Stream Computing : In-motion Data Analytics

2011/06/16
JeongKwon Lee

IBM Information
On Demand
Comes to You 2011

똑똑한 정보, 똑똑한 비즈니스

Agenda

Big Data ?

Big Data Strategy of IBM

Stream Computing ?

InfoSphere Streams

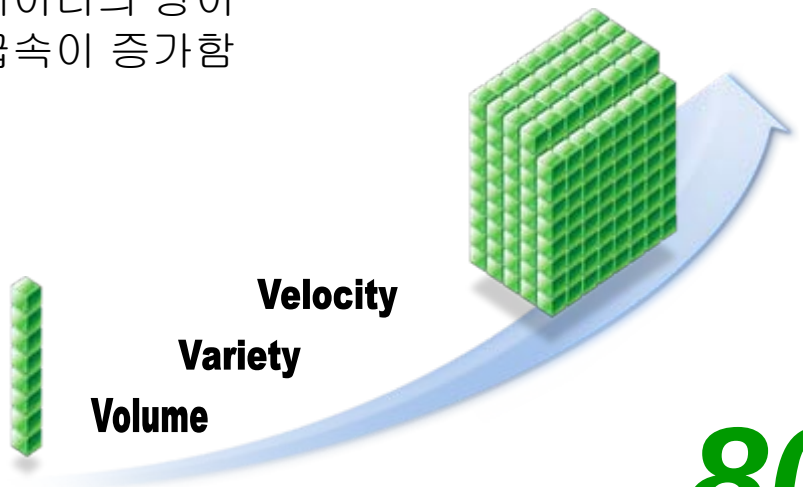
Use Case

정보의 증가와 분석의 필요

44x

데이터의 양이
급속이 증가함

2020
35 zettabytes



2009
800,000 petabytes

80%

의 데이터가 비정형
구조이다.

1 in 3 신뢰하기 어려운 정보에 기반하여 **decision making**하는 경우가 많다.

1 in 2 실제 필요로 하는 정보에 대해 액세스하기 어렵다.

83% 경쟁력 확보를 위해서 **business intelligence**와 분석이 중요하다고 생각한다.

60% 신속한 결정을 위해 정보에 대해 보다 빨리 처리하여 이해할 필요가 있다.

Big Data의 속성

정형/비정형 데이터
유형에 대한 분석

배치 형태의 처리에서
실시간 처리 요건

테라바이트에서
제타바이트까지 확장

Big Data 접근 방식

Traditional Approach
정형 & 반복적인 분석

Business Users

무엇을 요청할지 결정

IT

해답을 위한 데이터 구조화

- 월 영업 보고서
- 이윤 분석
- 고객 서베이

Big Data Approach
반복적 & 탐구적인 분석

IT

분석을 할 수 있는 플랫폼 전달

Business

어떤 것들이 분석 가능한지 찾아본다.

- 소비자 성향
- 제품 전략
- 자산의 최대 활용

IBM invest on Big Data Research

IBM to invest $100 million for big-data analysis research

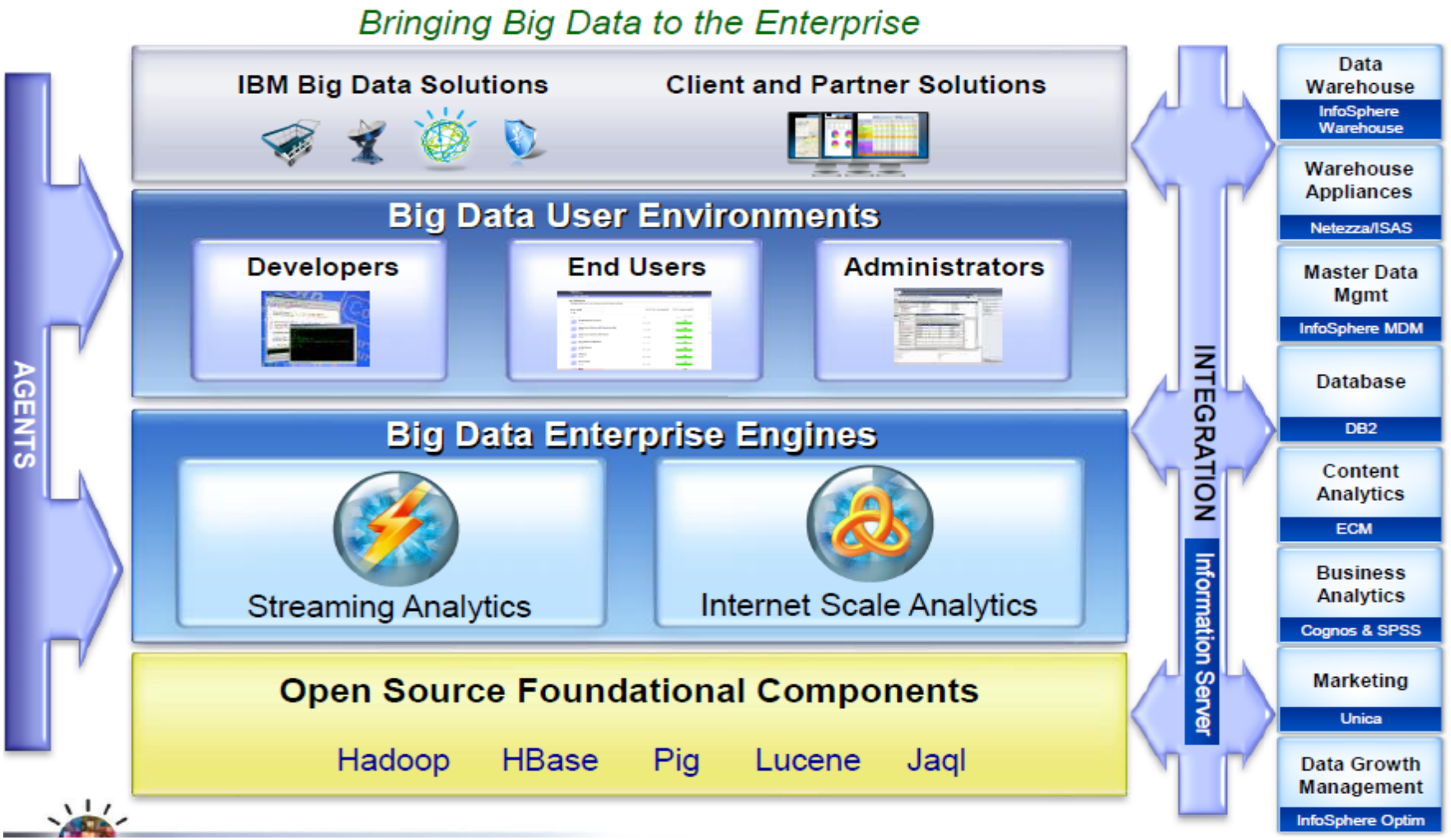
The company has released new tools for data analysis

May 20, 2011, 6:02 PM — Sensing a greater need in big-data analysis tools, IBM will invest $100 million to research advanced large-scale analytics, the company announced Friday.

Watson Inspires New IBM 'Big Data' Analytics Tools

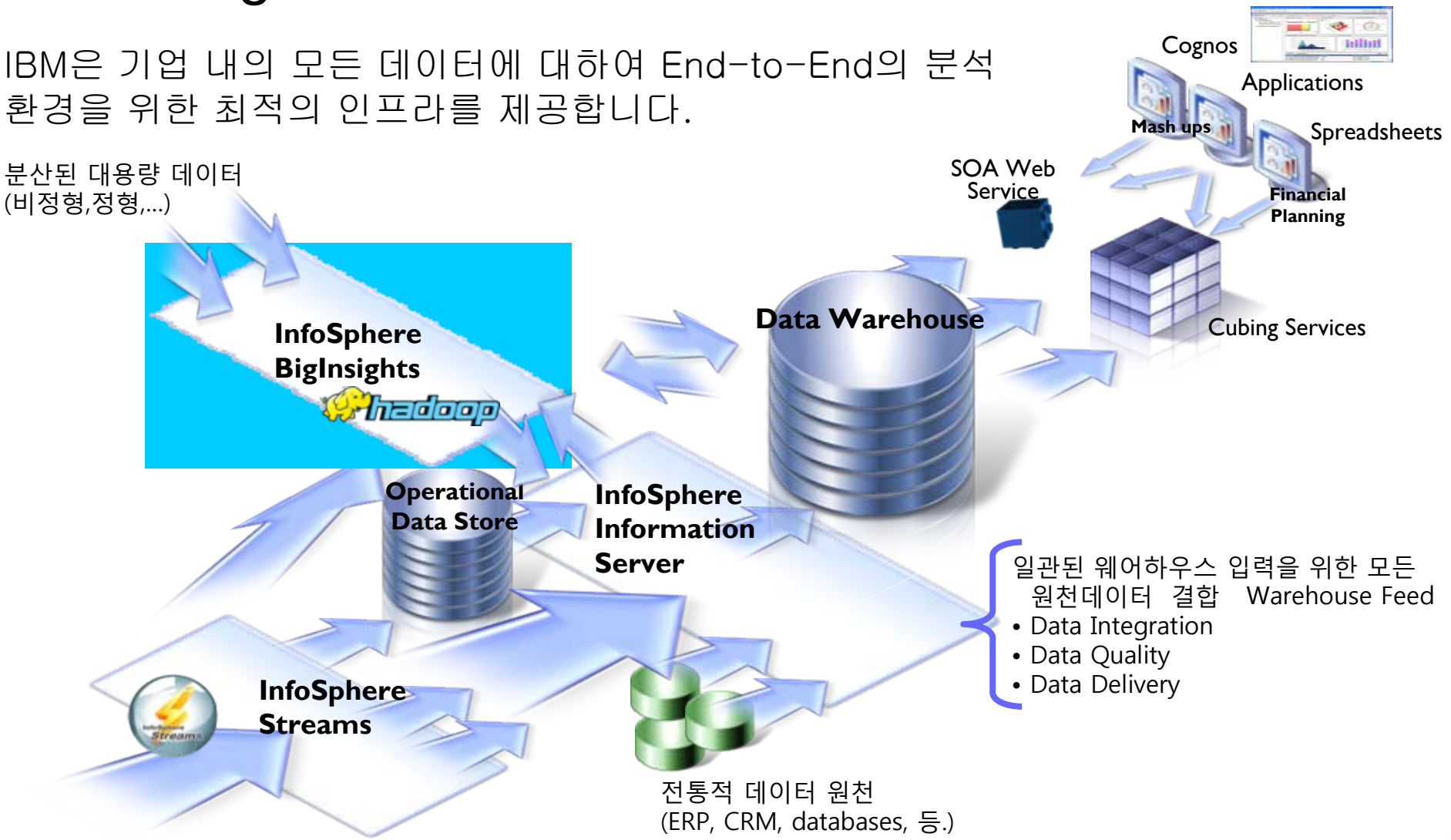
Remember Watson, the IBM computer that bested two human competitors on TV's Jeopardy? Building on what it learned with Watson and other research projects, IBM is unveiling new software and services to analyze tens of petabytes of data -- with subsecond response times.

IBM's Big Data Platform



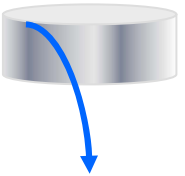
IBM's Big Data Initiative

IBM은 기업 내의 모든 데이터에 대하여 End-to-End의 분석 환경을 위한 최적의 인프라를 제공합니다.



Streaming Computing

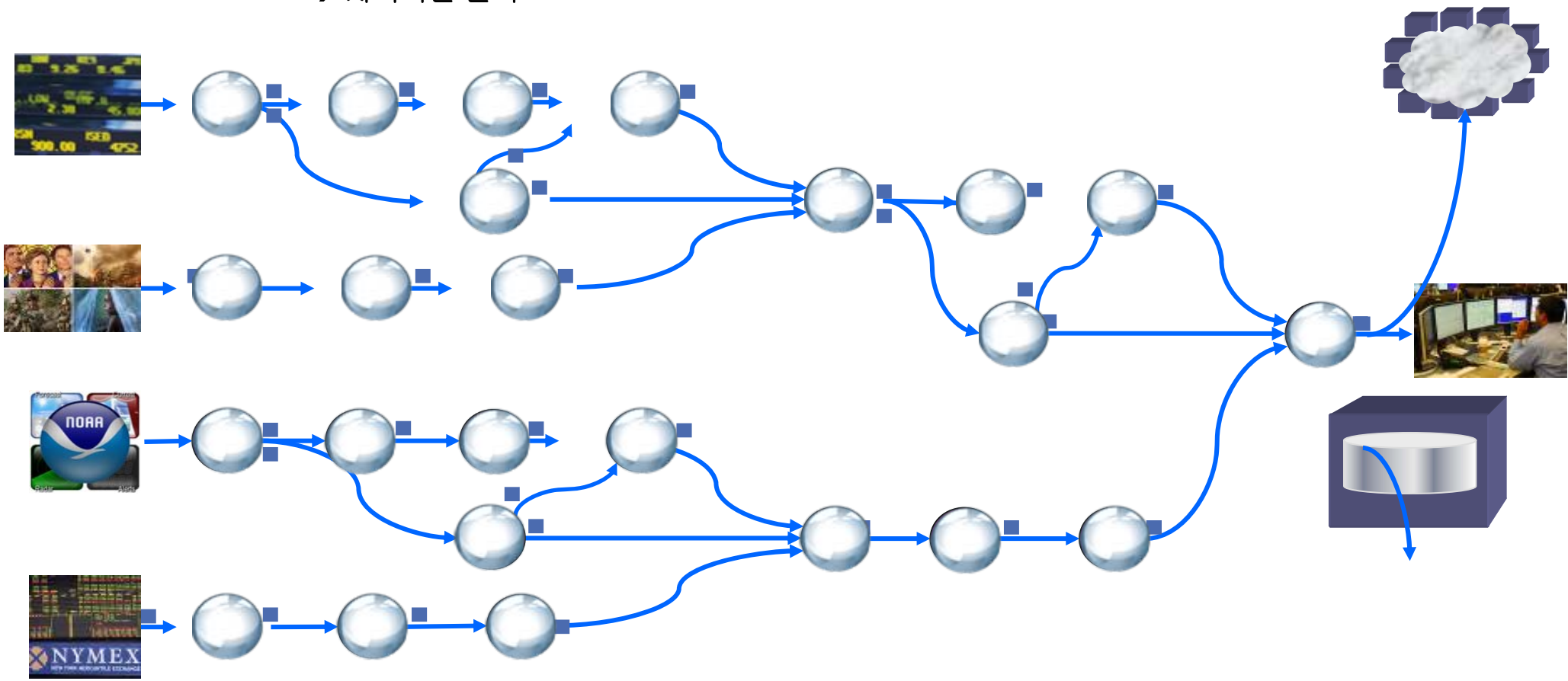
→ 지속적인 데이터 적재 → 지속적인 분석



Stream Computing

→ 지속적인 데이터 적재

→ 지속적인 분석



확장성 -

어플리케이션들을 개별 수행 요소로 파티셔닝

다양한 데이터 유형

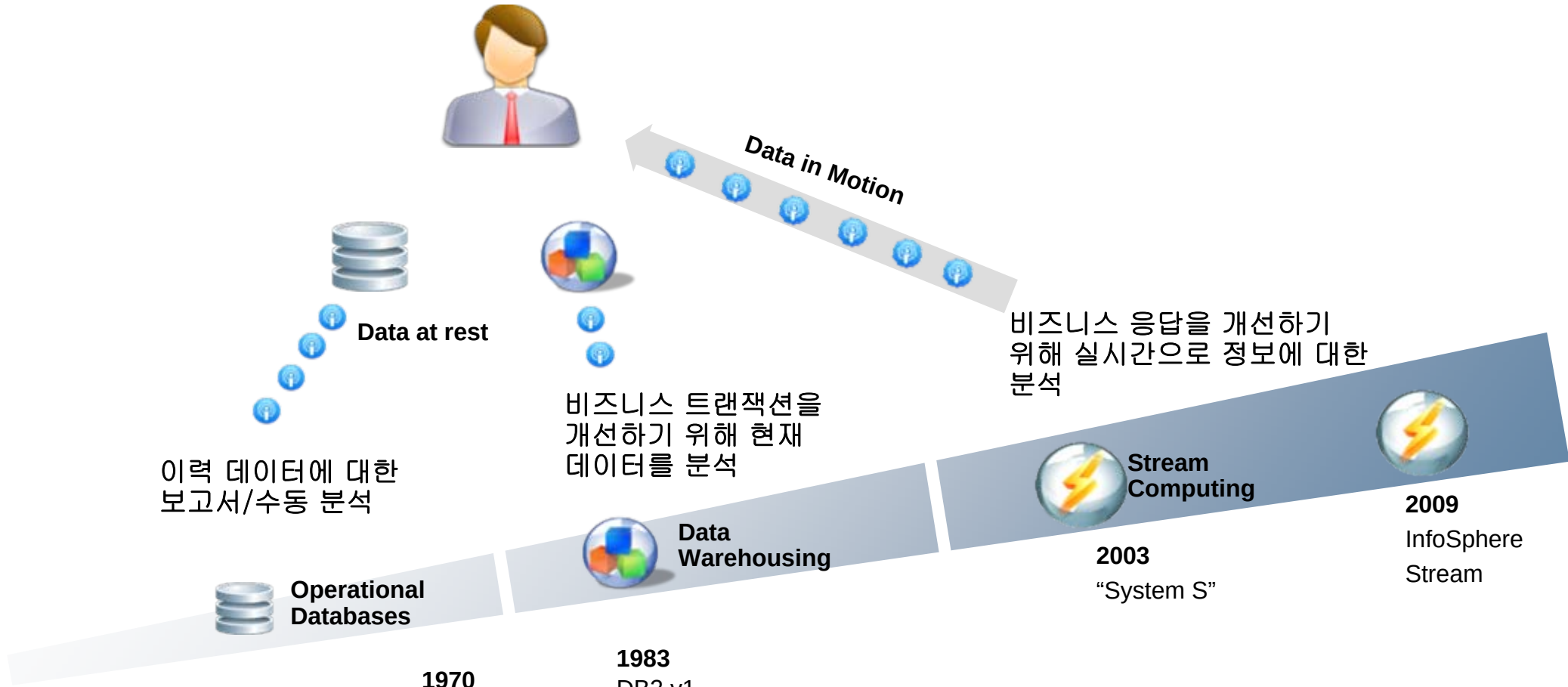
이벤트와 데이터의 스펙트럼

정형 데이터 ⟷ 비정형 데이터

- 활용 가능성이 높음
- 단순 분석
- 잘 정의된 이벤트
- 빠른 속도 (million events / sec)
- 매우 빠른 응답시간

- 활용 가능성이 적음
- 복합 분석
- 이벤트가 감지될 필요는 있다. - 대용량 (TB/sec)
- 빠른 응답시간

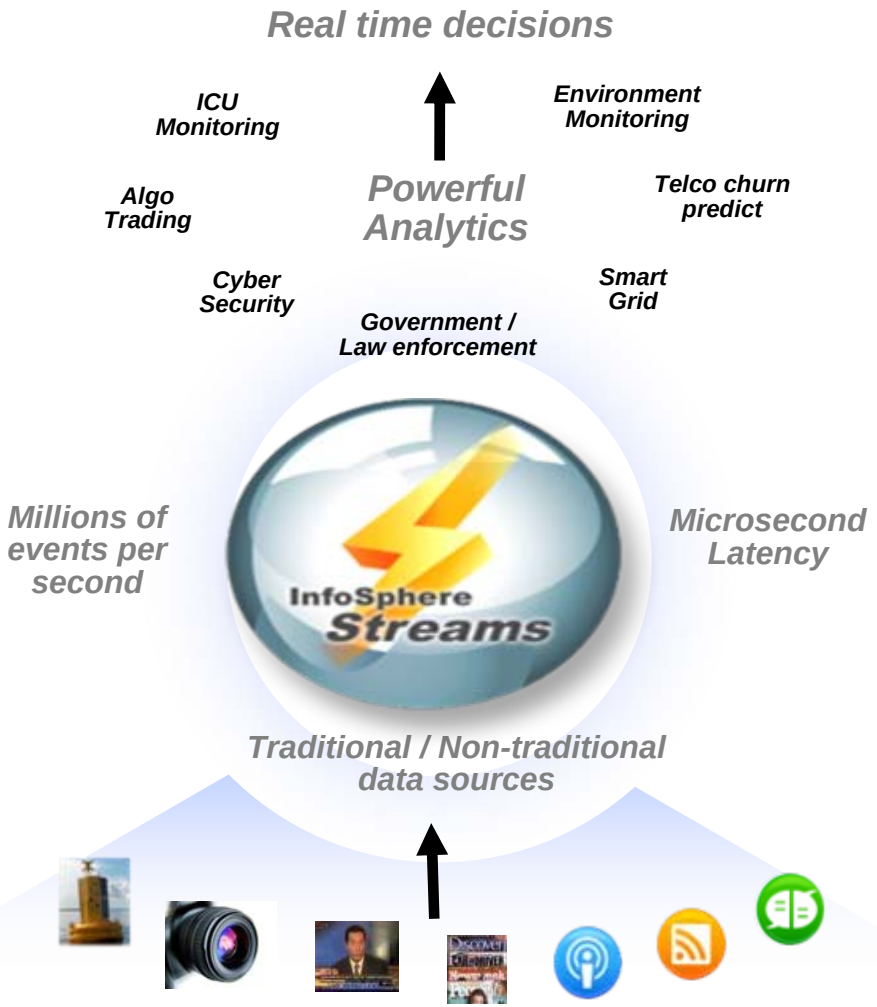
데이터 분석의 발전 방향



IBM’s history of innovation

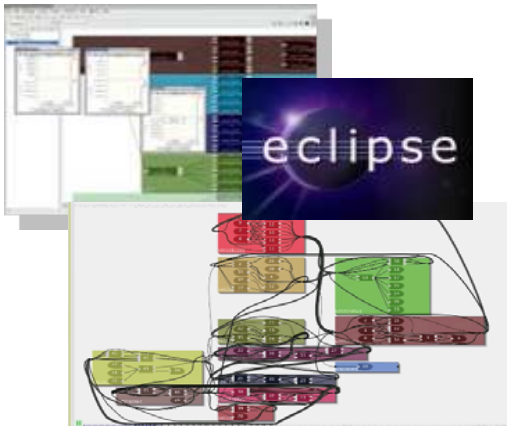
In-motion Analysis

- **Volume**
  - Terabytes / 초
  - Petabytes / 일
- **Variety**
  - 모든 유형의 데이터
  - 모든 유형의 분석
- **Velocity**
  - Microsecond 수준의 응답 시간
- **Agility**
  - 동적으로 deploy
  - 빠른 어플리케이션 개발



InfoSphere Stream 2.0

개발 환경



- Eclipse IDE
- Streams Live Graph
- Streams Debugger

운영 환경

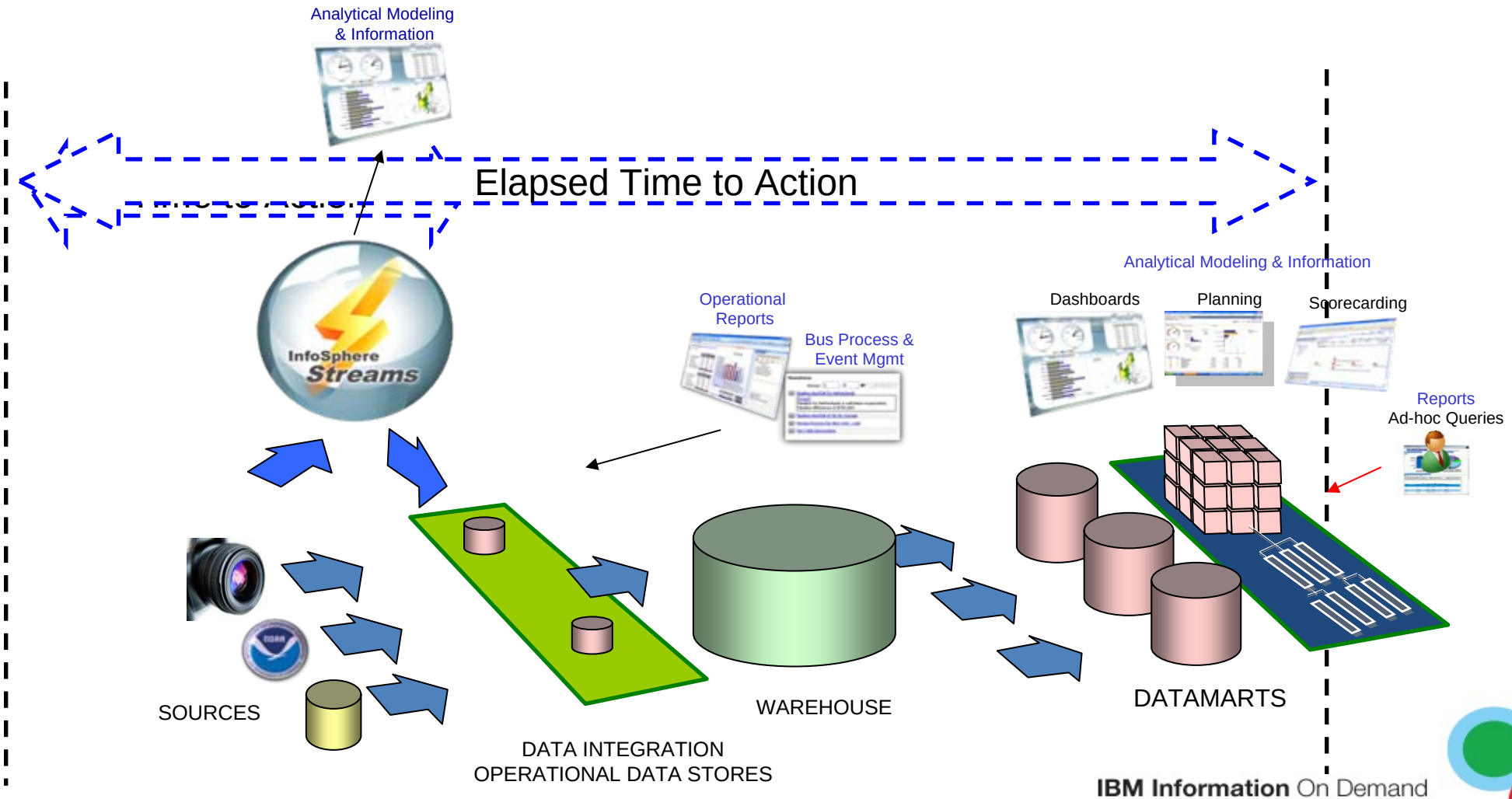
- RHEL v5.3 and above
- x86 multicore hardware
- InfiniBand support
- Clustered runtime for near-limitless capacity

툴킷 & 통합 기술

- Standard Toolkit
- Internet Toolkit
- Database Toolkit
- Financial/Mining Toolkit
- Stream Live Monitoring
- Math & Text Processing

InfoSphere Stream : Velocity

InfoSphere Stream은 보다 빠른 분석 환경을 제공합니다.



InfoSphere Stream : Variety

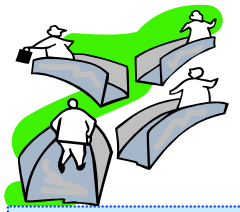
InfoSphere Stream은 다양한 데이터 유형에 대한 분석 환경을 제공합니다.

Punctor operator

스트림에 punctuation mark을 삽입

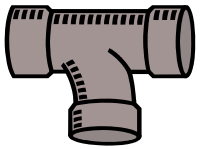
Aggregate operator

스트림 데이터에 대한 그룹핑 및 요약 작업



Split operator

하나의 스트림을 여러 개의 스트림으로 분해



Join operator

2개의 스트림 간의 연관 기능

Stream-Relational Toolkit

Sort operator

스트림 데이터에 대한 윈도 기반하의 정렬

Functor operator

스트림 내의 튜플 레벨에서의 데이터 조정

Delay operator

스트림의 속도를 일부로 지연시키는 기능

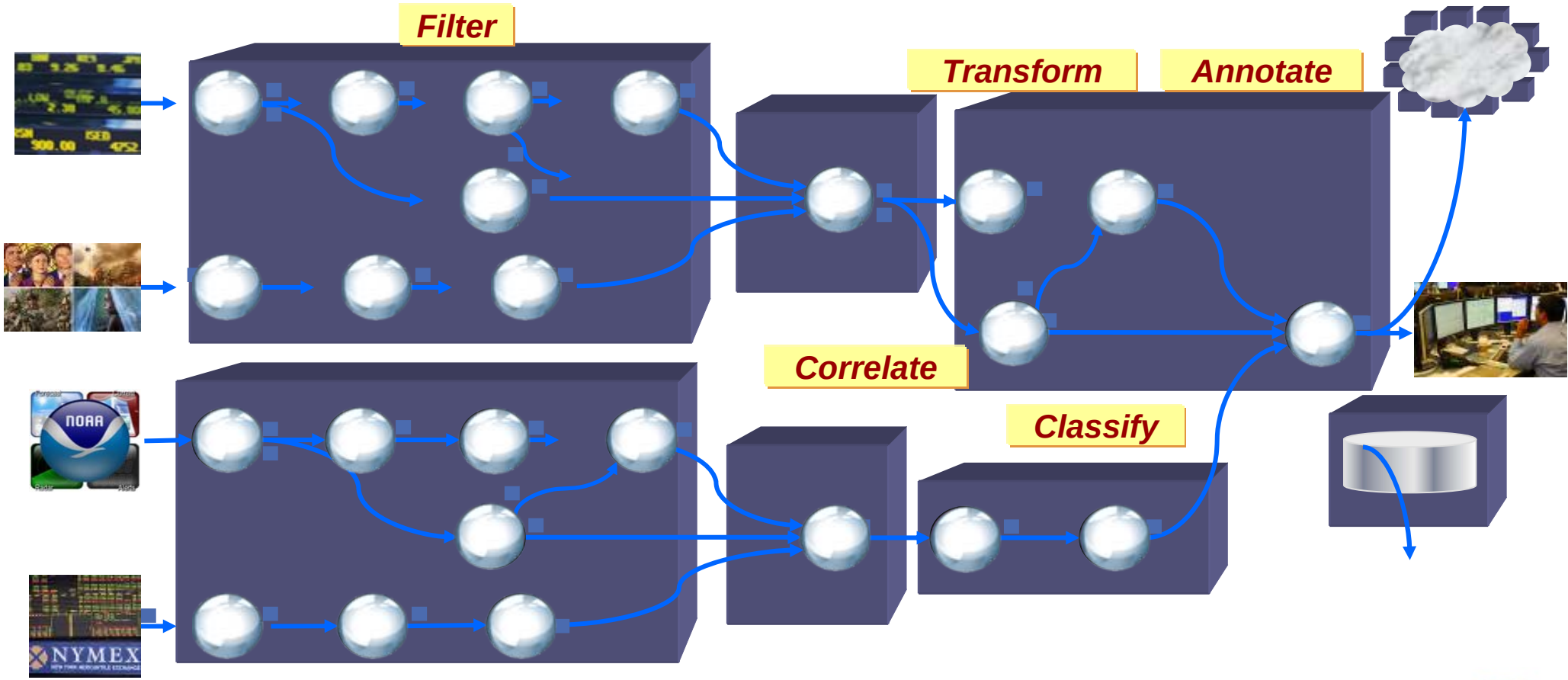
Barrier operator

동기화 포인트를 찍는 기능

InfoSphere Stream : Volume

InfoSphere Stream은 대용량의 데이터를 처리할 수 있습니다.

→ 지속적인 데이터 적재
→ 지속적인 분석



확장성 -
개별 수행 요소들을 여러 개의 하드웨어 노드로 분산

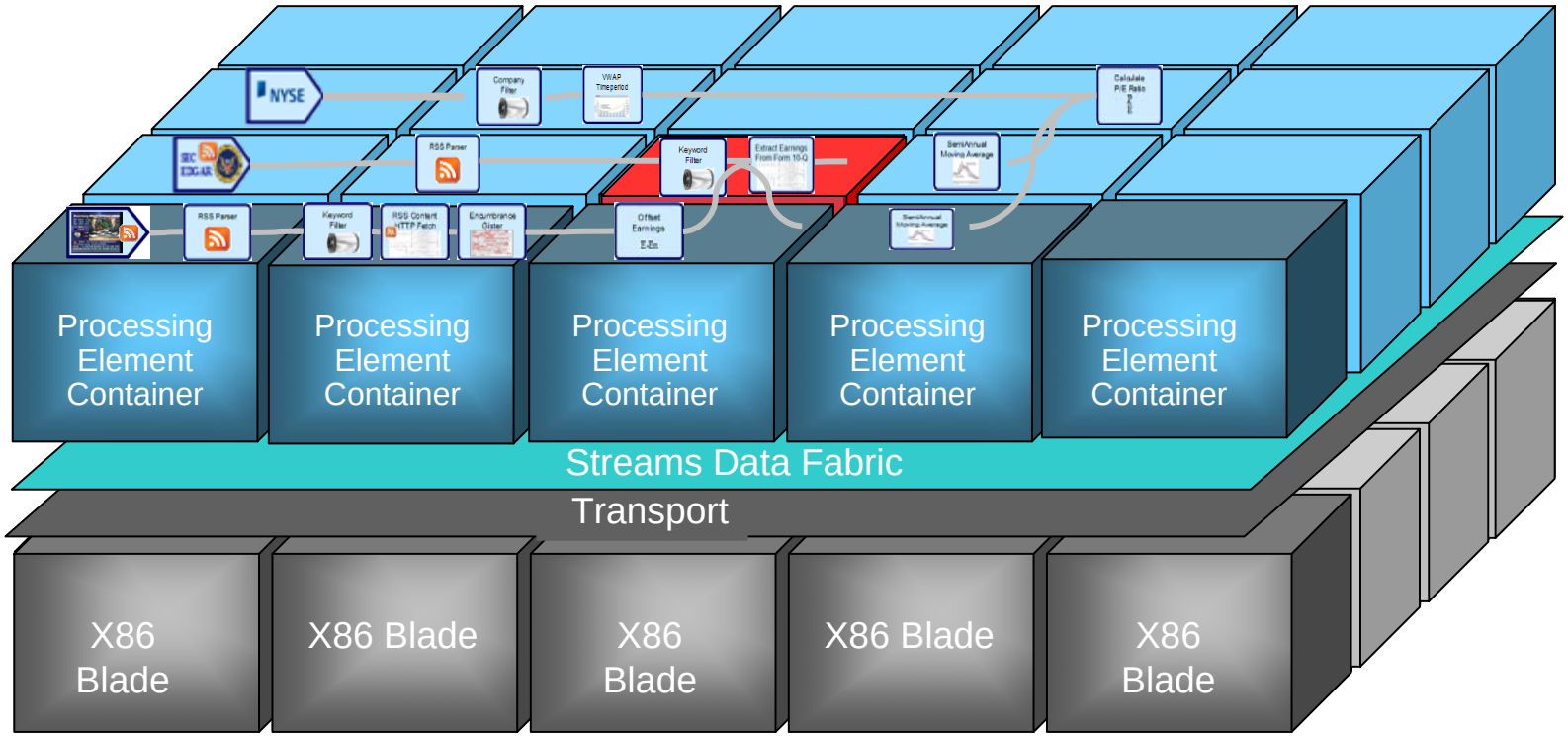
InfoSphere Streams : Agility

InfoSphere Stream은 어플리케이션의 동적 Deploy 및 시스템 자원 현황에 따라 유동적으로 수행됩니다.

스케줄러가 PE에 대한 할당 및 지속적인 자원 관리 수행

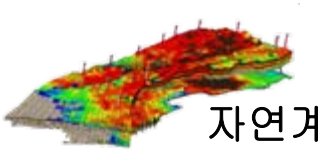
한 개 노드에서 블레이드 센터까지 multi-rack환경에서 수행됨

자원, 워크로드, 데이터 양에 따라 동적으로 변경



Smart Applications

Use Case



자연계 시스템

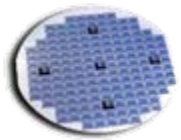
- 지진 모니터링
- 산불 관리
- 웨이퍼 관리

운송

- 지능적인 트래픽 관리

제조

- 마이크로칩 제조에 대한 프로세어 제어



주식 시장

- 주가에 영향을 미치는 날씨 점검
- 마켓 데이터에 대한 빠른 응답시간 필요

Law Enforcement

- 실시간 다 모델 감시

사기 예방

- 다자간 사기 방지
- 실시간 사기 감지

Radio Astronomy

- 순간적인 이벤트 감시

Health & Life Sciences

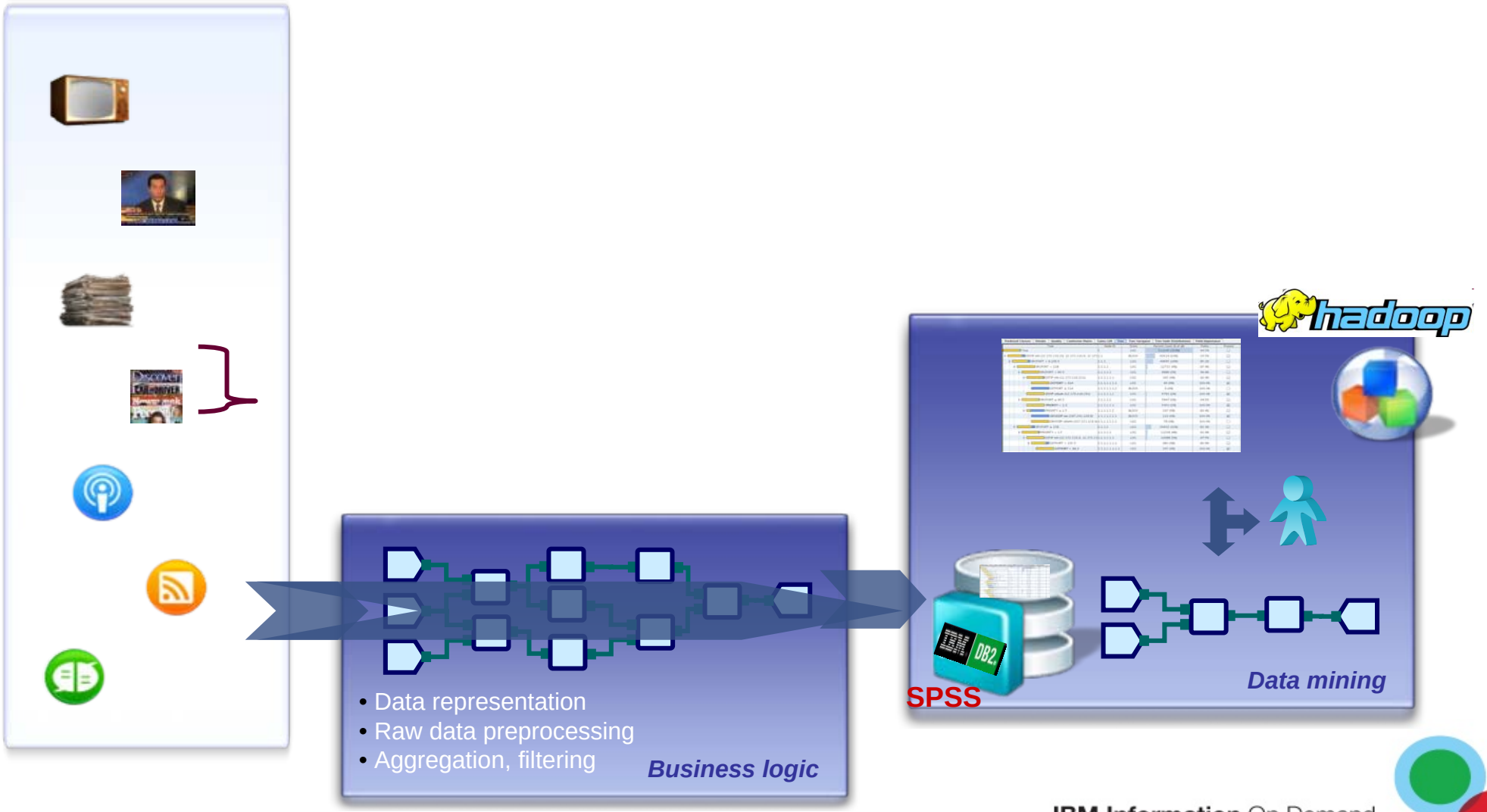
- 신생아 ICU 모니터링
- 전염병 조기 경보 시스템
- 원격 healthcare 모니터링

Telecom

- **콜 데이터 분석**
- **실시간 서비스, 청구, 광고**
- **비즈니스 인텔리전스**
- **Churn Analysis, Fraud Detection**

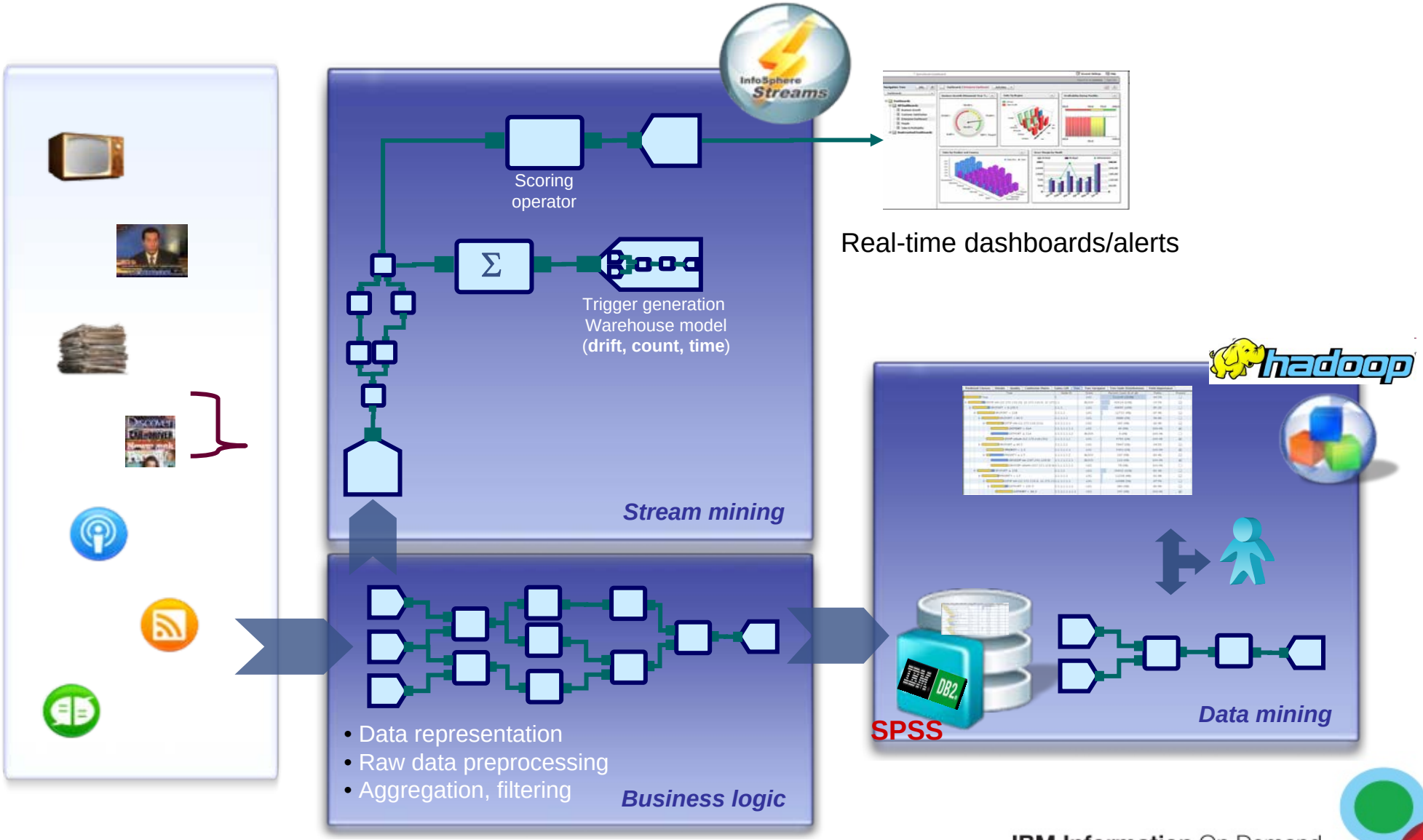
Stream Processing

Use Case



Stream Mining

Use Case



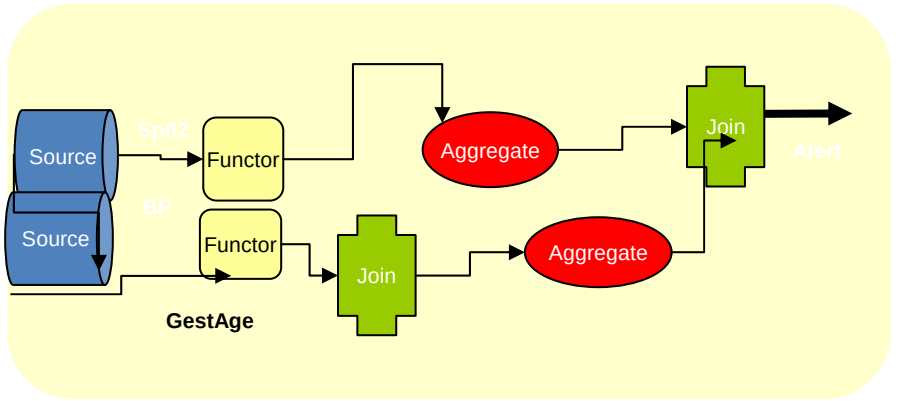
고객 사례 : Neonational ICU

Use Case

Challenge

- 기존의 모니터링 환경에서는 조산아들의 위험을 감지하는데 소요되는 시간이 많이 소요됨
- 결과적으로 조치를 늦게 취하게 되어 심각한 문제를 야기시킴

UOIT
CHALLENGE INNOVATE CONNECT



Results

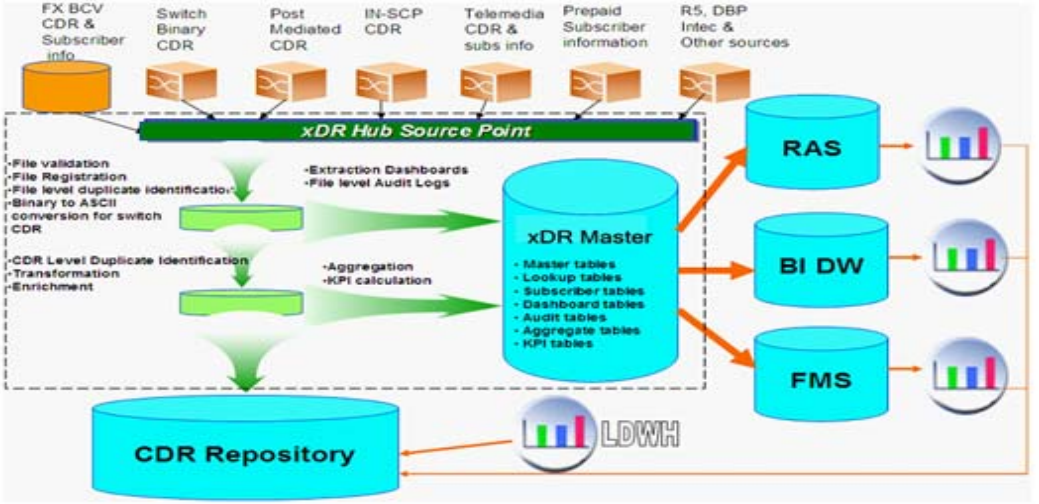
- 조산아들의 생명에 위험을 줄 수 있는 상황에 대한 조기 판단을 할 수 있는 인프라 구축
- 건강 관리 효율이 높아짐
- 건강 관리 비용 감소
- 사망자 수와 사망률이 감소

- 생리학상의 데이터 스트림에 대한 연관 관계 및 실시간 분석
  - *혈압, 온도, EKG, 혈중 산소 농도,...*
- 생명을 위협할 수 있는 상황에 대한 조기 진단
  - *이전보다 24 시간 빨리 진단*
  - *조기 진단을 통해 환자의 치사율을 낮춤*
- 의사들에게 새로운 치료 방안을 검증해 볼 수 있는 환경 제공

고객 사례 : Telco Company

Challenge

- 사용자 증가에 따른 데이터 볼륨 증가
- DataWarehouse 내에서의 CDR 프로세싱에 의한 부하



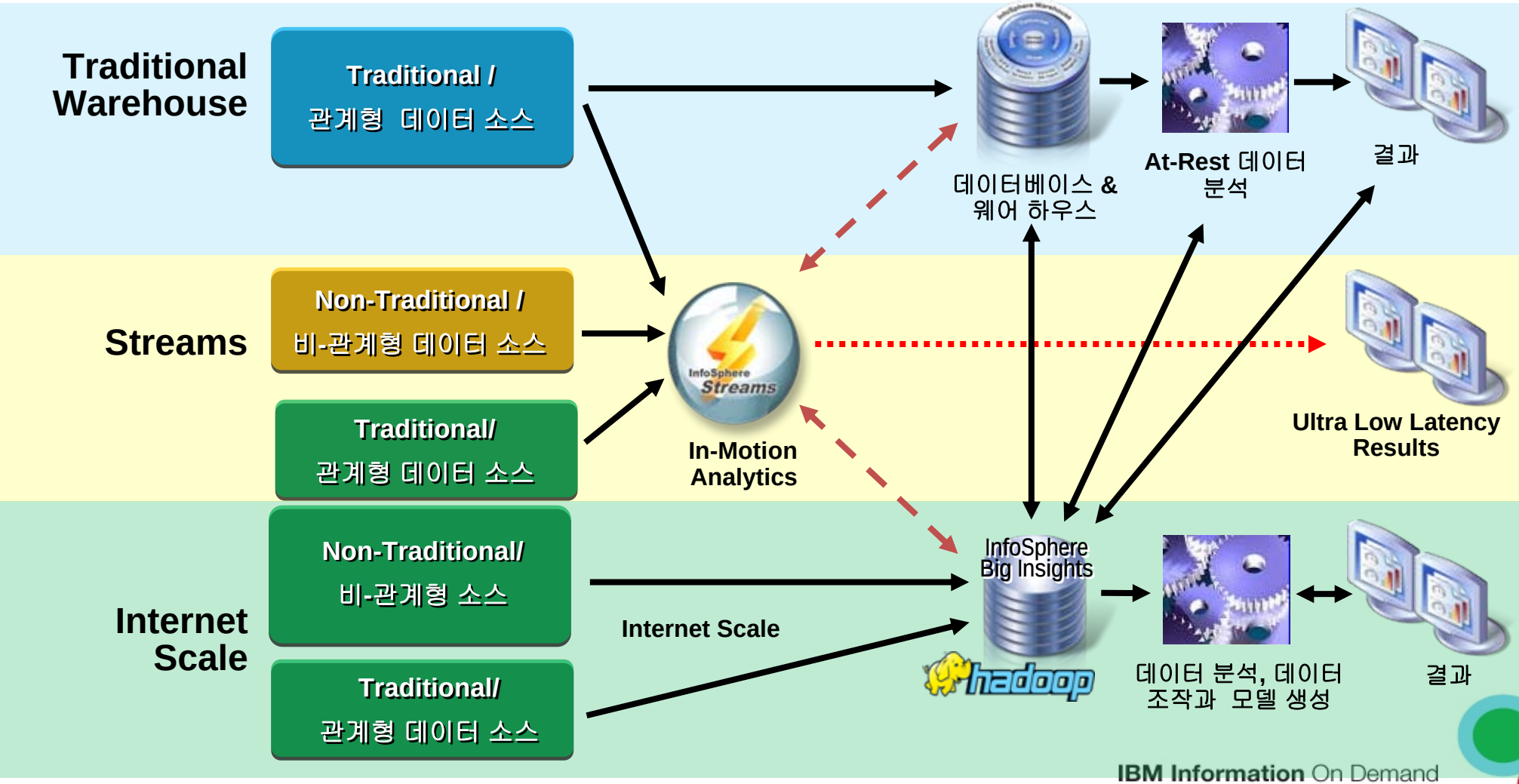
Results

- CDR 데이터에 대한 pre-processing를 통해 데이터웨어하우스의 부하를 줄여줌
- 기존 시스템 대비 8배 성능이 좋아짐
- 데이터 지연이 12시간에서 1초로 줄어듬

To-be Use Case

- 실시간 고객 이탈 예측 및 방지를 위한 마케팅
- 실시간 사기 방지
- 실시간 문맥 인식 마케팅
- 실시간 네트워크 장비/콜 트래픽 모니터링
- …

기존 Data Warehouse를 넘어서...



감사합니다.

IBM Information
On Demand
Comes to You 2011

똑똑한 정보, 똑똑한 비즈니스