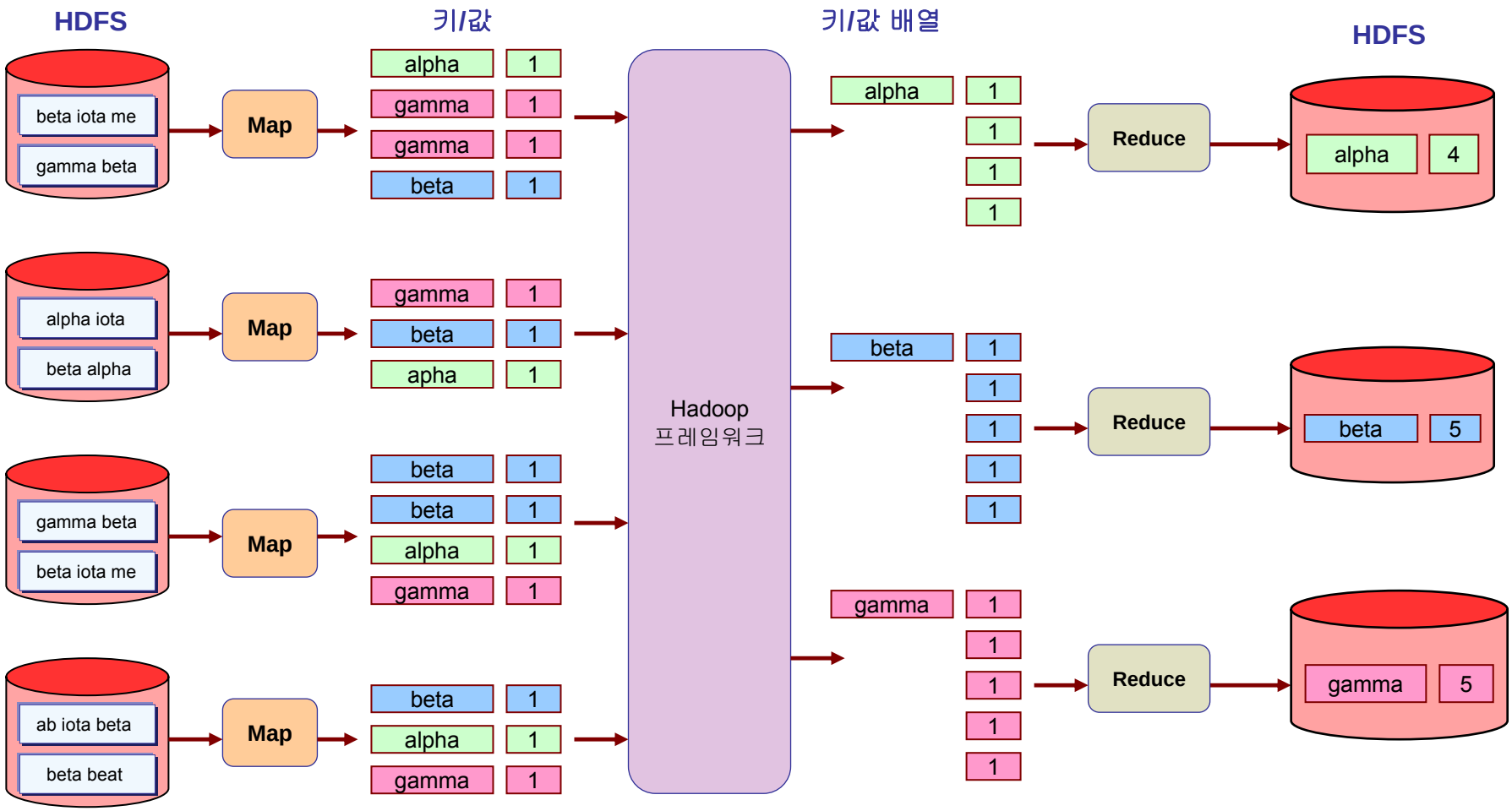
MapReduce 프레임워크는 프로그램을 데이터 노드로 전송



- MapReduce 작업은 개별 노드로 전송됨
- Map-Reduce 태스크는 여러 노드에서 동시에 실행됨
- Hadoop 프레임워크는 많은 양의 “대규모 이동”을 수행
  - 예: map-reduce 태스크 간 데이터 이동

간단한 **MapReduce** 예: 텍스트에서 문자열 발생 횟수 계산

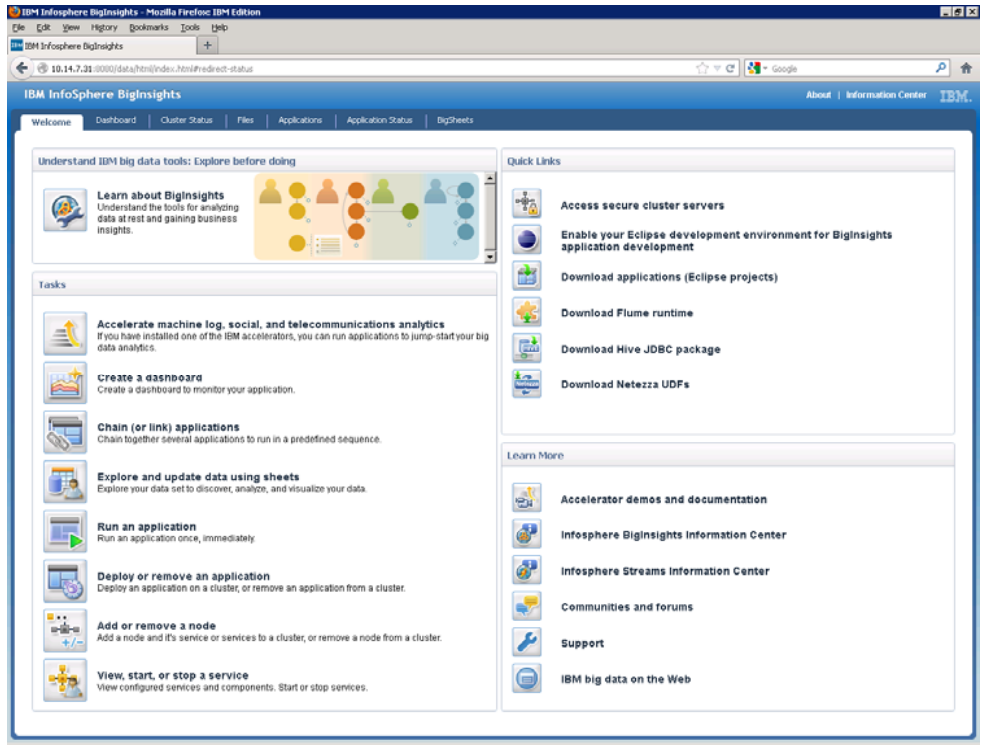
*목표: 텍스트 파일에서 **alpha, beta** 및 **gamma** 발생 횟수를 계산합니다.*



BigInsights를 통해 모든 빅데이터 역할을 용이하게 수행

- 관리자 역할
  - 클러스터 전체 관리
    - 구성요소 모니터/시작/중지
    - 노드 추가/제거
  - 포털 스타일 대시보드
- 개발자 역할
  - Eclipse 기반 도구
  - HDFS에 대한 읽기/쓰기 액세스
  - 시스템의 작업 및 워크플로우에 대한 통합된 뷰
  - 애플리케이션 스테이징, 런칭 및 스케줄링 센터
  - 여러 내장 액셀러레이터
- 비즈니스 사용자 역할
  - Java 프로그래밍 스킬 필요 없음
  - 스프레드시트 도구
  - 시각화

InfoSphere BigInsights Console



고객 불만사항을 분석하려 하는 **Service Oriented Finance**

고객이 무엇에 불만을 가지고 있는지 알아야 합니다.

Service Oriented Finance CMO

IBM은 **BigInsights**를 통한 감성 분석으로 도움을 드립니다.

IBM

감성 분석 – 빅데이터의 과제이자 기회

방대한 규모의 비정형 데이터

다음 사항을 결정

- 제품 수요
- 신제품 수용
- 경쟁 위협
- 브랜드 위상에 대한 위협
- 광고 목표

소셜 미디어 데이터에서 감성 파악

데모: BigInsights로 Twitter에서의 부정적 감성 분석

데이터 소스
Twitter

주제
Service Oriented Finance

호감

- 수표 보호 기능에 만족
- 온라인 청구서 지급 기능에 만족
- ATM이 도시 곳곳에 배치되어 있는 것에 만족
- 서비스 담당자에 만족

반감

- 온라인 뱅킹 기능을 신뢰하지 않음
- 오래 대기하는 것을 꺼림
- ATM 수수료에 불만족
- 당좌대월 수수료에 불만족

BigInsights는 다른 Hadoop 배포에 없는 기능을 제공

- 보안
  - LDAP 인증
  - 역할 기반 권한 부여
- 성능 및 최적화
  - 어댑티브 MapReduce
  - 고급 스케줄러
  - 대규모 인덱싱을 위한 BigIndex
  - 빠르고, 분할 가능한 압축
- Optim Development Studio
  - Eclipse 기반 Java IDE
- 빅데이터 통합
  - Information Server, InfoSphere Streams, Netezza, DB2
- 분석 액셀러레이터
  - BigSheets 스프레드시트 및 시각화
  - 장비 데이터
  - 소셜 미디어
  - 고급 텍스트 분석
  - JAQL 쿼리 언어

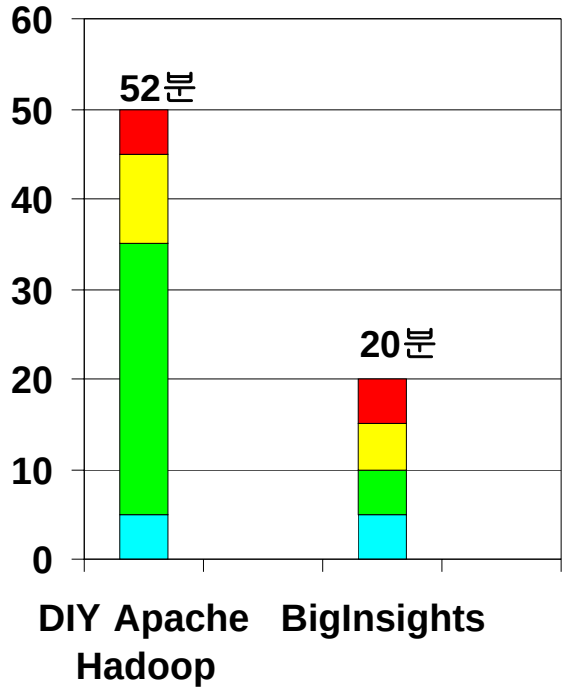
GIGAOM PRO

*"**Cloudera**는 어플리케이션 수준에서의 **Hadoop** 처리를 위한 스택 확장 및 제품 개발에 대한 계획이 없습니다... **IBM**은 엔터프라이즈급 오퍼링에 근접한 배포에 초점을 두고 있습니다."*

개발자의 생산성을 크게 향상시키는 **Machine Data Accelerator**

로그 파일 분석

작업	DIY	MDA
IDE 설치	5분	5분
코드 개발	30분	5분
패키징 및 배포	10분	5분
코드 테스트	7분	5분
코드 라인 수	57	7

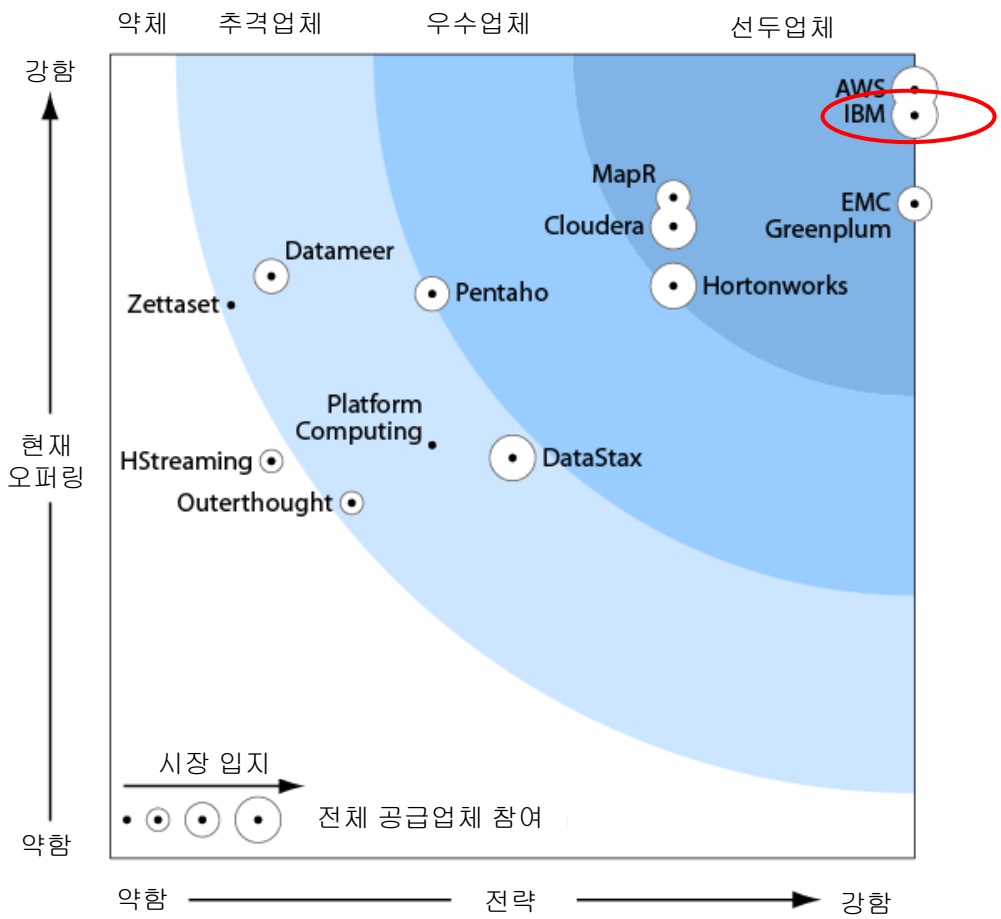


Machine Data Accelerator는 로그 파일 분석 작업에서 개발 시간을 절반으로 단축

새로운 코드가 **8배** 더 적게 필요함

생산성 향상은 테스트 케이스 범위에 따라 달라지며, MDA 모듈을 더 많이 재사용할수록 향상률이 커집니다.

Forrester, IBM을 Hadoop 솔루션 분야의 최고로 선정



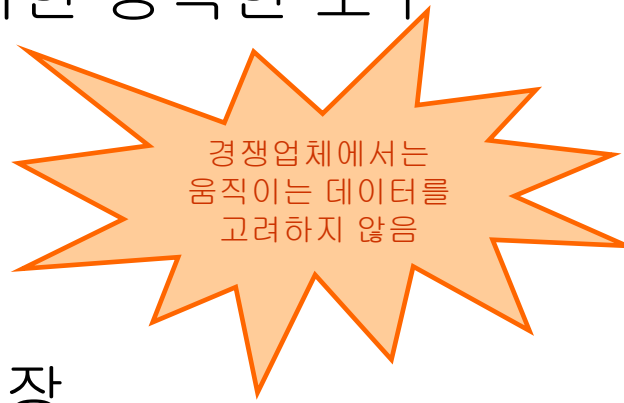
약함 전략 강함

"IBM은 가장 깊이 있는 Hadoop 플랫폼 및 애플리케이션 포트폴리오를 보유하고 있습니다."

Hadoop 에코시스템을 위한 엔터프라이즈급 SQL 지원

IBM은 가장 완전한 빅데이터 플랫폼 보유

- Streams는 빠른 속도의 실시간 분석을 위한 강력한 도구
  - Drag and Drop 방식의 간편한 개발
  - 광범위한 시각화 기능



- BigInsights는 Hadoop을 엔터프라이즈급 빅데이터 플랫폼으로 확장
  - 고급 액셀러레이터가 빠른 가치 실현 지원
  - Hadoop 데이터에 대한 ANSI SQL 지원 제공

FORRESTER®

"IBM은 가장 깊이 있는 Hadoop 플랫폼 및 애플리케이션 포트폴리오를 보유하고 있습니다."