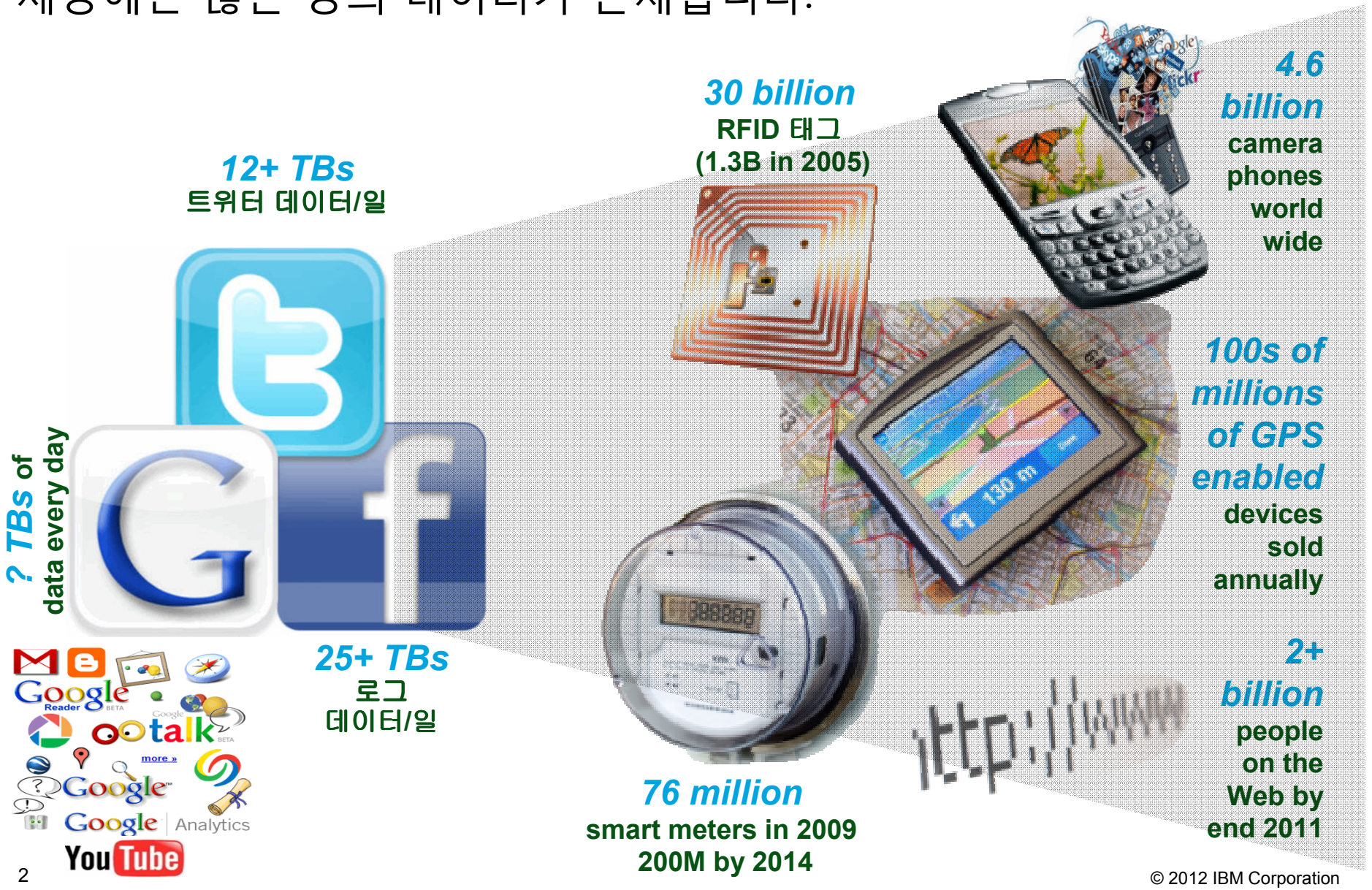
# ‘빅 데이터’를 사용하여 새로운 인사이트 얻기

한국IBM 소프트웨어그룹, 정보관리사업부(Information Management)
이정권 실장(jkwonl@kr.ibm.com)

# 세상에는 많은 양의 데이터가 존재합니다.



# 세상에는 많은 양의 데이터가 흘러다니고 있습니다.

Algorithmic Trading Forum

**알고리즘 옵션 거래를 위해 초당 5M의 마켓 데이터간의 연관 관계를 분석합니다.**

ipdr

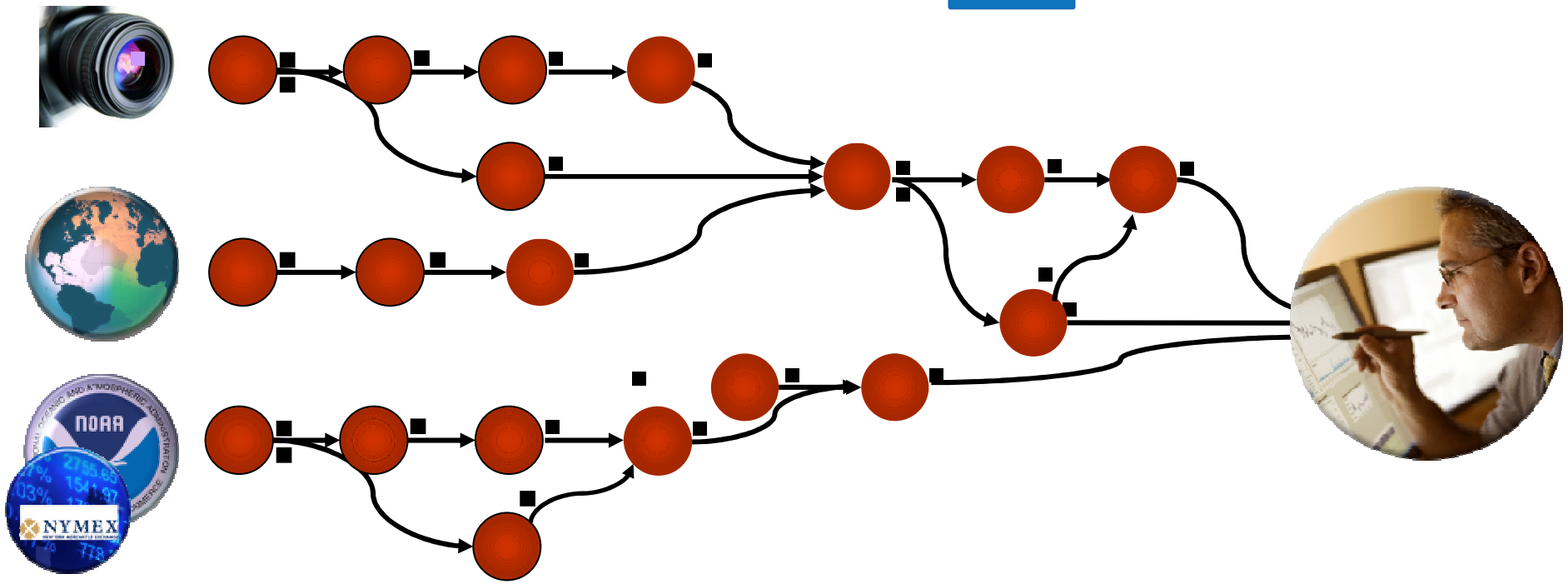
**초당 500K, 매일 6B+ IPDRs 에 대한 분석이 이루어지고 있습니다.**

TELECOM

**RTAP decision making을 위해서 초당 12M CDRs을 수집하고 분석합니다.**

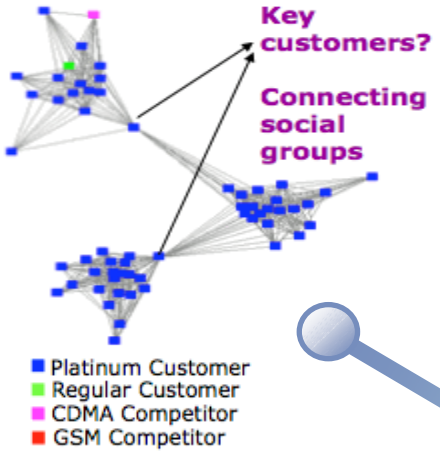
IRF

**초당 6GB, 시간당 21.6TB의 날씨 데이터를 분석하여 구름의 이동 경로와 영향도를 예측합니다.**

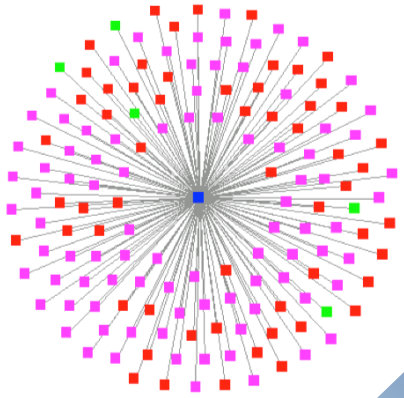


# 그리고 데이터 간에는 연관 관계가 존재합니다.

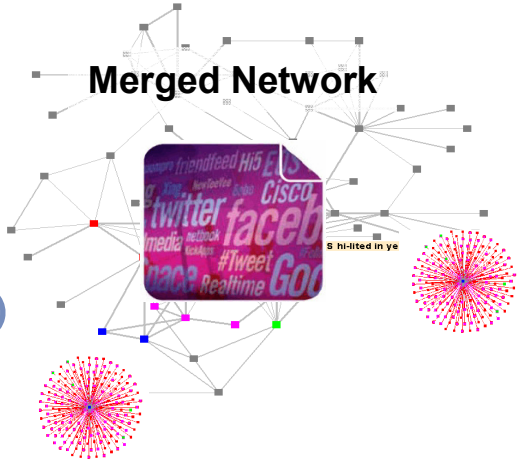
**Calling Network**

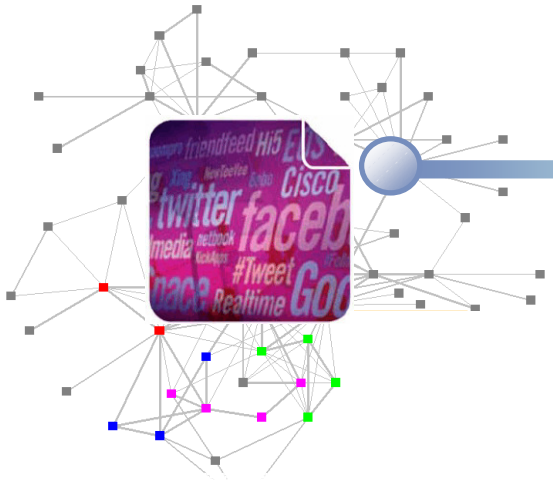



*How do we retain this customer ? (Why is she calling so many competitors ?)*

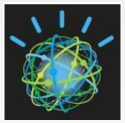


**Merged Network**



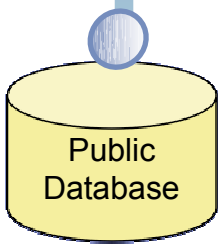


**Social Network**



**홍 길동**

나이, 혼인, 직장, 자녀
- 소셜 네트워크
- 통신 네트워크
- 소비 성향




**Telco company**

스마트 폰을 활용하는 영역은 ?

**Retailer**

소비자 성향이 어떠한지 ?

# 빅 데이터의 특징 – v3 (Volume, Variety, Velocity)

*기존에 가능했던 영역의 범위를 넘어 v3(Volume, Variety, Velocity)의 속성을 가지고 있는 데이터들에서 인사이트를 얻을 수 있습니다.*


**Variety:** 다양한 관계형 및 비 정형 구조의 데이터에 대한 분석

**Velocity:** 스트리밍 데이터/대용량 데이터의 이동

**Volume:** 테라바이트에서 제타바이트까지 확장

# 다양한 대용량 데이터를 분석하여 새로운 통찰력을 얻을 수 있습니다.


- 방대하고 다양한 데이터 분석
- 스트리밍 데이터에 대한 통찰력
- 대량의 구조화된 데이터 분석

IBM 빅 데이터 플랫폼

- Variety
- Velocity
- Volume

*다중 채널 고객 감성 및 경험 분석*

*신생아실에서 예측 분석을 통해 생명을 위협하는 상황을 최대한 빨리 감지*

*기후/지이 데이터 분석으로 풍력 발전기 및 풍력 발전소 부지 플래닝과 날씨 패턴 분석*


*실시간 트랜잭션 데이터를 기반으로 리스크 측정 및 이에 대한 의사 결정*

*산재되어 있는 비디오, 오디오와 데이터 피드를 통해 범인과 위협 감지*

# 빅 데이터를 바라보는 새로운 접근 방식

## 전형적인 접근 방식
*구조적 & 반복적인 분석*

**비즈니스 사용자**

어떤 보고서를
볼지 결정

**IT**

보고서에 필요한
인프라를 구성

- 월 세일즈 리포트
- 이윤 분석
- 고객 서베이

## 빅 데이터 접근 방식
*반복적 & 탐구적인 분석*

**IT**

데이터의 모델을
분석하기 위해 플랫폼
구성

**비즈니스**

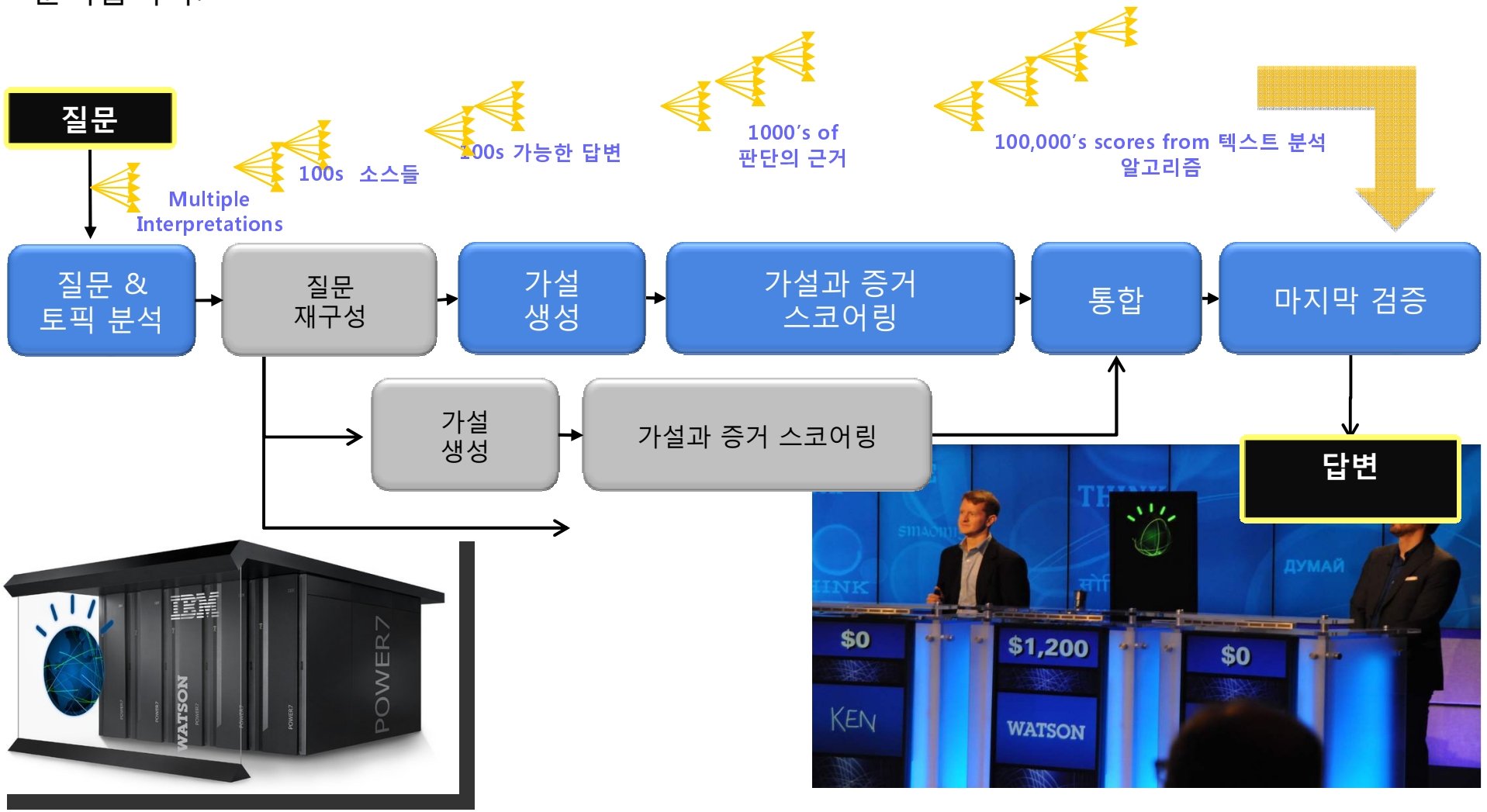
무엇을 확인할 수
있는지 탐구

- 브랜드 정서
- 제품 전략
- 자원의 최대 활용

# Scenarios

# 빅 데이터 분석을 활용한 Wastson

Watson의 고급 분석 기술 능력은 3초 내에 답을 알아내기 위해서 2억 페이지의 데이터를 분석합니다.



# Petabytes의 정보를 활용하여 투자 최적화


## *3주 -> 3일*

- **수 백개의 변수를 가진 1x1 킬로미터 그리드** 내에서 최적의 터빈 위치를 찾기 위한 날씨 모델링

## *운영 최적화*

- 운영중인 터빈 장비에서 센서 데이터를 수집하여 실제 결과를 이해
- **운영 기간, 서비스 중지 시까지의 시간, 바람과 터빈 사이의 상호 작용 최적화**

## ***Vestas HPC Cluster***

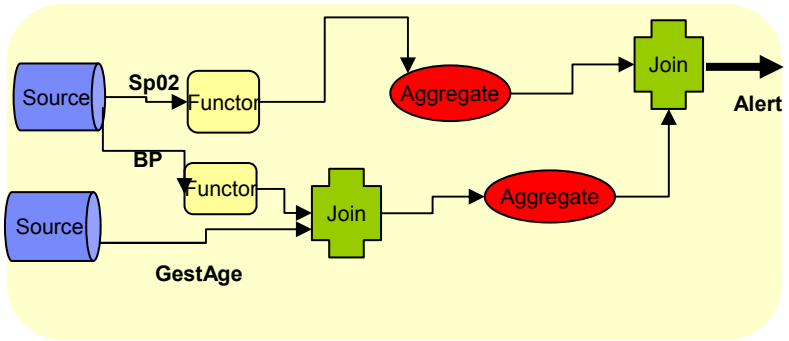
- 1,222 iDataplex Servers
- #53 on Top 500 Supercomputer list
- grow to 20+ PB over 4 years

- Leveraging JAQL, open source, column store, Text Analytics from BigInsights

# 조산아들의 위험 감지를 위한 빅 데이터 분석

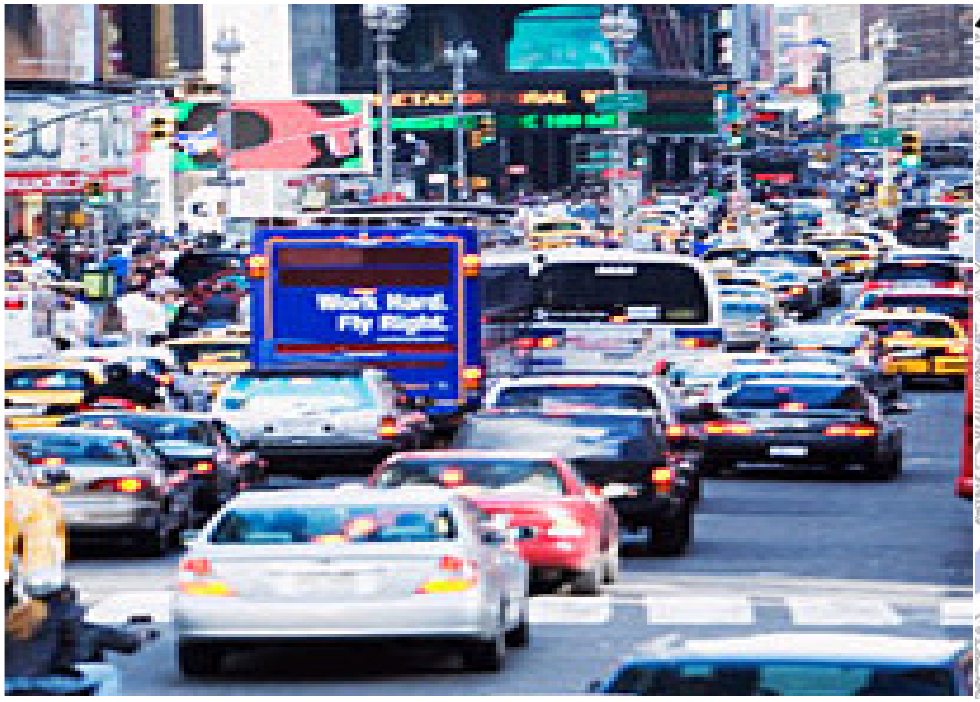

## 조산아들의 상태 모니터링

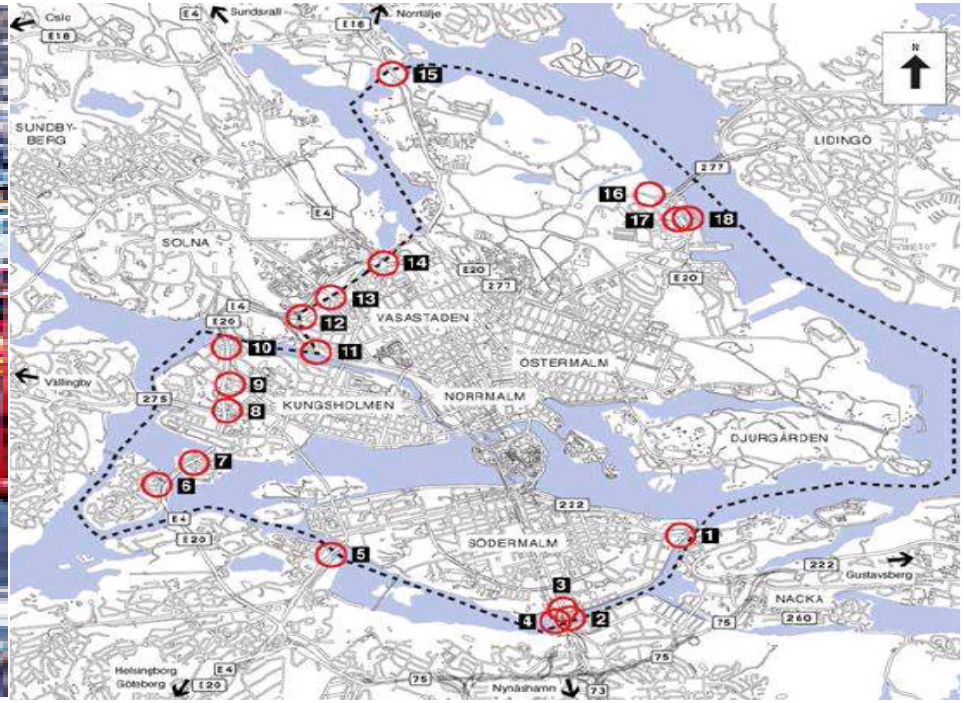
- 조산아의 위험상태를 check하기 위해 평균 BP와 Sp02간의 연관 관계를 활용
  - **Sp02 < 85%**
  - **BP < GestAge for 20sec**




❖ Sp02 = 산소 포화도, BP - 혈압

# 교통 트래픽 관리를 위한 빅 데이터 분석





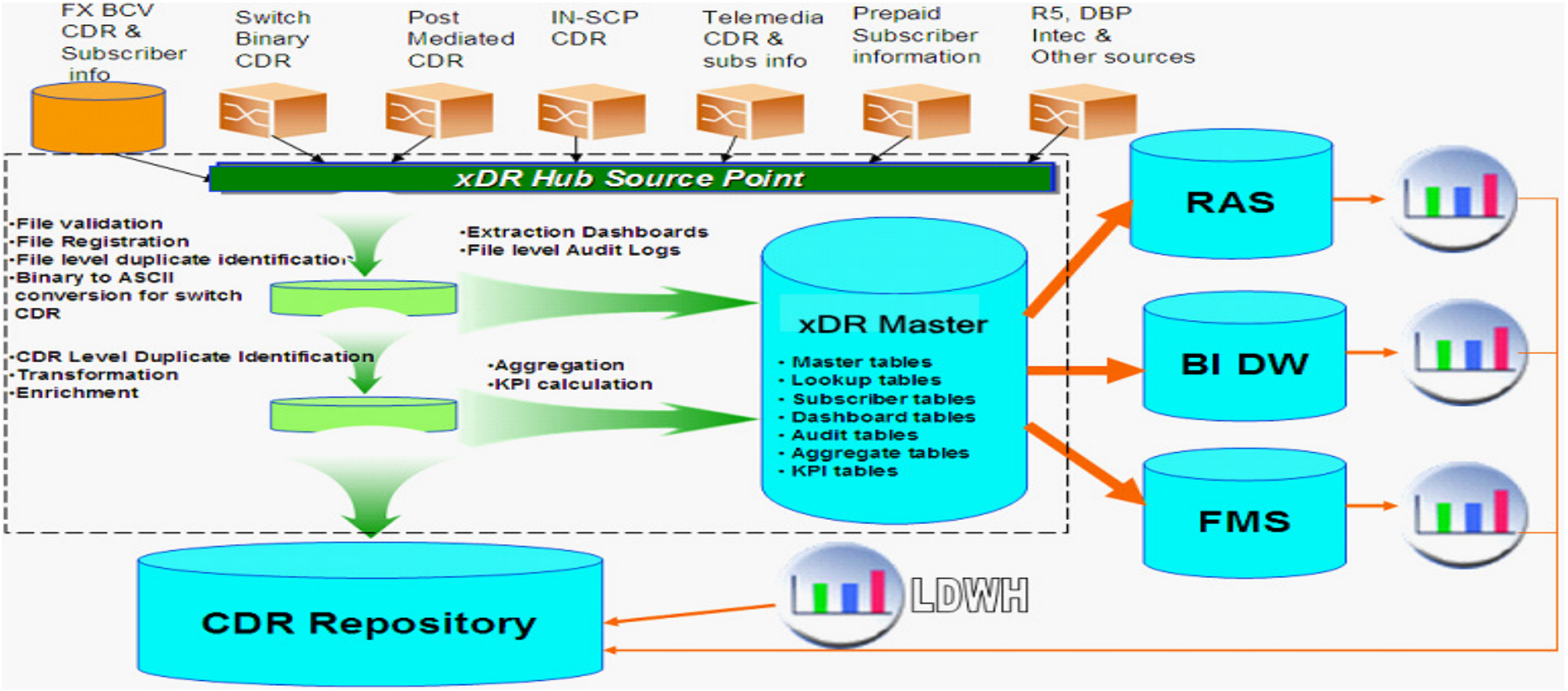

## *다양한 모델의 데이터 스트림*

- GPS
- Cell-Phone
- 공공 운송수단
- 운행 시간
- Induction loop detector data
- 사고
- +++

## *결과*

- 25%의 트래픽 감소
- 수송 고객이 40,000명 이상 증가
- 공공 운송 수단이 보다 효율적으로 운영됨
- 10~20%의 택시 수익 증가
- 40%의 배기 가스 감소

# 실시간 xDR 분석을 위한 빅 데이터 분석




- 실시간 요금 청구를 위해서는 수 billions CDR 데이터에 대한 조정 작업이 필요함
- 이전에 웨어하우스 내에서 중복 제거 작업을 하기 위해서는 12시간이 소요되었으나 1초 만에 완료됨

- 6 billions CDRs / 일
- 웨어하우스의 성능이 개선됨
- IT의 복잡성이 감소됨

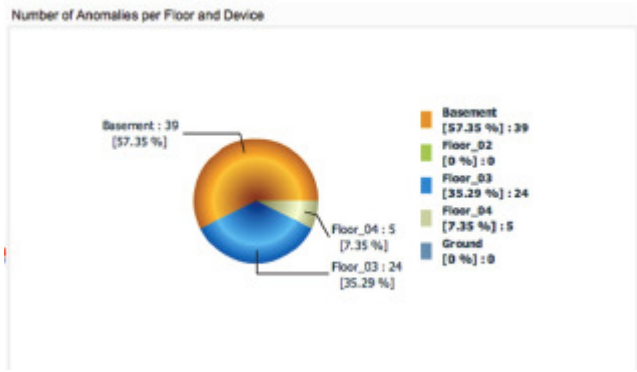
# 스마트 인프라 관리를 위한 빅 데이터 분석

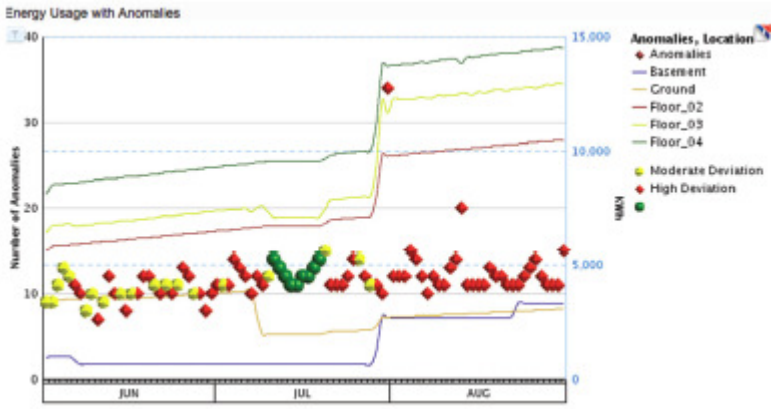
## 시나리오

- 차세대 서비스 플랫폼 구축
- IIMS (지능적인 인프라 관리 서비스)
- IMMS (지능적인 유지 관리 서비스)

## 요구사항

- 빌딩 내의 에너지 관리
- 사용 양과 이력 분석에 의한 장비 관리 및 관리 예측
- 데이터 센터내에 IP로 관리 가능한 어플리케이션들의 모니터링 및 예측 분석
- 1 TB / 일, 50k / 초 로그
- 다양한 데이터 소스 – Text, XML, no schema



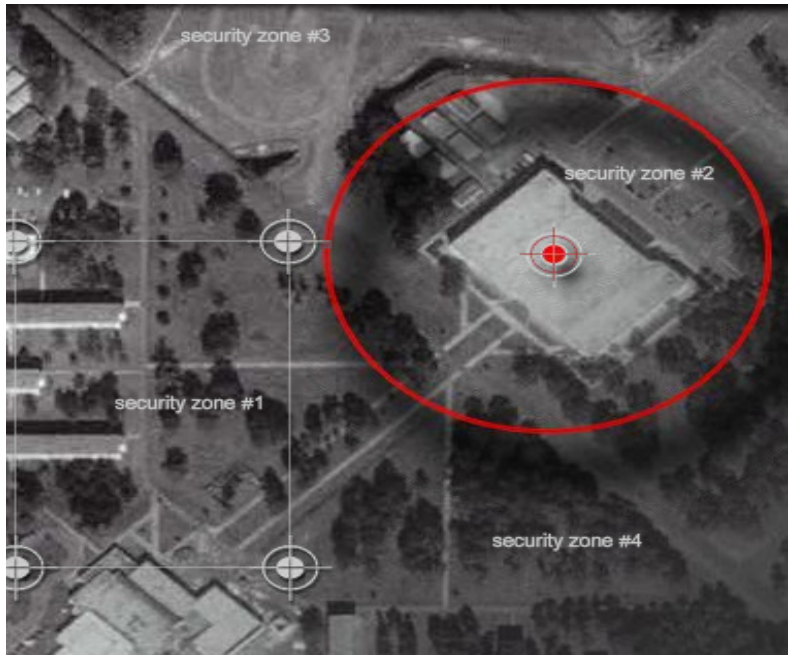
# 시설 감시와 보안을 위한 빅 데이터 분석

## 시나리오

- 스트림 플랫폼에 기반하여 최첨단 기술을 활용한 감시 시스템
- 매설된 광 케이블에서 나오는 음향 시그널을  실 시간으로 모니터링, 분석하고 보고함
- 바이너리 데이터에 대하여 1600개의 스트림까지 확장 가능하도록 디자인됨

## 요구사항

- 다양한 모델의 시그널 (음향, 비디오, 등등)에 대한 실 시간 처리가 필요
- 확장이 쉬어야 하며 동적으로 진행되어야 함
- 3.5M data / 초



# And more

로그 분석 및 보관

Smart Grid / Smarter Utilities

RFID 트랙킹 & 분석

사기 방지 / 리스크 관리 & 모델링

고객에 대한 360° 뷰

Warehouse의 확장

Email / 콜 세터 트랜스크립트 분석

Call Detail Record 분석

+++

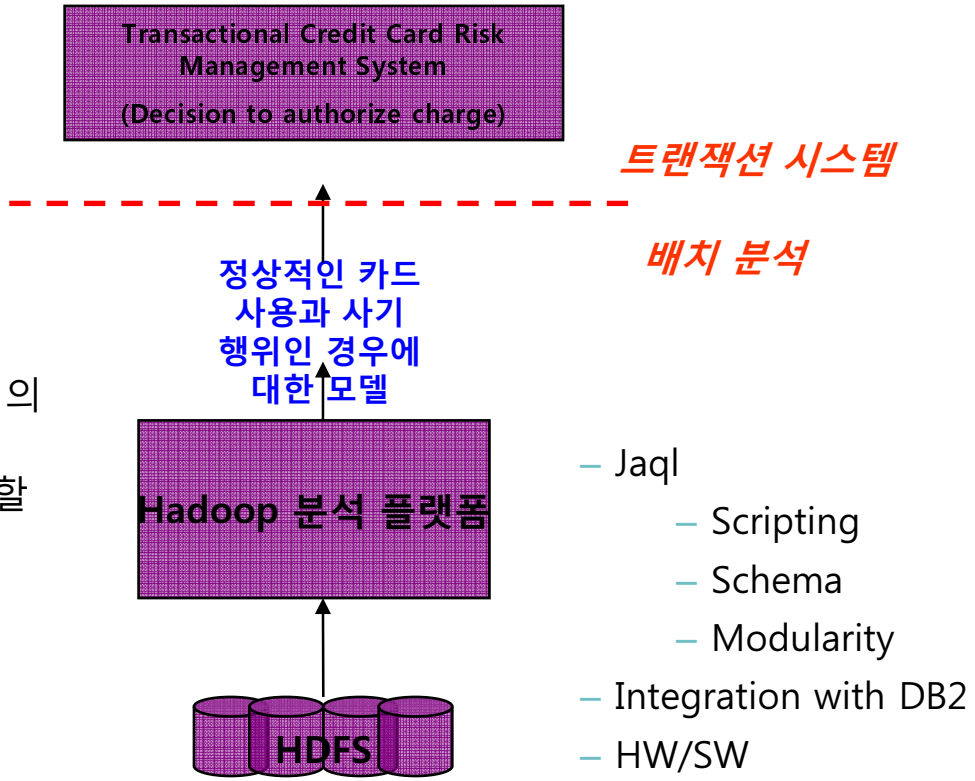
# 향상된 사기 방지 시스템

## Customer Applications

- 신용 카드 리스크 관리
  사기 행위에 대한 모델 생성
- 타겟 마케팅

## User Scenario

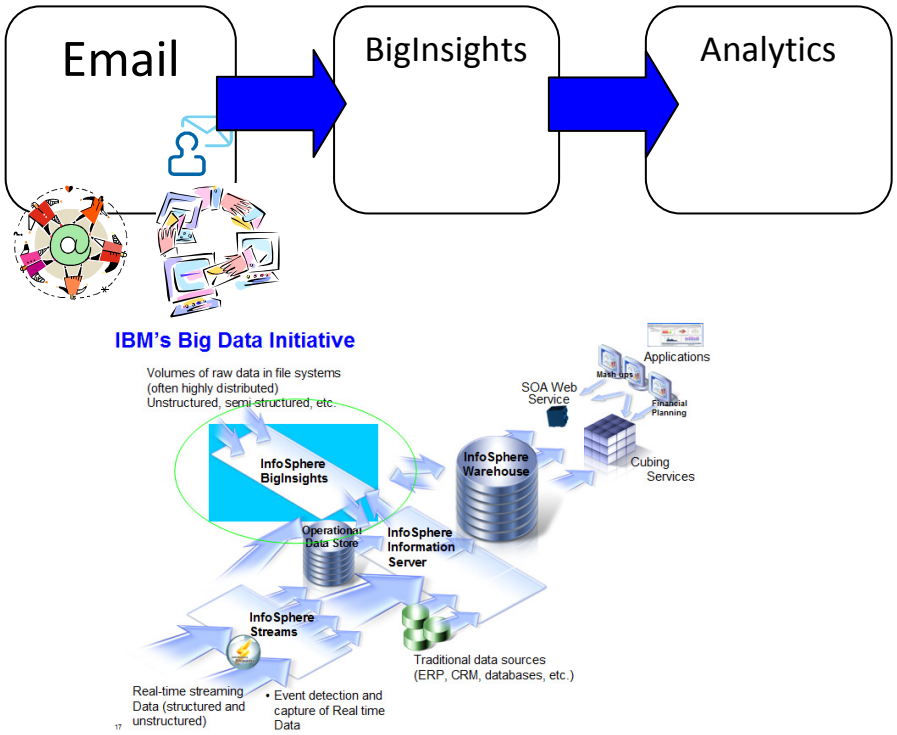
- 2억 트랜잭션 / 일
- 신용 카드 승인 요청 시에 트랜잭션 시스템의 정보로 활용될 수 있는 모델 필요
- 트랜잭션의 패턴을 분석하여 사기 방지를 할 수 있기를 원함
- 분석은 자원을 최대한 활용
- 수 년간의 데이터를 활용해야 함
- 새로운 사기 패턴을 찾는데 도움을 줄 수 있어야 함

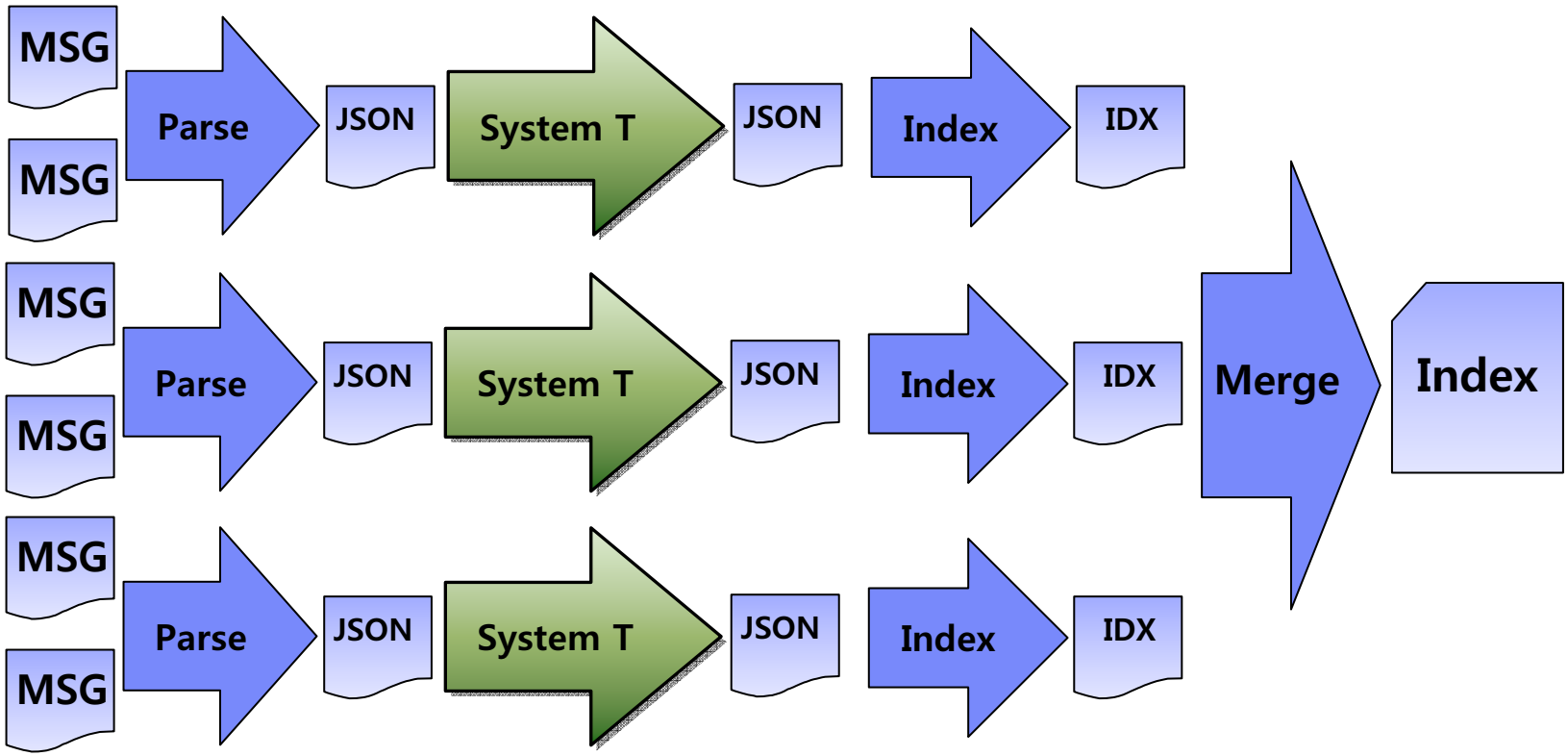


# E-mail 분석

**대용량의 e-mail 아카이브를 보유하고 있는 고객이 보다 효율적이면서 유연한 저장 장치와 검색 및 조회를 원함**

- 기존에 100TB의 아카이브 존재
- 매년 2TB씩 증가
- 추가적으로 매년 6TB를 더 수집하여야 함
- 아카이브 영역은 일/주/월 단위로 갱신 작업등이 가능하여야 하며 조회 작업을 할 수 있어야 함
- E-mail에 대하여 조회, 탐색 및 패키징이 가능하여야 함
- 헤더, 본문, 첨부 영역까지 단순 조회에서 복잡한 형태의 조회 기능도 요구됨






***IBM Research의 기술력은  이와 같은 작업들이 가능하게  합니다***

- Text Analytics (systemT)
- JAQL
- BigIndex

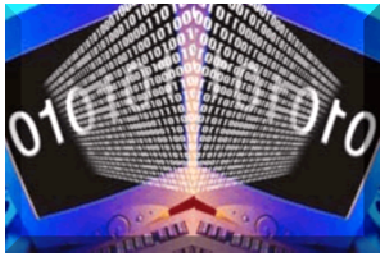
# 빅 데이터를 활용한 분석 영역은 무한합니다.

## Smarter Healthcare

## Multi-channel

## Finance

## Log Analysis



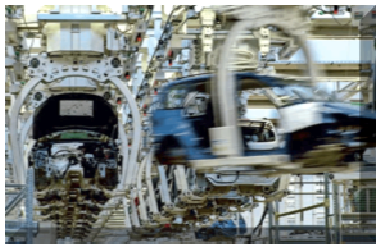

## Homeland Security


## Traffic Control

## Telecom

## Search Quality


## Manufacturing



## Trading Analytics

## Fraud and Risk

## Retail: Churn, NBO

# Technology

# IBM 빅 데이터 플랫폼

**InfoSphere BigInsights**
비정형 데이터를 포함하는 하둡 기반하의 분석

**Hadoop**

**InfoSphere Information Server**
대용량 데이터 통합 및 변형

**Information Integration**

**Stream Computing**

**InfoSphere Streams**
스트리밍 데이터에 대한 분석

**MPP Data Warehouse**

**IBM InfoSphere Warehouse**
대용량의 구조화된 데이터 분석

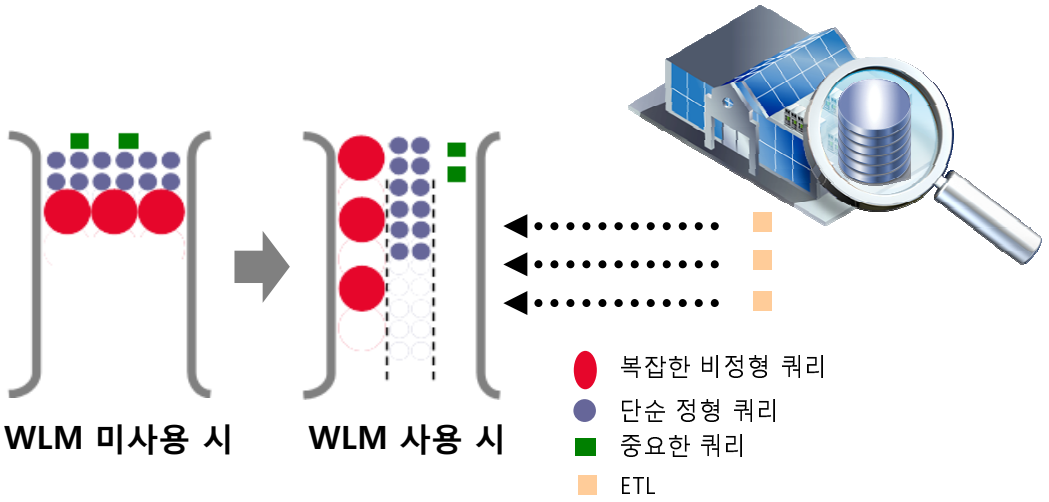
**IBM Netezza High Capacity Appliance**
조회용의 아카이브 형태의 구조화된 데이터

**IBM Netezza 1000**
BI+Ad Hoc
정형 데이터에 대한 분석

**IBM Smart Analytics System**
정형 데이터에 대한 운영 분석

**IBM Informix Timeseries**
시계열 데이터 분석

# Operational 분석 환경을 위한 ISAS (Smart Analytics System)

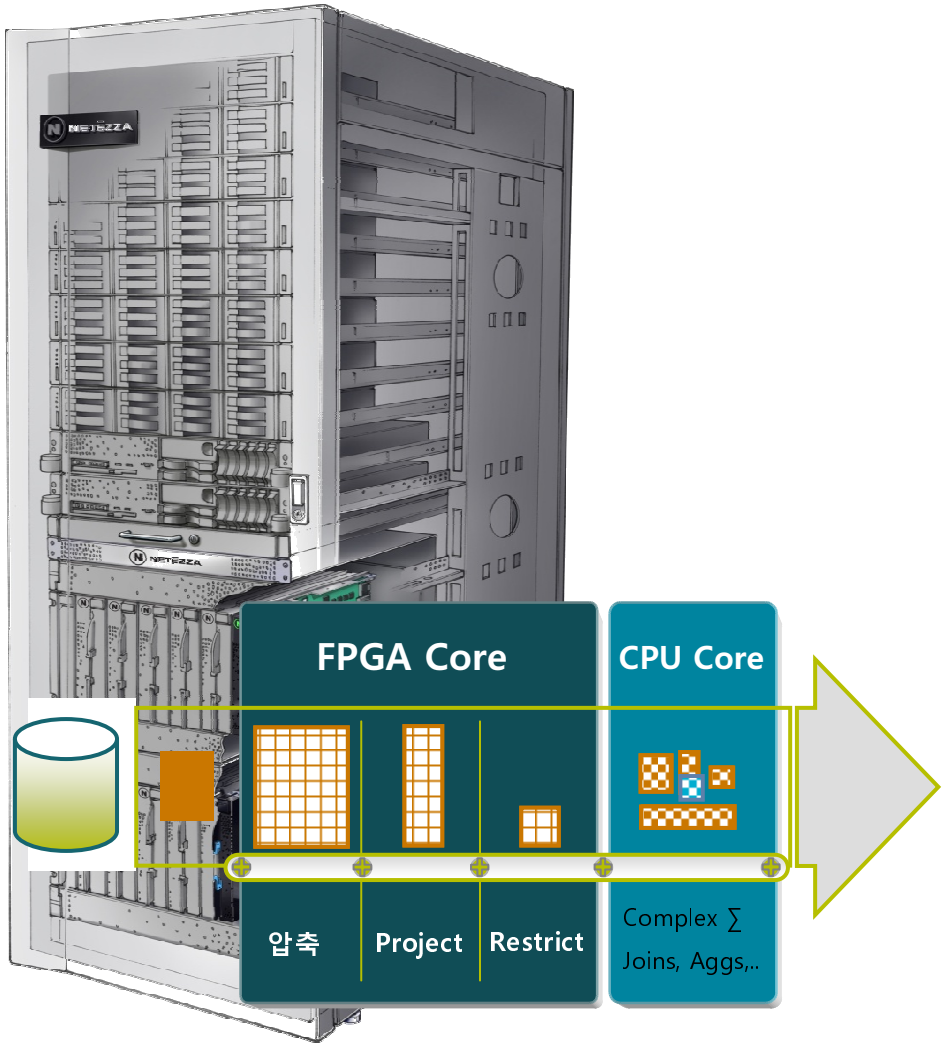

**Mixed Workload**

실시간, 조회, 적재, 혼합 워크로드와 다수의 사용자 환경에 적합

**Workload Management**

뛰어난 워크로드 관리 기능을 통해 시스템 자원에 대한 통제 기능 제공

# 대용량 정형 데이터 분석 전용 솔루션인 Netezza




## Speed

DW에 특화된 True Appliance

## Simple

DBA 작업을 최소화

## Smart

업계 unique한 Data Streaming 기술을 통해 쿼리와 상관 없이 고성능 분석 지원

## Scalability

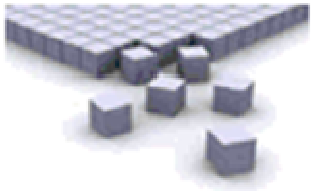
MPP 구조의 확장성을 가진 업계 unique한 Data Streaming 기술을 적용한 FPGA와

# 대용량의 정형/비정형 데이터 분석을 위한 InfoSphere BigInsights

## Platform for V3

Enhanced Hadoop foundation

Storage, 보안, 클러스터 관리



Semi-structured data

## Usability

웹 콘솔

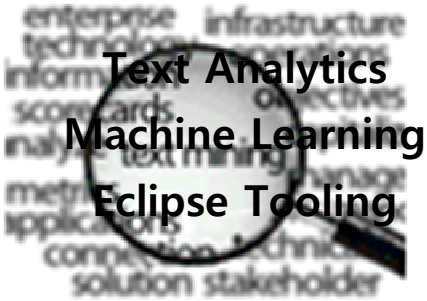
통합 설치

Big Sheet

Ready-made “app”

InfoSphere BigInsights

## Analytics for V3



Text Analytics

Machine Learning

Eclipse Tooling

## Performance

Adaptive MapReduce

FAIR 잡 스케즐러

압축 기술

빅 인덱스

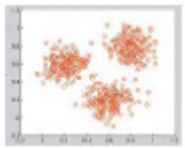
## Integration

Connectivity to DB2, Netezza, JDBC 지원 상용 DBMS

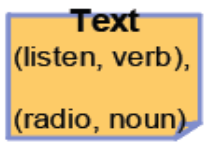
Structured data

# In-motion 분석 기능을 제공하는 InfoSphere Streams

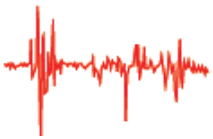
데이터 마이닝
(In Streams)



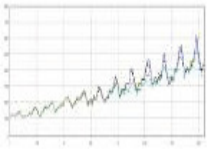
텍스트 분석
(In Streams)




음향 분석
(Research)



예측 분석
(Research)

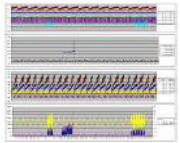


통계 분석
(In Streams)

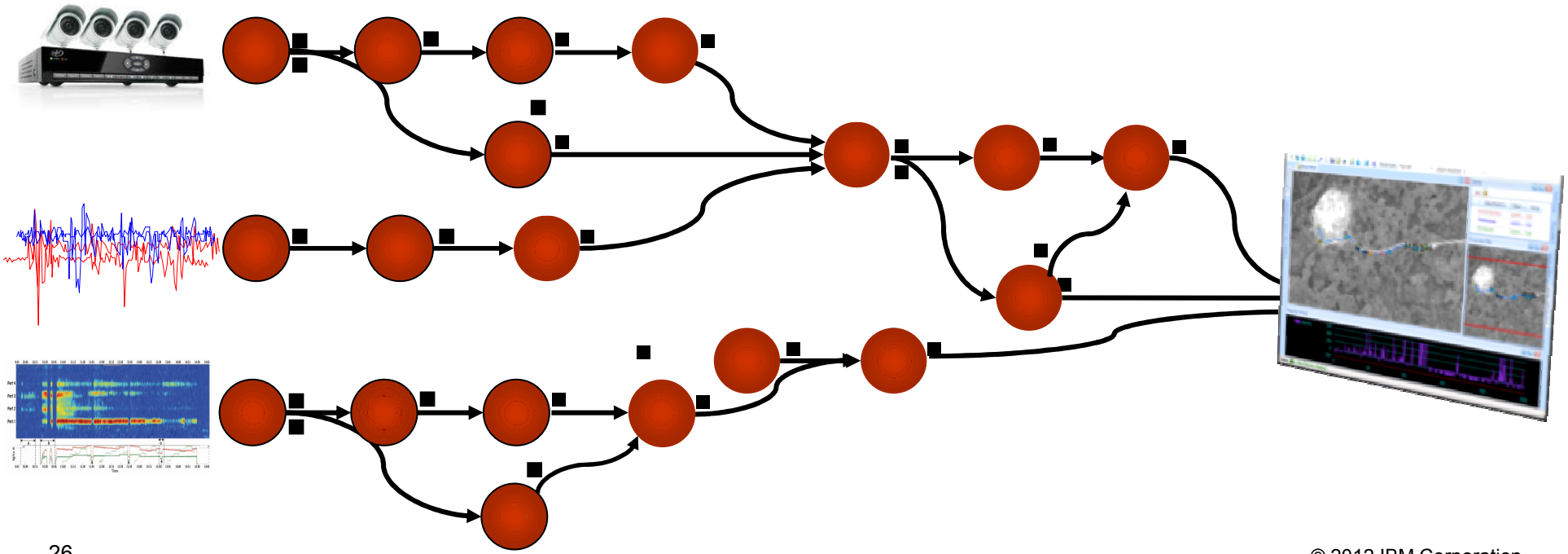
$$\sum_{population} R(s_t, a_t)$$

지리 데이터
(Research)

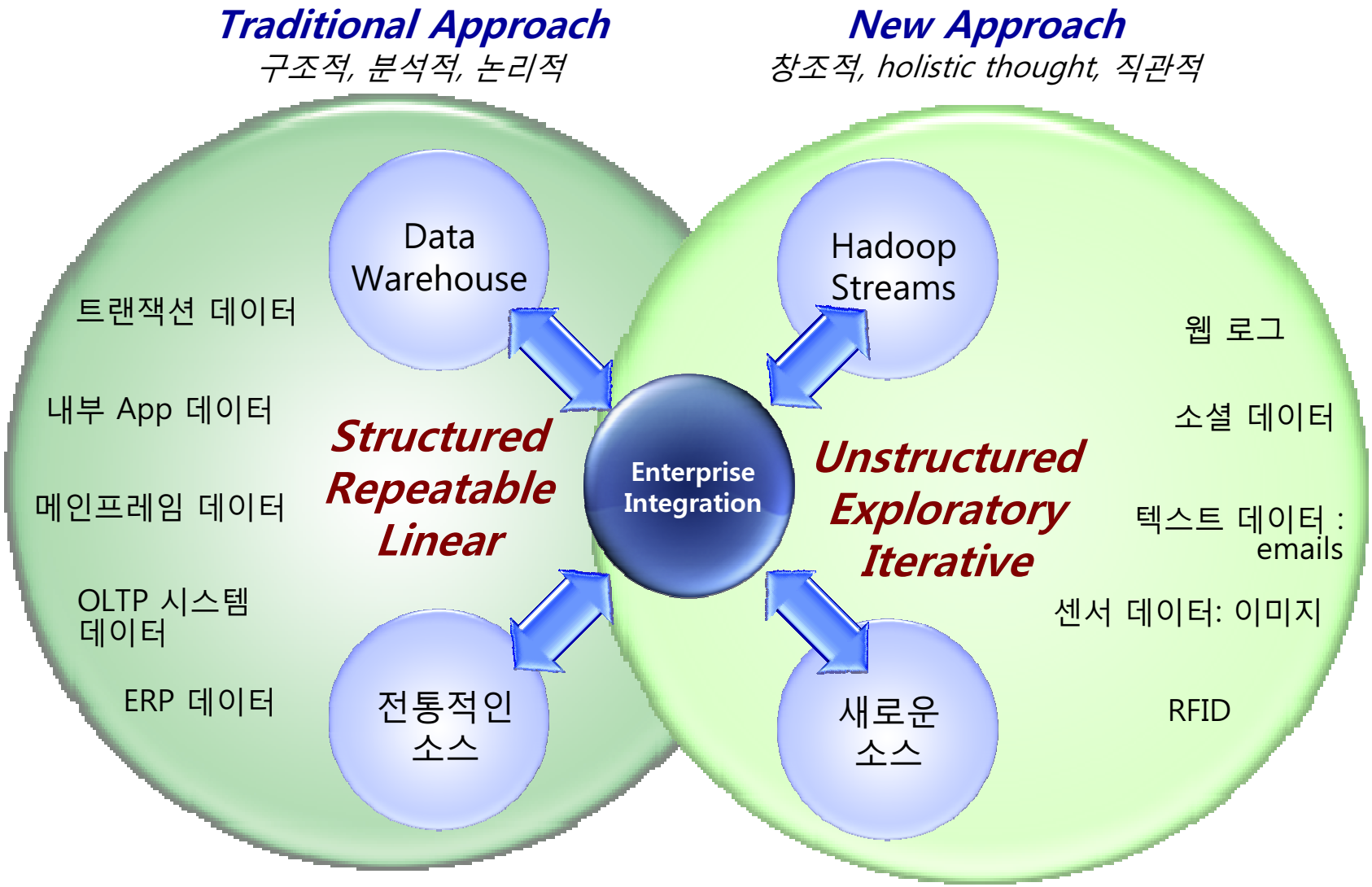
고급 산술 모델
(Research)



이미지, 비디오
(Open.

# 빅 데이터를 분석하기 위해서는 분석 플랫폼간의 연계가 중요합니다.



# 감사합니다.