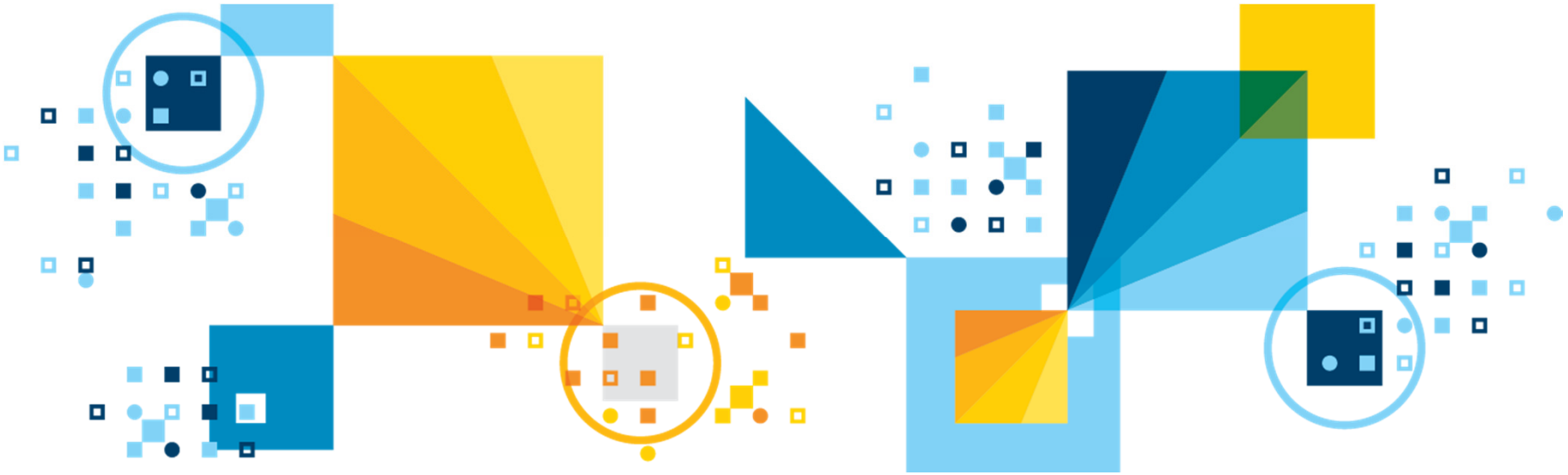
하둡, 비즈니스 기업을 위한 빅데이터 플랫폼으로 거듭나기

이정권 실장
한국IBM



© 2014 IBM Corporation

모든 산업군에서 Big Data와 Analytics를 활용할 수 있습니다.

Banking
- Optimizing Offers and Cross-sell
- Customer Service and Call Center Efficiency
- Fraud Detection & Investigation
- Credit & Counterparty Risk

Insurance
- 360˚ View of Domain or Subject
- Catastrophe Modeling
- Fraud & Abuse
- Producer Performance Analytics
- Analytics Sandbox

Telco
- Pro-active Call Center
- Network Analytics
- Location Based Services

Energy & Utilities
- Smart Meter Analytics
- Distribution Load Forecasting/Scheduling
- Condition Based Maintenance
- Create & Target Customer Offerings

Media & Entertainment
- Business process transformation
- Audience & Marketing Optimization
- Multi-Channel Enablement
- Digital commerce optimization

Retail
- Actionable Customer Insight
- Merchandise Optimization
- Dynamic Pricing

Travel & Transport
- Customer Analytics & Loyalty Marketing
- Predictive Maintenance Analytics
- Capacity & Pricing Optimization

Consumer Products
- Shelf Availability
- Promotional Spend Optimization
- Merchandising Compliance
- Promotion Exceptions & Alerts

Government
- Civilian Services
- Defense & Intelligence
- Tax & Treasury Services

Healthcare
- Measure & Act on Population Health Outcomes
- Engage Consumers in their Healthcare

Automotive
- Advanced Condition Monitoring
- Data Warehouse Optimization
- Actionable Customer Intelligence

Chemical & Petroleum
- Operational Surveillance, Analysis & Optimization
- Data Warehouse Consolidation, Integration & Augmentation
- Big Data Exploration for Interdisciplinary Collaboration

Aerospace & Defense
- Uniform Information Access Platform
- Data Warehouse Optimization
- Airliner Certification Platform
- Advanced Condition Monitoring (ACM)

Electronics
- Customer/ Channel Analytics
- Advanced Condition Monitoring

Life Sciences
- Increase visibility into drug safety and effectiveness

© 2014 IBM Corporation

Big Data 와 Analytics이 패러다임의 변화를 가져옵니다.

전통적인 분석 방식

구조적 & 반복적
Structure built to store data

자원 한계로 인해 데이터 sampling 활용

분석 전에 데이터에 대한 정재

빅 데이터 분석 방식

Iterative & Exploratory
Data is the structure

가용한 많은 데이터를 활용

Raw 상태의 정보 & 필요에 따라 정제

© 2014 IBM Corporation

Hadoop이 점점 진화하고 있습니다.

1. 화두가 아닌 현실 *The Hype is Over*
2. 더욱 큰 규모로 성장하고 있는 빅데이터 *Big Data Grows Bigger*
3. 새로 등장하는 분석 앱 *Support for New Analytic Apps*
4. **하둡의 진화 *Hadoop Matures***
5. 보안/개인정보보호 *Security/Privacy*
6. 인지컴퓨팅으로의 발전 *Designing for Cognitive Computing*
7. 마케팅에 효과를 드러내는 Big Data *Big Data in Marketing*
8. 최고 데이터 경영자 *Chief Data Officer*
9. 데이터 사이언티스트 *Data Scientists*
10. 데이터 품질 *Data Quality*

Source : Horizon Watch 2014 / IBM Software Strategy and Directions

© 2014 IBM Corporation

5가지 주요 Use Cases

Big Data 탐구

비즈니스 지식을 확장하기 위한 탐색, 시각화, 이해

고객에 대한 360° 확장된 뷰

내부 데이터와 외부 소스를 연동하여 폭 넓은 고객에 대한 이해

Security/Intelligence Extension

실시간으로 사이버 보안을 모니터하여 위험 감소와 사기 감지에 활용

Operations Analysis

비즈니스 결과를 극대화하기 위해 다영한 장비 데이터에 대한 분석

Data Warehouse Augmentation

운영의 효율성을 높이기 위해 빅 데이터와 데이터 웨어하우스 역량을 통합

© 2014 IBM Corporation

AS-IS 전사 데이터 분석 환경

Data Sources

Information & Insight

+ In-database transformations (ELT faster than ETL)
+ Provides some structure, enabling queries
- Adds significant cost and overhead to EDW

Structured Operational

Staging Area

Expanded EDW

Marts

Archive

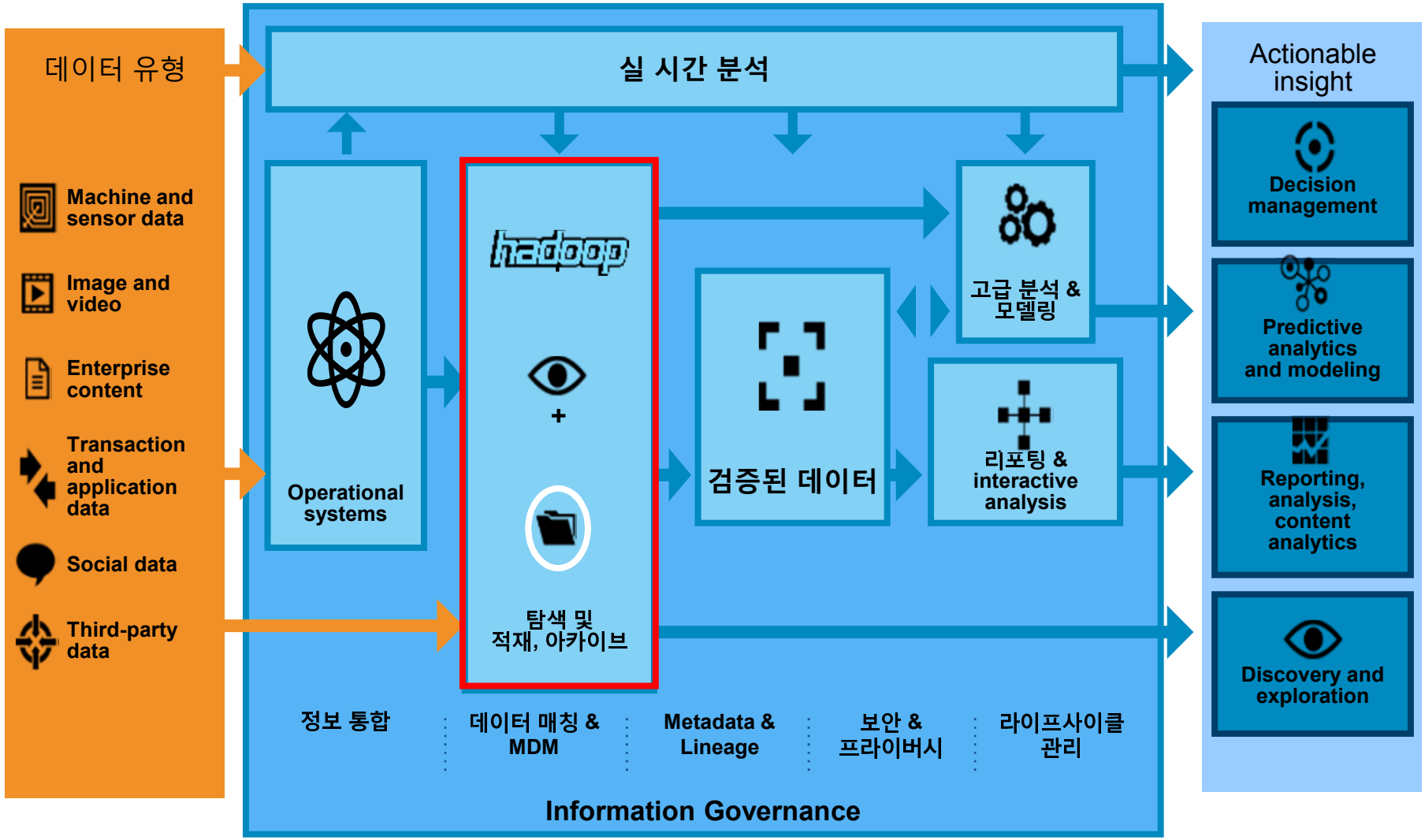
Information Movement & Transformation

Predictive Analytics & Modeling

BI & Performance Management

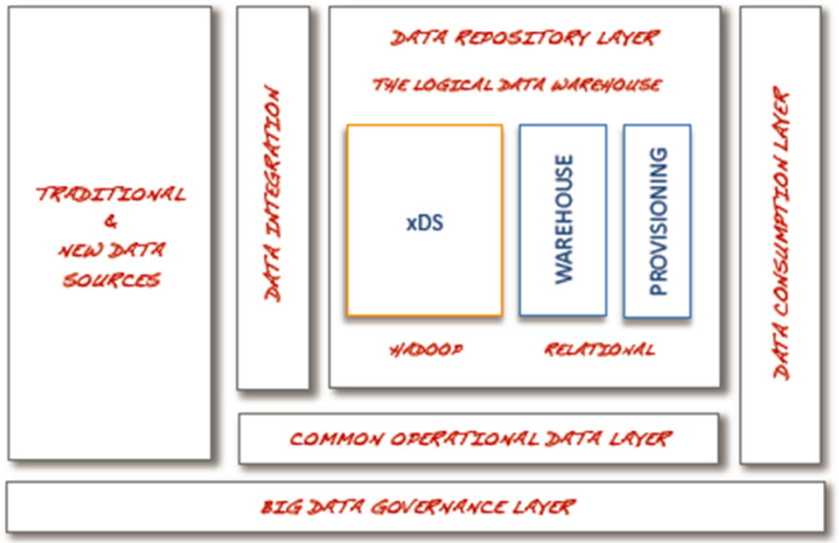
© 2014 IBM Corporation

Next Generation Enterprise Warehouse



© 2014 IBM Corporation

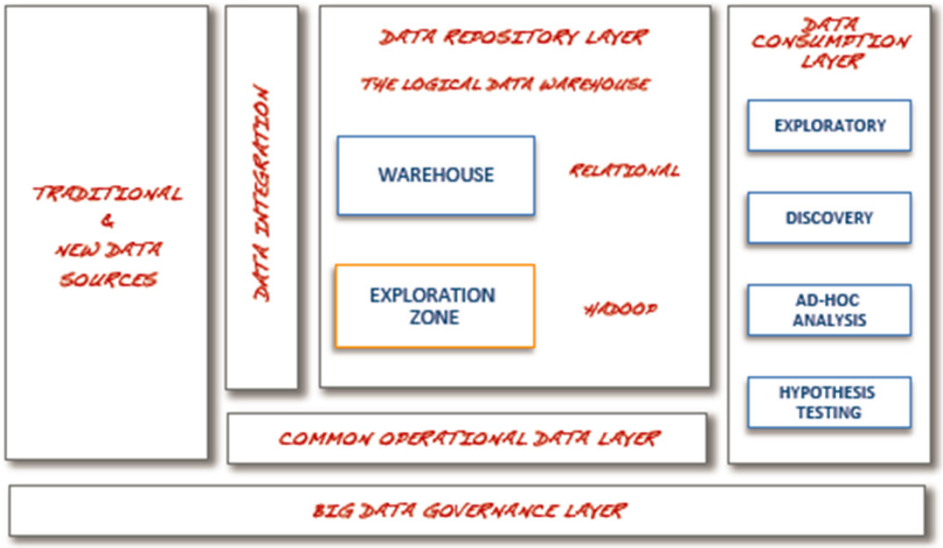
The xDS



- Notion of the Information Superset
- Raw 형태로 모든 데이터 저장
- xDS (SOR)
  - Raw Data 영역
  - Universal Data 영역
  - Big Data 영역
  - 공유 데이터 영역
  - Global Data Store
- Pre-processing 허브
  - Landing and Staging
  - HDFS
- Universal ODS
  - 정형 & 비정형
  - HBASE/HIVE
- 추출된 데이터를 전송
  - ODS/DWH/CMS

© 2014 IBM Corporation

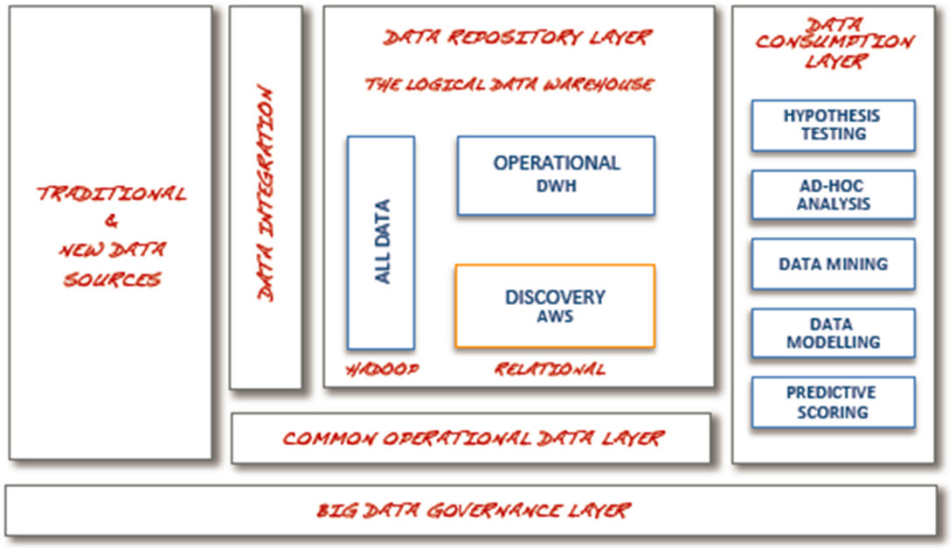
The Exploration Zone



- Data Scientist's sand-pit
- Mine for value
- 연관 관계와 컨텍스트 탐색
- 웨어하우스 데이터와 연계
  - Connectors
  - JAQL / UDF
- Analytic freedom
  - Data untouched
  - Schema-less / Schema on run
  - Fail-fast, Fail-smart
- Mined value
  - Propagate to warehouse
  - Operationalise

© 2014 IBM Corporation

Discovery & Operational Analytics

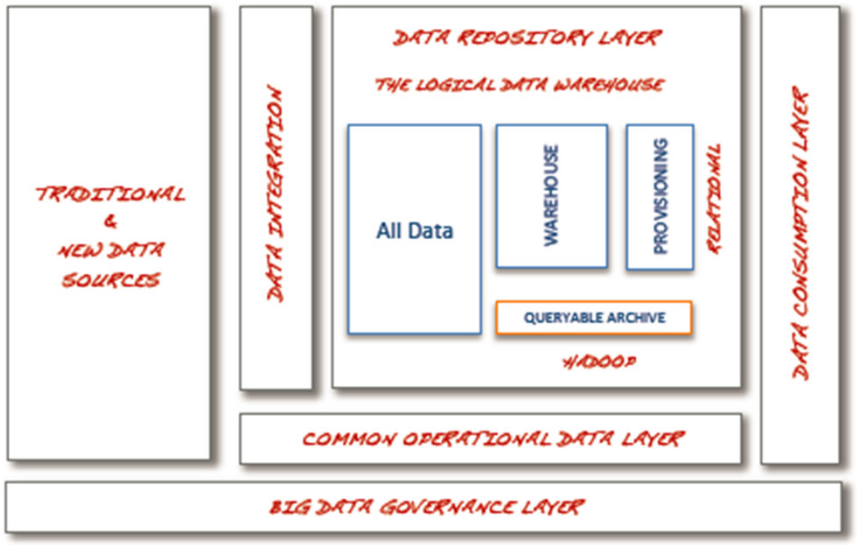


- Analyst & Data Miner's workspace
- 관계형 분석 엔진 (appliance)
- Discovery environment
  - 데이터 마이닝 및 모델링
  - 가설 테스트
  - Self-service
  - More control of structures and data
  - Ability to incorporate additional data
  - 모델 생성 e.g. scoring
  - Propagate warehouse data
  - Run the model in operational environment

- Operational environment
  - Managed, governed, secured
  - Repeatable / enterprise use

© 2014 IBM Corporation

The Queryable Archive



- Data always available
- Create active and online archive
- 비용 효율적인 compute and storage
  - Hadoop framework
  - Commodity server and storage
- Move data from DWH to Hadoop
  - ETL or managed archive load
  - HIVE data-warehouse storage
- BI and Reporting from Hadoop
  - HDFS, HBASE, HIVE
  - BigSQL and ODBC
- Federated query
  - DWH & Hadoop (archive)
  - Current and historical

© 2014 IBM Corporation

Automobile and Manufacturing Quality Control and Customer Satisfaction

기존 IT 솔루션의 유연성과 확장성에 제약 사항이 많음

A new solution is needed **to improve customer insights, quality and operational efficiency**

- 부품 제고 제어
- 장비 제조와 조립 라인의 데이터
- 딜러의 보증 및 서비스 데이터
- 차량의 Telemetry 데이터
- 고객 서비스와 소셜 미디어 데이터

차 세대 엔터프라이즈 데이터 웨어하우스 :

- 5~10년간의 데이터를 저장 및 분석하기 위한 영역 : Data landing zone and analytic zone
- 고 성능의 리포트를 제공하기 위한 영역 : Warehouse reporting zone

IBM Corporation

Vestas optimizes capital investments based on **2.5 Petabytes** of information.

활용된 기술 :

InfoSphere BigInsights

InfoSphere Warehouse

- 전기 생성를 최대화하고 장기간 운용할 수 있도록 터빈의 위치를 최적화하기 위해 날씨 모델을 활용
- 터빈의 위치를 결정하는데 소요되는 시간을 수 주에서 수 시간으로 단축
- 2.5 PB의 정형 및 반 정형 데이터를 제어.
- 데이터 양이 6 PB까지 증가할 것으로 예상됨

Vestas

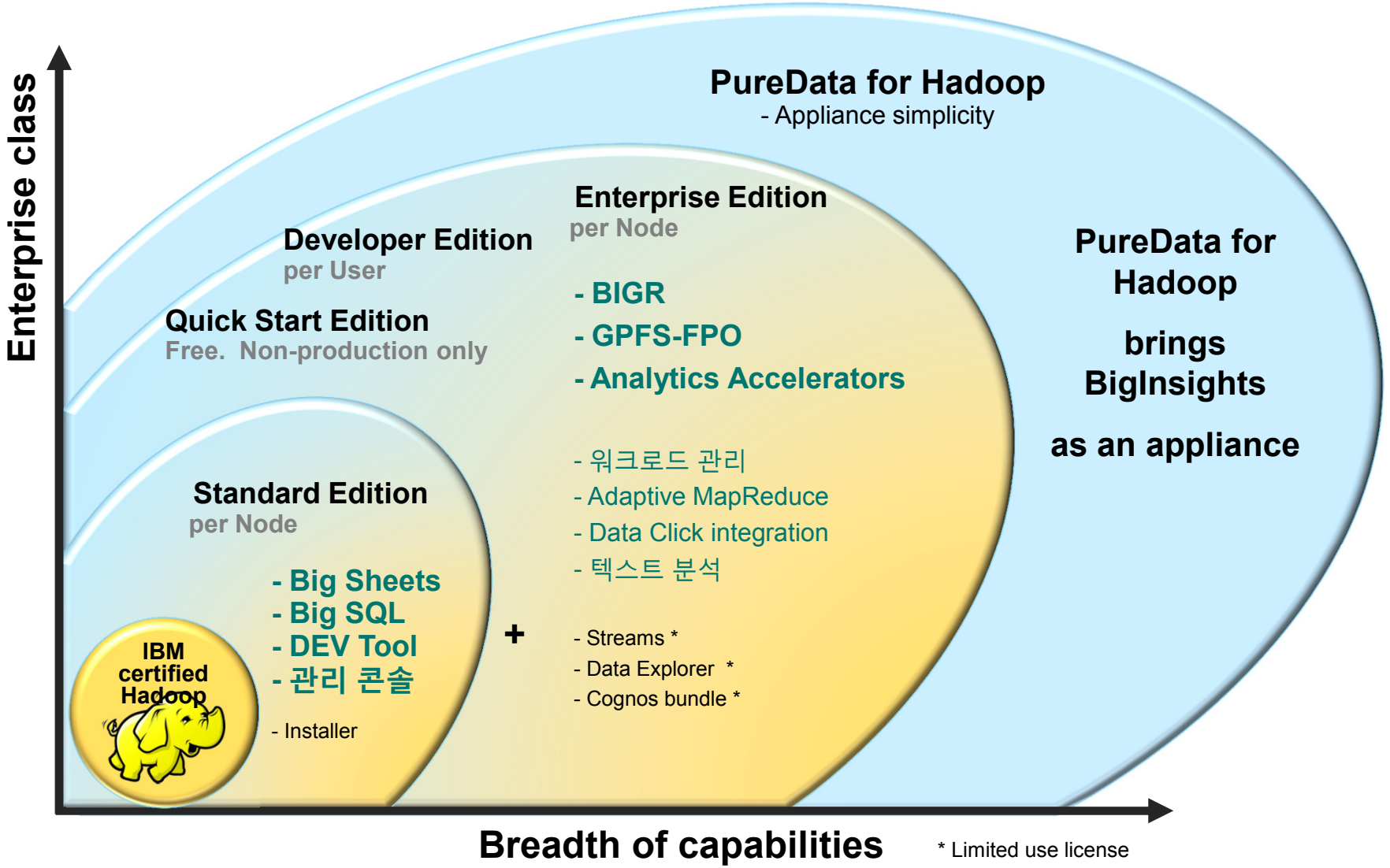
IBM은 Hadoop 기술을 보완하여 Enterprise에 적용할 수 있게 해 줍니다.

- **확장성**
  - 신규 노드 추가를 온라인 중에 수행
- **효율성**
  - 상용 서버 위에 대용량 병렬 컴퓨팅 구형
- **유연성**
  - Hadoop은 schema-less하여 모든 유형의 데이터를 처리할 수 있습니다.
- **장애 대응**
  - MapReduce software framework을 통해

IBM Innovation

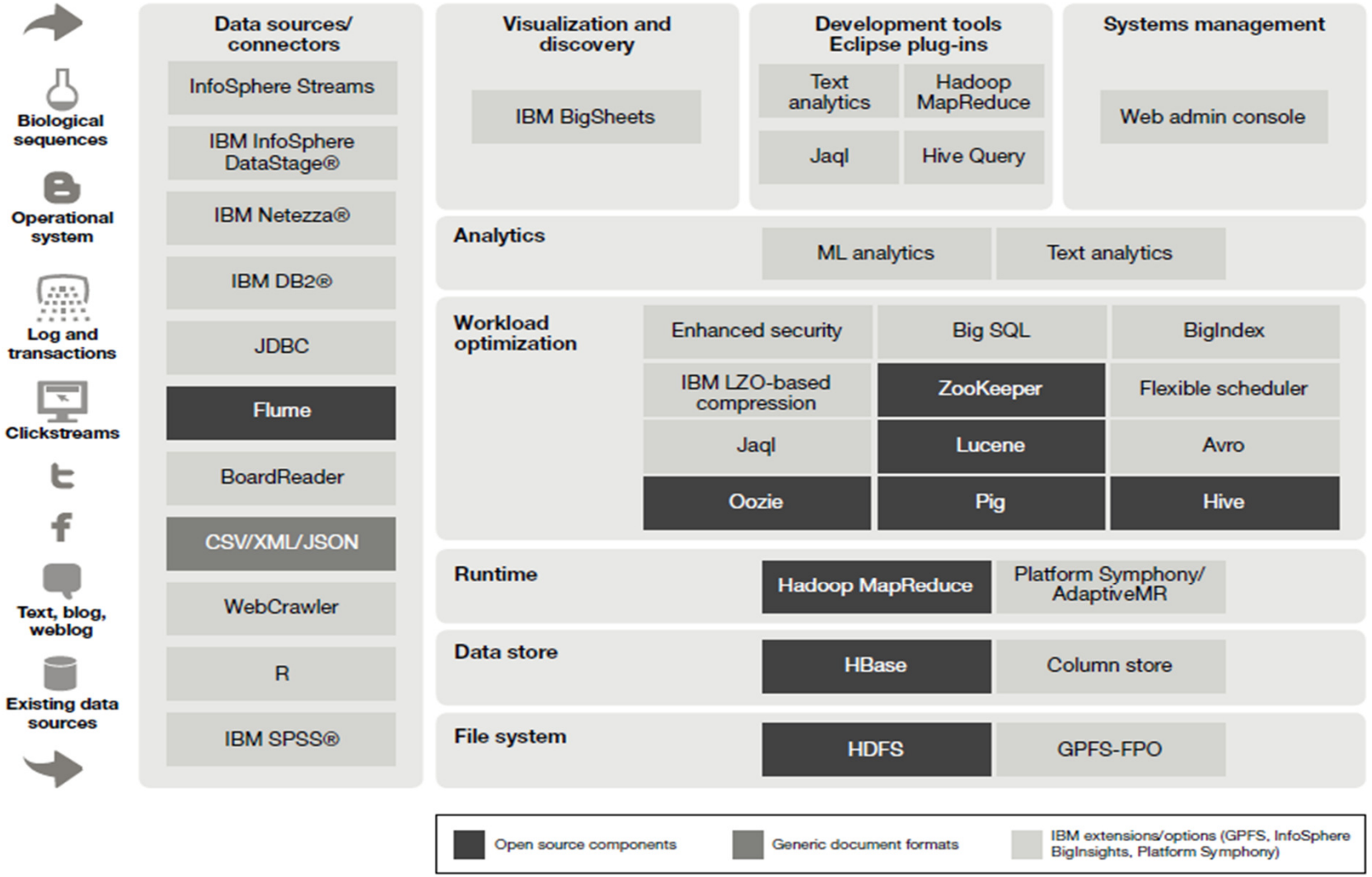
- **성능 & 안정성**
  - Adaptive MapReduce, Compression, Indexing, Flexible Scheduler, GPFS-FPO
- **분석 가속기**
- **생산성 향상**
  - BigSQL
  - Web-based UIs
  - Tools to leverage existing skills
  - End-user visualization
- **엔터프라이즈 통합**
  - To extend & enrich your information supply chain.

IBM은 다음과 같이 Hadoop Offering을 제공합니다.



© 2014 IBM Corporation

Solution Components



© 2014 IBM Corporation

BigSQL : SQL on Hadoop

- ANSI SQL 92+ 지원
- 범용 JDBC/ODBC 드라이버 제공
- 쿼리에 따른 성능 최적화
  - Heavy 쿼리에 대해서는 MR를 통한 병렬 처리 수행
  - 작고 가벼운 query에 대해서는 MR를 pass
- 다양한 데이터 소스 지원
  - HIVE, HBASE, CSV, JSON,…

생성

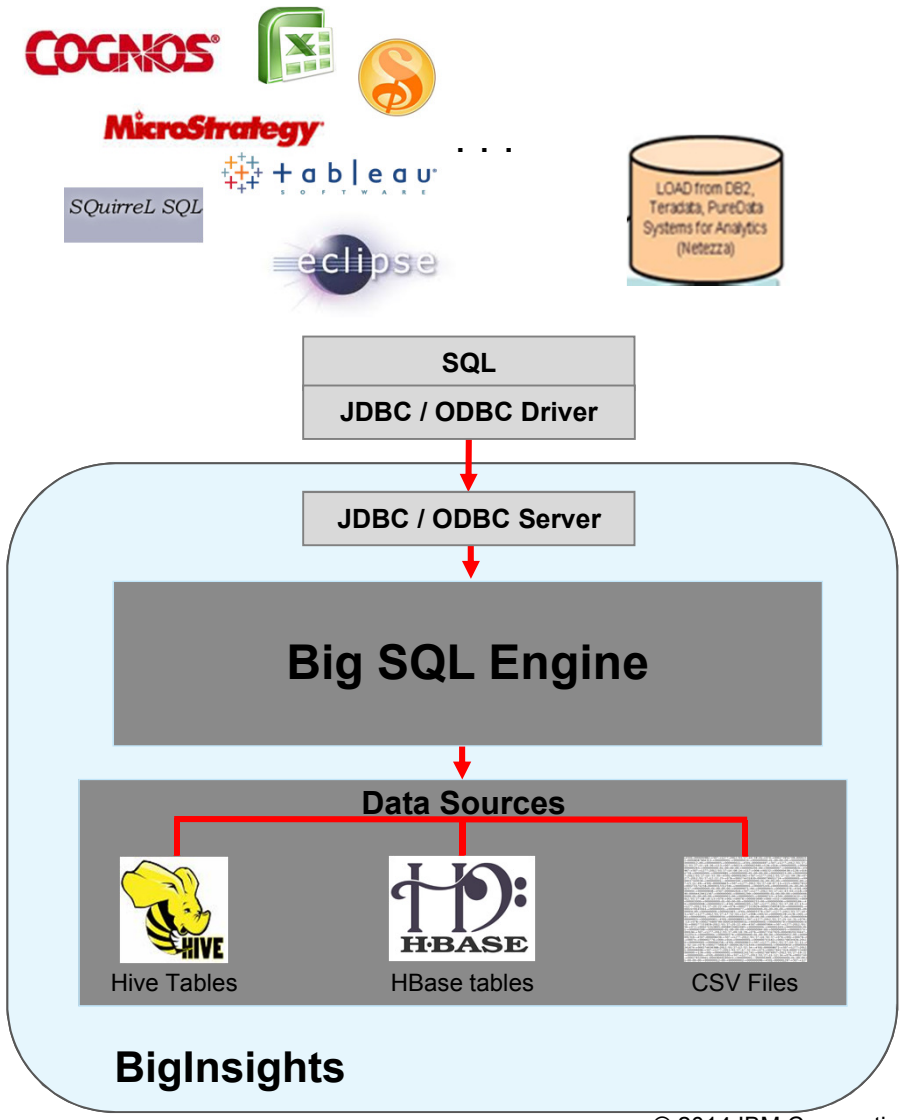
```
CREATE HBASE TABLE table-name
(column_name data_type, ... )
COLUMN MAPPING
KEY  MAPPED BY (keys) ENCODING BINARY,
cf:next_column_name MAPPED BY (next_column_name)
ENCODING BINARY,
     ... )
```

적재

```
LOAD USING FILE URL
"sftp://biadmin:password@myserver.abc.com:22/home/biadmi
n/mydir/staff/"
 INTO HBASE TABLE STAFF APPEND
WITH TARGET TABLE PROPERTIES (hbase.load.method =
"put", hbase.disable.wal = true)
```

조회

```
select e.lname, e.fname   from employees e
where e.salary > 30000
order by e.lname
```



© 2014 IBM Corporation

BigSheet : Spreadsheet 분석

- Spreadsheet 스타일의 환경을 제공하여 보다 손쉽게 데이터의 조작함은 물론 시각화를 할 수 있는 기능을 제공합니다.

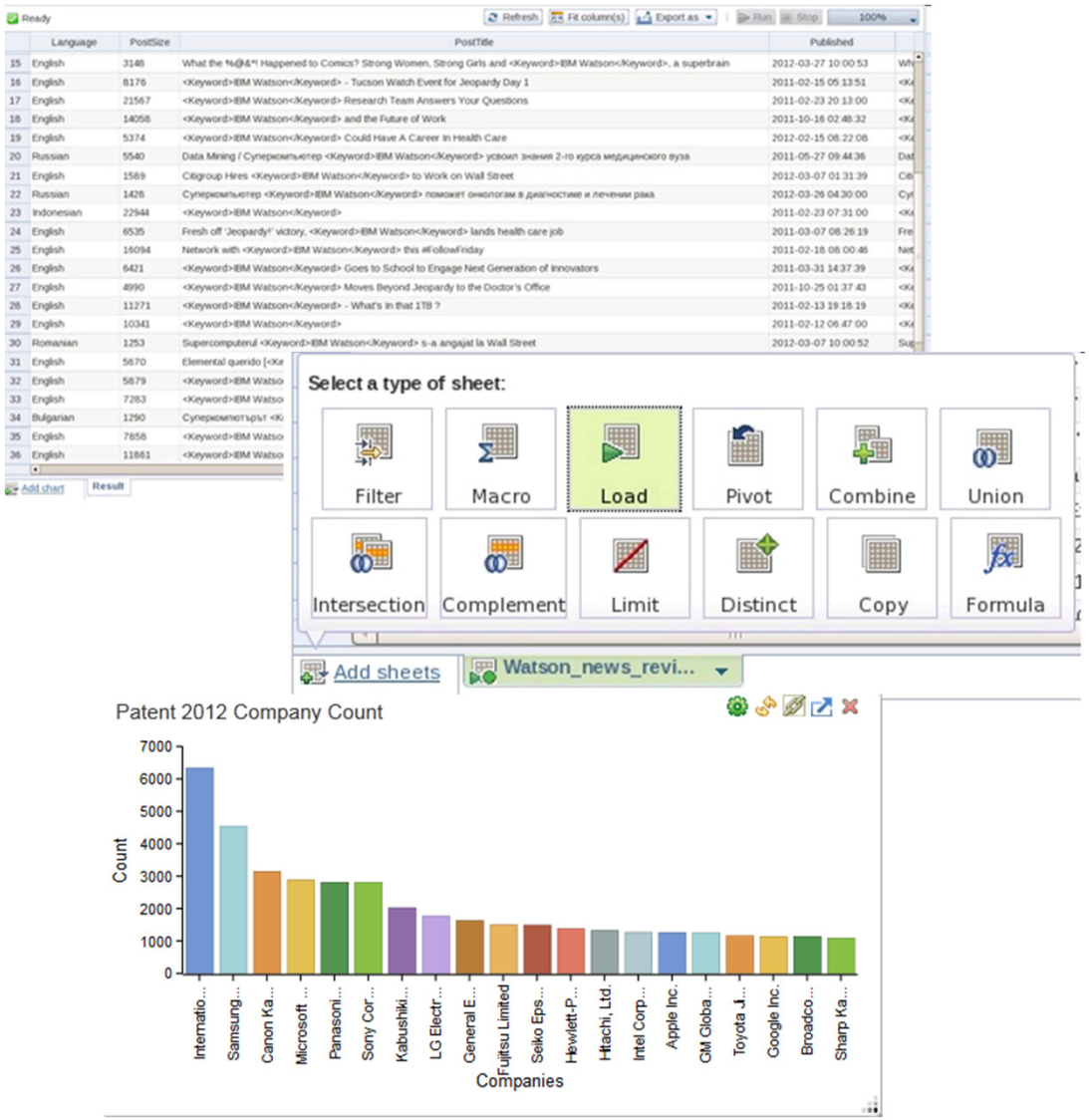
- 다양한 데이터 소스에서 수집된 데이터를 Spreadsheet 형태로 데이터 모델링

- 내장된 다양한 함수들을 활용하여 필터링, 매크로 수행 등 변형 작업 수행

- 여러 개의 워크북들에 있는 데이터를 혼합

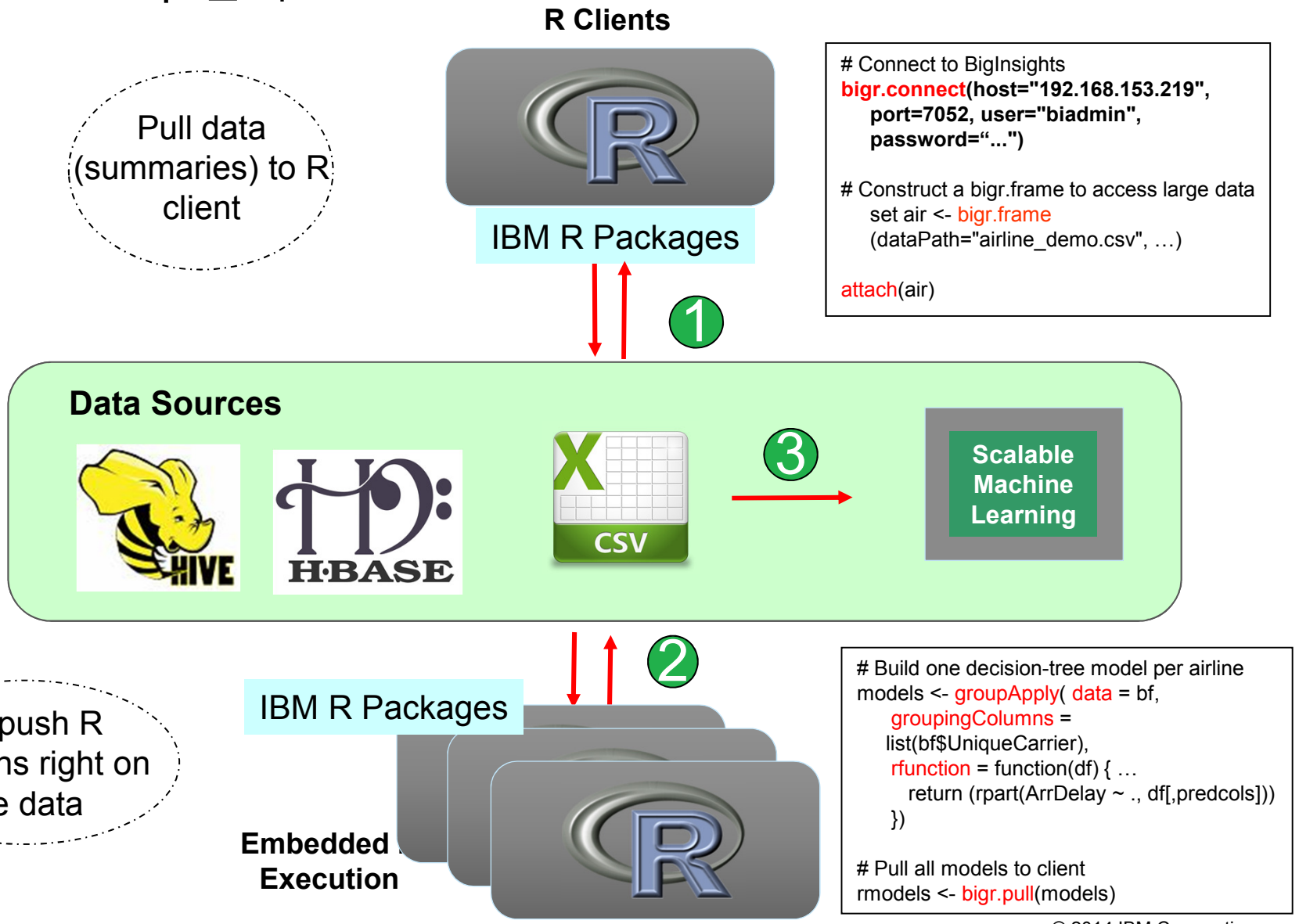
- 내장된 차트를 통해 데이터의 시각화

- 외부 데이터 연계를 위해 데이터 export



© 2014 IBM Corporation

BigR : in-Hadoop 분석




```
# Connect to BigInsights
bigr.connect(host="192.168.153.219",
    port=7052, user="biadmin",
    password="...")

# Construct a bigr.frame to access large data
    set air <- bigr.frame
    (dataPath="airline_demo.csv", …)

attach(air)
```

```
# Build one decision-tree model per airline
models <- groupApply( data = bf,
     groupingColumns =
    list(bf$UniqueCarrier),
     rfunction = function(df) { …
       return (rpart(ArrDelay ~ ., df[,predcols]))
     })

# Pull all models to client
rmodels <- bigr.pull(models)
```

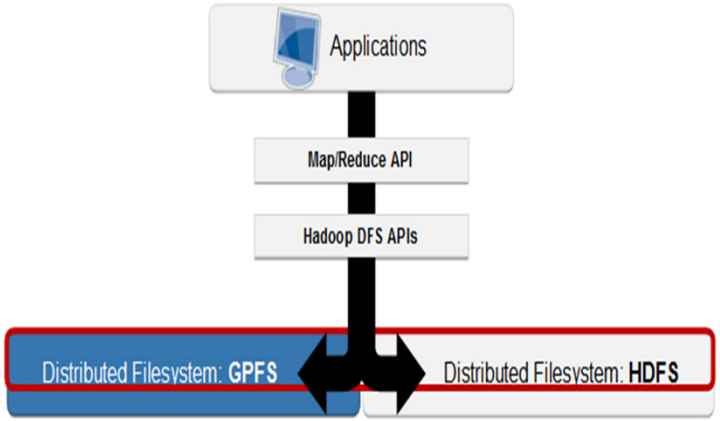

© 2014 IBM Corporation

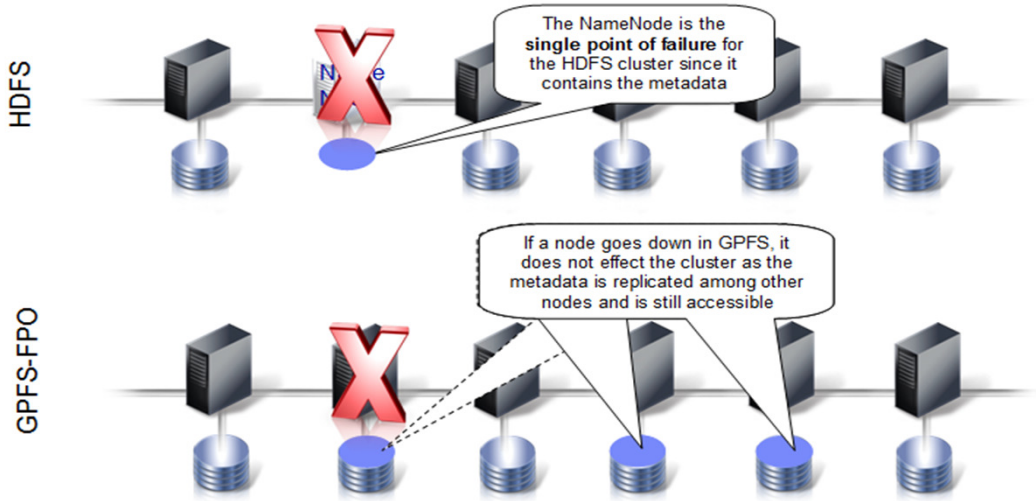
GPFS-FPO : Hadoop reliability

- GPFS-FPO을 통한 NameNode에 대한 고가용성 보장
- POSIX 호환성을 통해 Linux에서 사용하는 모든 유틸리티 활용 가능
- 백업을 위한 Snapshot 기능
- ACL 기능을 통한 보안 기능 강화

파일 시스템	GPFS-FPO	HDFS
안정상	No single point of failure	NameNode 장애에 취약
데이터 정합성	High	데이터 유실의 가능성
확장성	Thousands of nodes	Thousands of nodes
POSIX 호환	지원	제한적임
데이터 관리	보안, 백업, 복제	제한적임
MapReduce 성능	Good	Good
Workload 제어 기능	디스크 할당을 통한 제어	지원 언함
전통적인 어플리케이션 설능	Good	Random 읽기/쓰기에 취약함

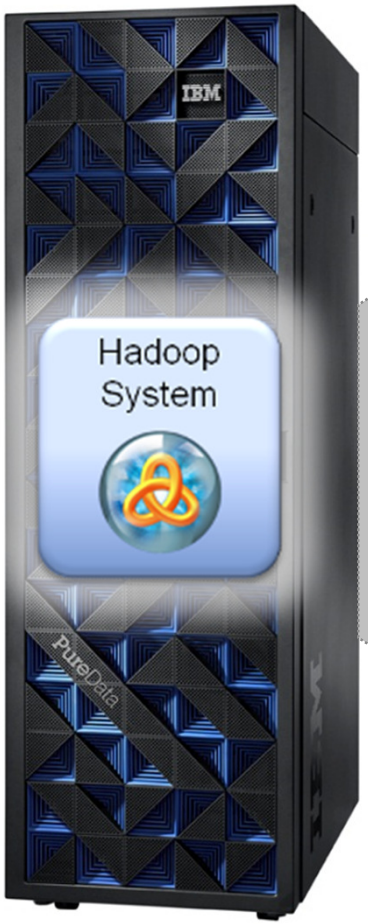
Hadoop Map/Reduce 어플리케이션은
GPFS와 HDFS를 동일하게 인지합니다.





© 2014 IBM Corporation

Appliance : Pure Data for Hadoop



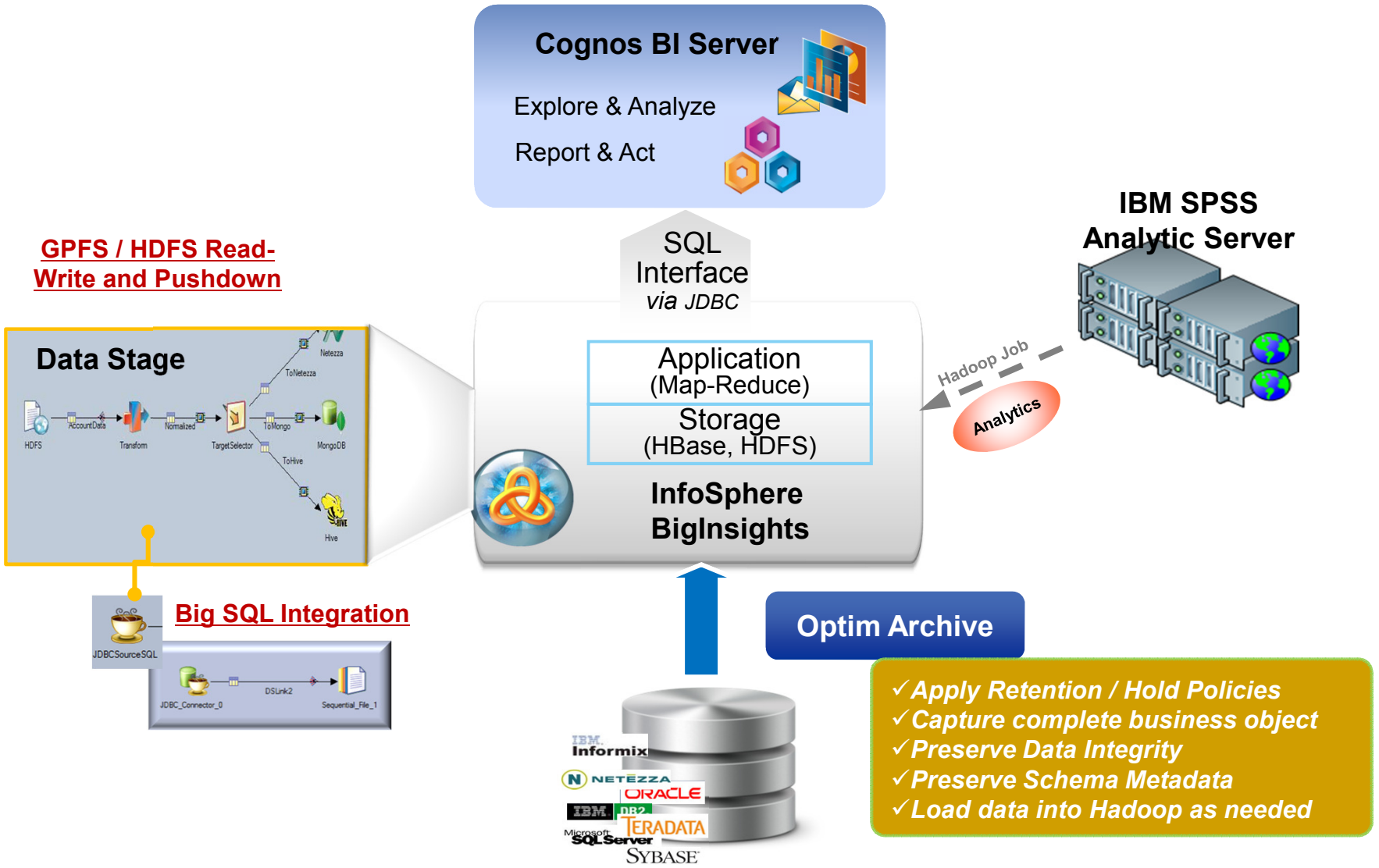
- **8x 빠른 구축**
  than custom-built clusters
- **내장된 시각화 기능**
  to accelerate insight
- **타 appliances와 달리, PureData System for Hadoop은 built-in analytic accelerators 기능을 제공합니다.**

- **싱글 시스템 콘솔**
  for full system administration
- **빠른 유지 보수**
  with automation
- **별도의 조합 없이, 빠른 시간 내에 데이터 적재 가능**

- **내장된 아카이브 툴과 통합된 Hadoop 시스템**
- **보다 강력한 보안 기능 제공**
  than open source software
- **고가용성 보장**
- **14TB/hr의 데이터 적재 성능**

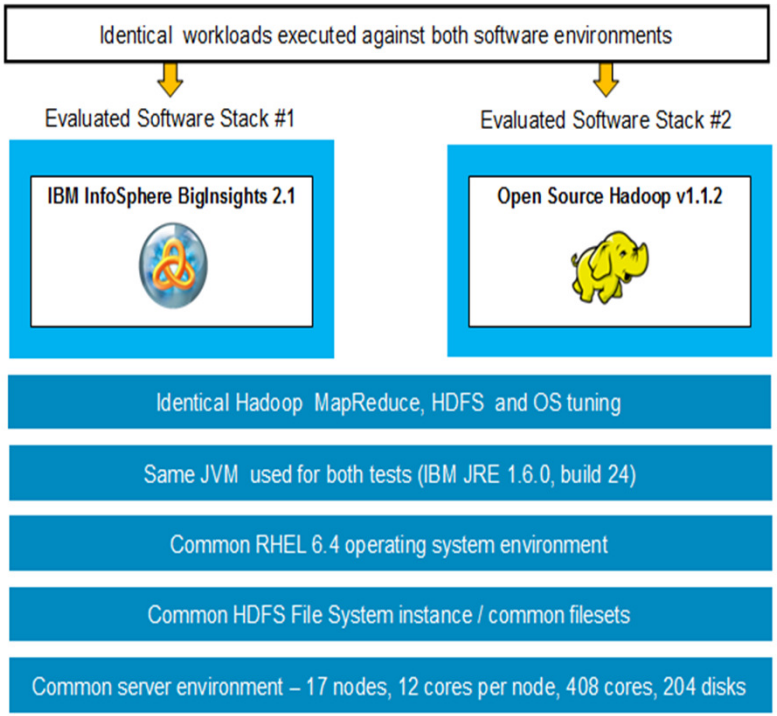
© 2014 IBM Corporation

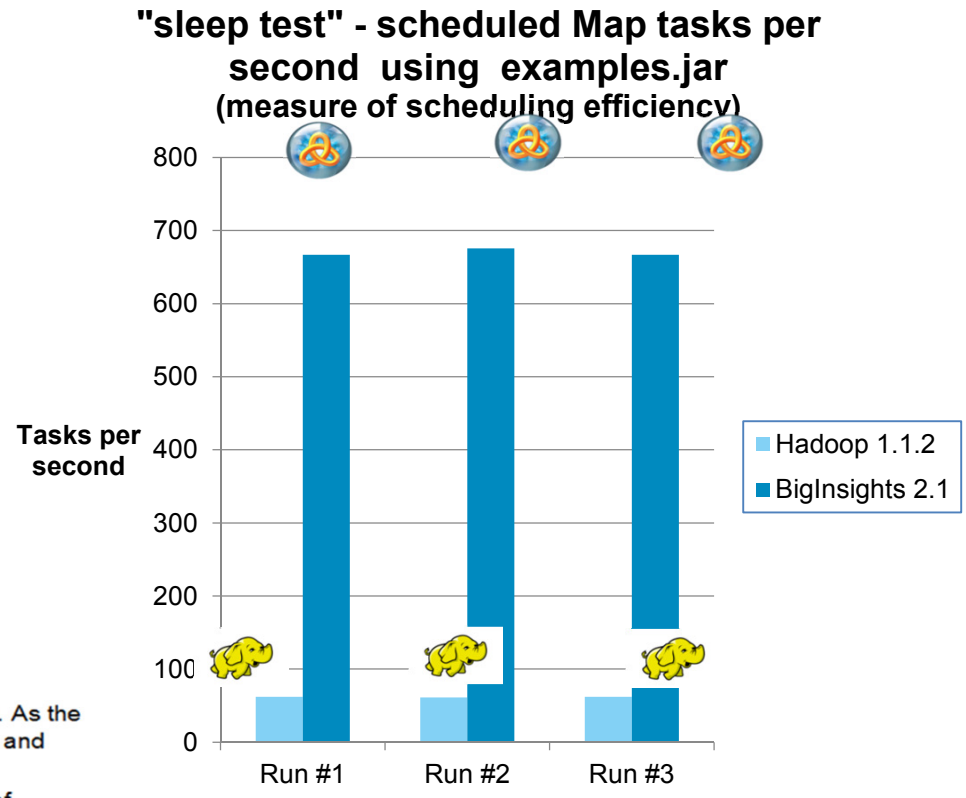
엔터프라이즈 솔루션과의 연계



© 2014 IBM Corporation

Performance

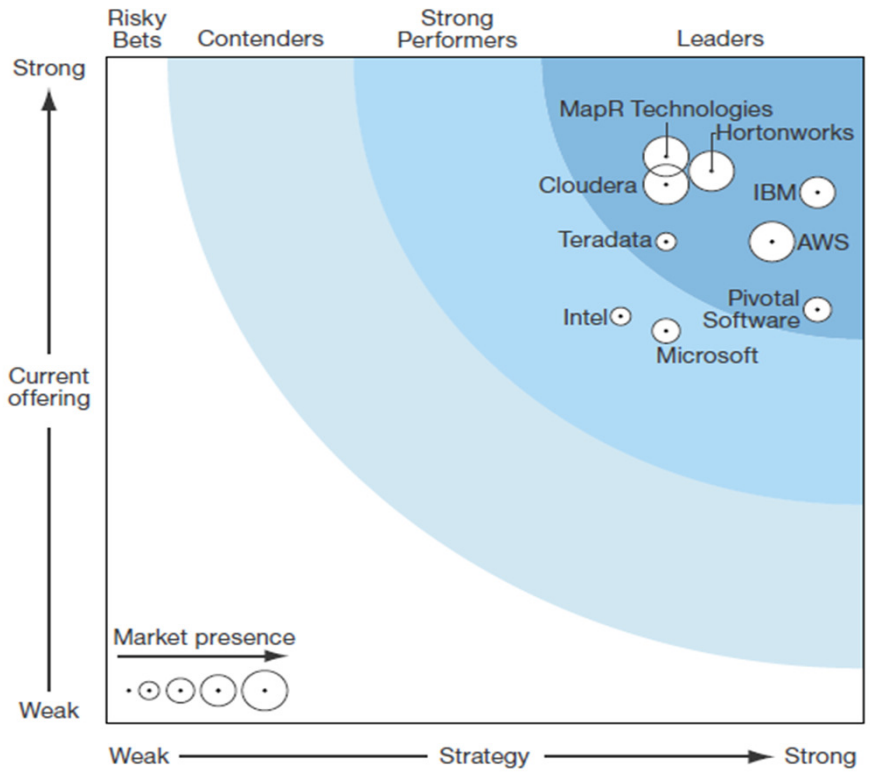




- "Sleep" is one of the standard tests included in the Hadoop distribution. As the name implies, tasks simply sleep for a specified duration for each map and reduce task dispatched to the cluster.
- Recognized as an effective way to measure the scheduling efficiency of Hadoop distributions - http://www.slideshare.net/cloudera/hadoop-world-2011-hadoop-and-performance-todd-lipcon-yanpei-chen-cloudera

© 2014 IBM Corporation

외부 평가 기관



- **IBM flexes its enterprise muscles with InfoSphere BigInsights.** Distributed computing platforms and data management are certainly not new to IBM. It has offerings in grid computing, databases, and many other data management technologies that it can bring to a comprehensive Hadoop solution. In addition, IBM has advanced analytics tools, a global presence, and implementation services, so it can offer a complete big data solution that will be attractive to many customers. IBM's road map includes continuing to integrate the BigInsights Hadoop solution with related IBM assets like SPSS advanced analytics, workload management for high-performance computing, BI tools, and data management and modeling tools. Today, IBM has more than 100 Hadoop deployments, some of which are fairly large and run to petabytes of data.

❖ the Forrester Wave : 2014 Q1

© 2014 IBM Corporation