

# **What is Big Data?**

**Bok, Jong Soon**  
**[javaexpert@nate.com](mailto:javaexpert@nate.com)**  
**<https://github.com/swacademy>**

**Now IS...**

Not



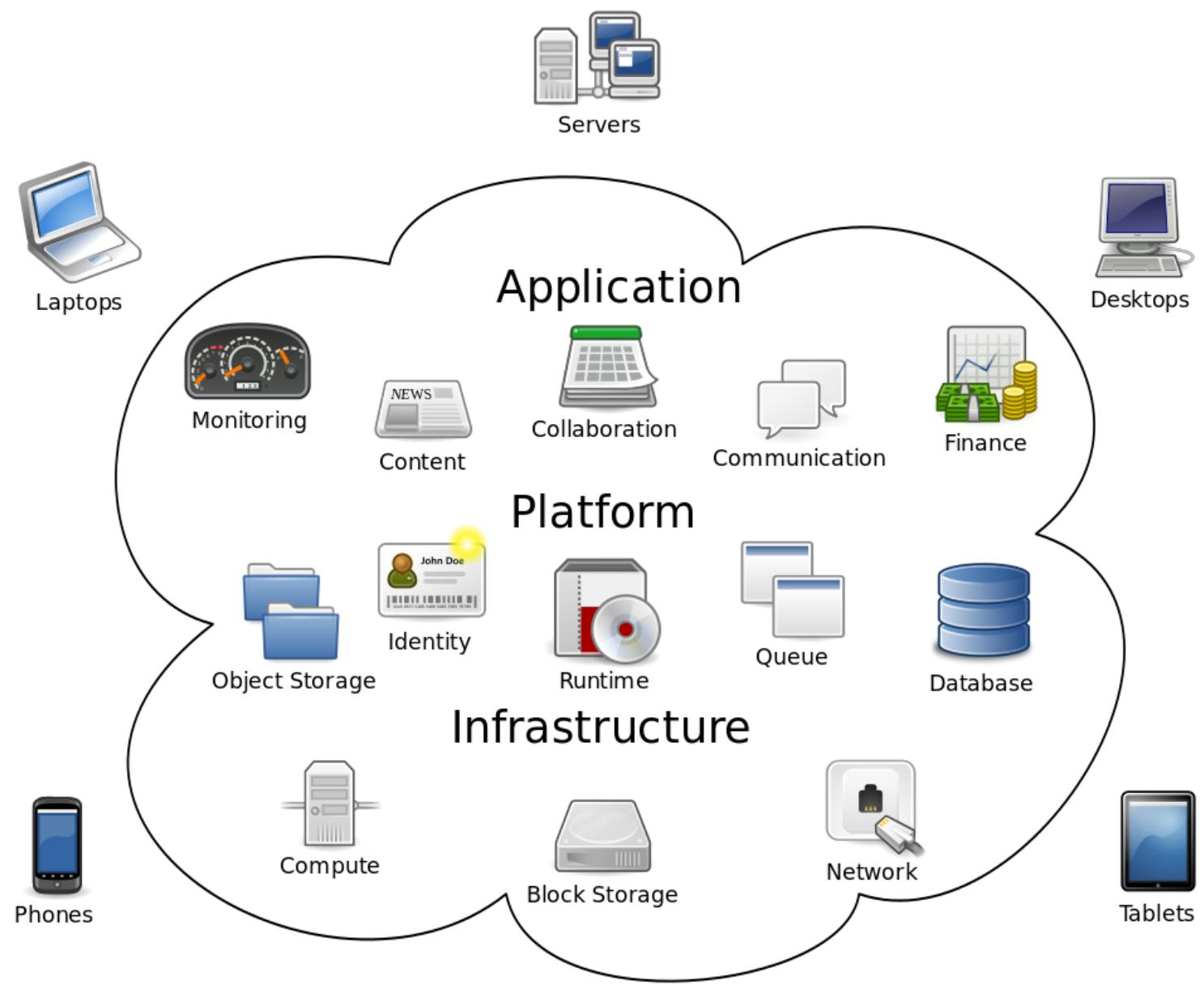
I nterC ontinental B allistic M issile



**But**

**I C B M**





# Cloud computing

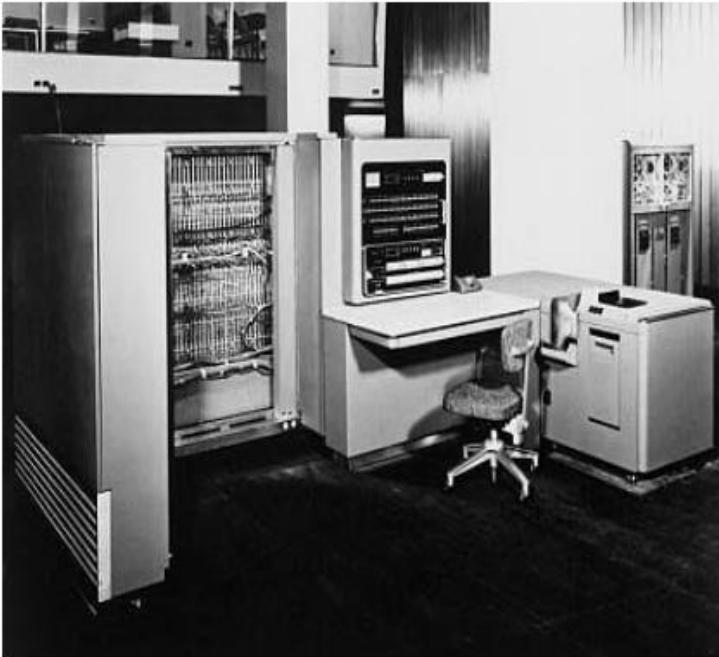




# 60여년전...

1952.5.21

The first IBM large-scale electronic computer



## IBM 701

Electrostatic storage capacity:

Magnetic drum capacity:

Magnetic tape capacity:

Addition and subtraction:

Multiplication and division:

Tape reading and writing speed:

Drum reading and writing speed:

Printed output:

Punched card input:

Punched card output:

20,480 digits.

81,920 digits.

More than **8 million** digits without changing tape.

More than 16,000 operations per second.

More than 2,000 operations per second.

12,500 digits per second.

8,000 digits per second.

180 letters or numbers per second.

600 digits per second.

400 digits per second.

# 지금은...

- 전세계 음악을 모두 저장할 수 있는 디스크 드라이브 가격 **\$600**
- 2010년 전세계 핸드폰 **50억**대 이상
- 매달 페이스북에서 주고받는 컨텐츠 **300억**건
- 매년 전세계 데이터 **40%** 씩 증가, IT 분야는 **5%** 씩 증가
- 미국 의회 도서관이 수집한 데이터 **235테라바이트** (2011.4 기준)



Google Datacenter

<http://www.google.com/about/datacenters/locations/index.html>

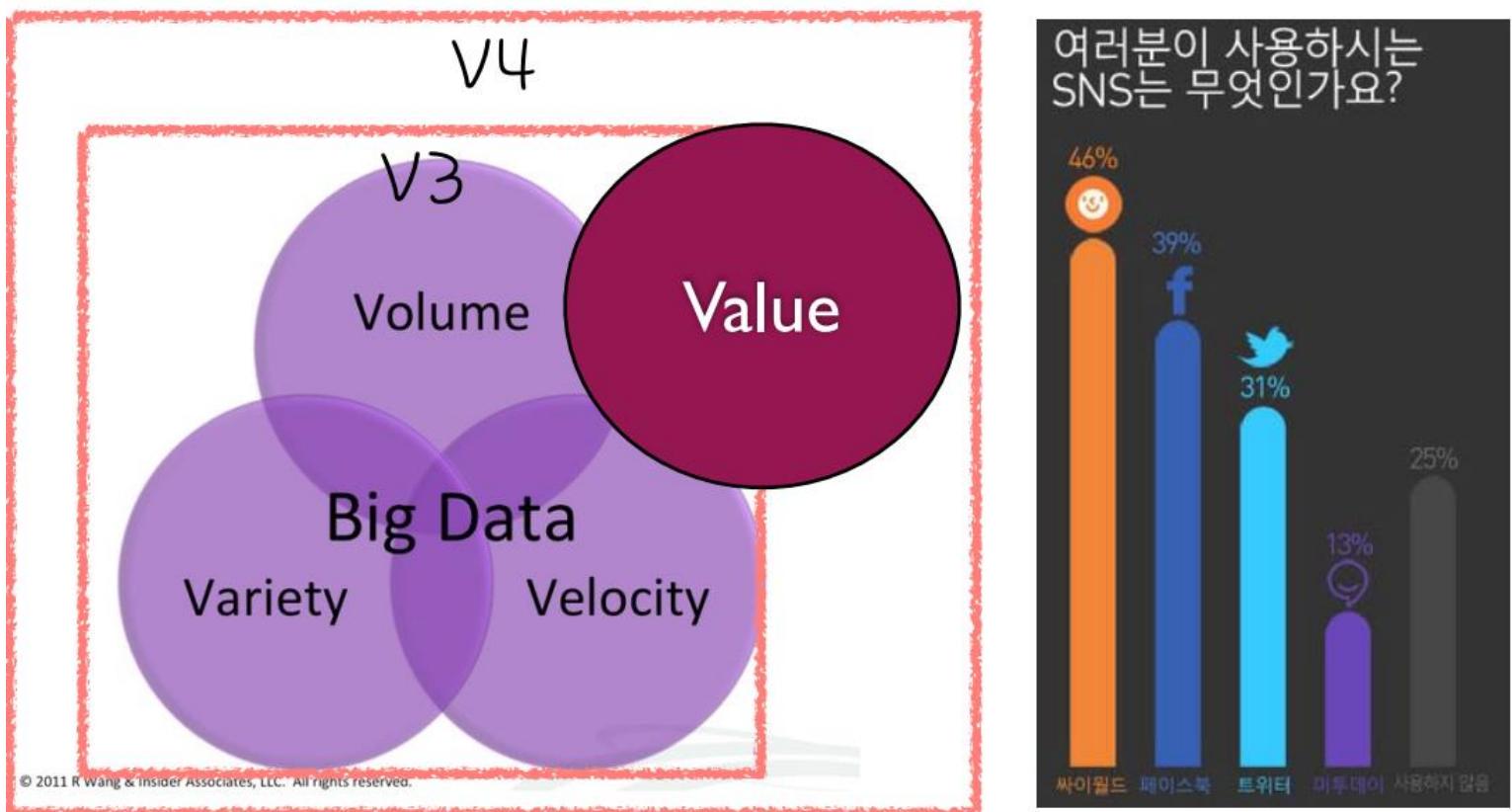


Facebook Datacenter

스웨덴 극지방 위치

# Big Data

- 기존 데이터에 비해 너무 방대해 이전 방법이나 도구로 수집, 저장, 검색, 분석, 시각화 등이 어려운 정형 또는 비정형 데이터 세트



# V3 + V1

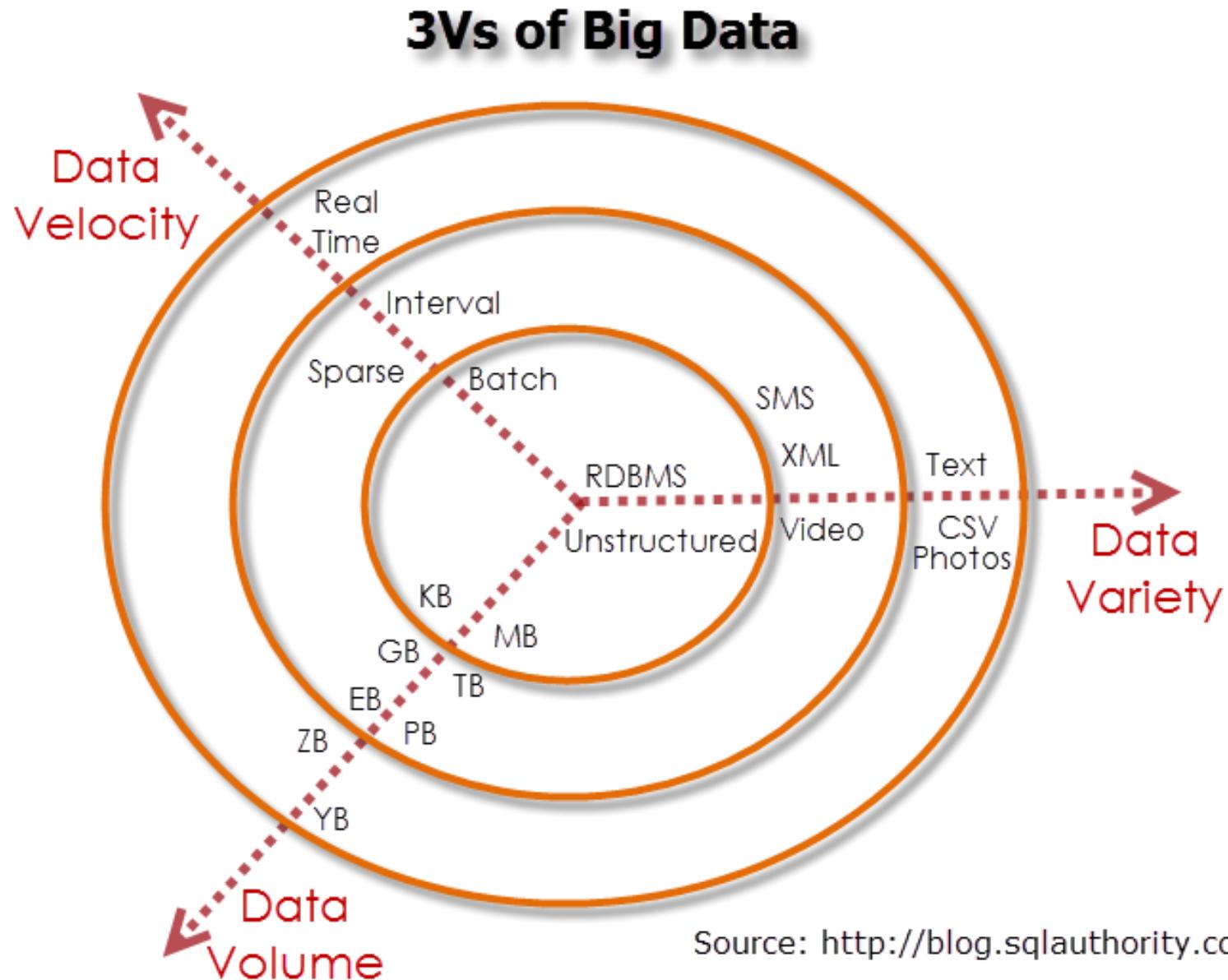
## ■ V3

- Variety : 다양
- Velocity : 유입속도 빠르고
- Volume : 대용량

## ■ V1

- Value : 분석을 잘하면

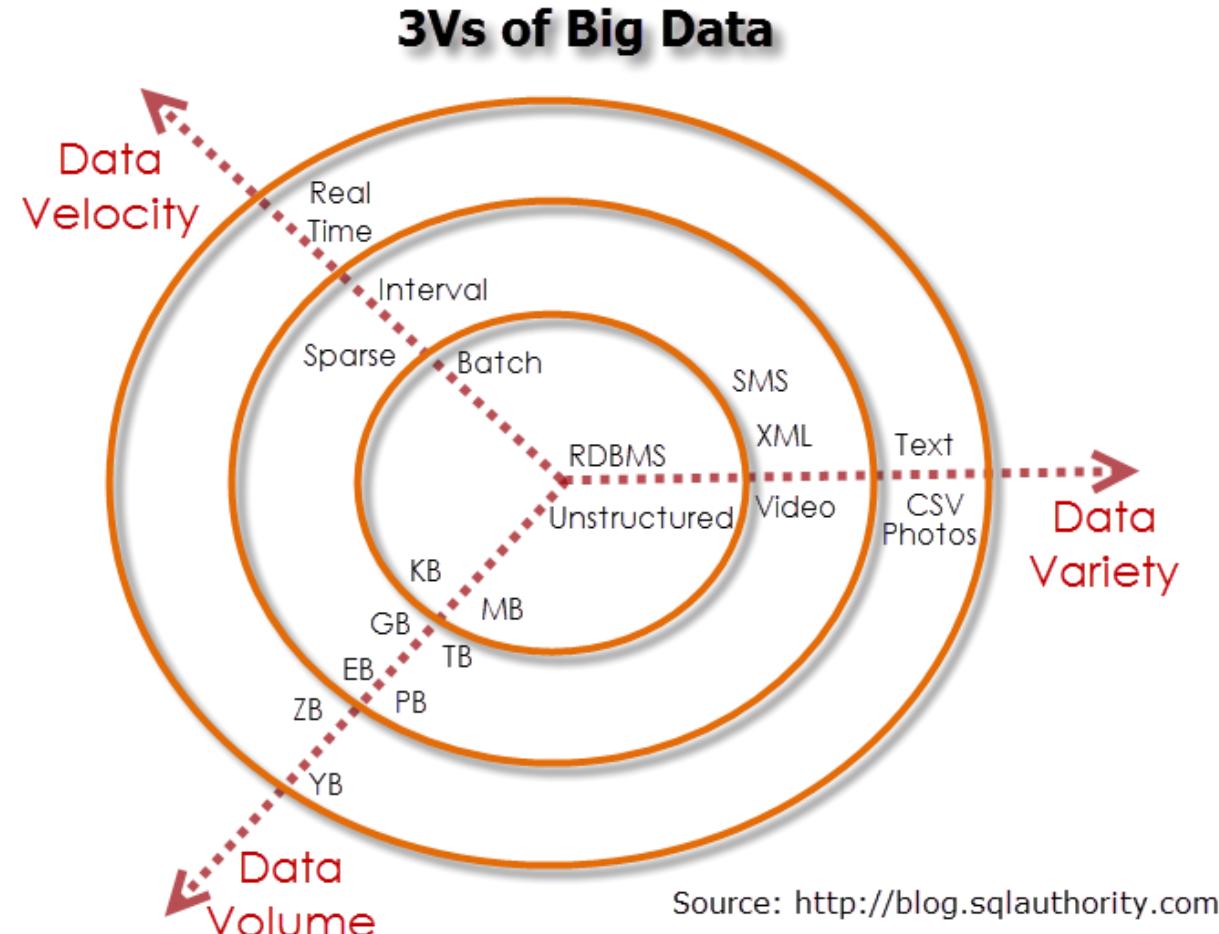
# V3 + V1 (Cont.)



# Volume – 데이터 볼륨

## ■ 그 어느 때보다 많은 Data를 생산

- 금융 거래 데이터
- 센서 데이터(IoT)
- 서버 로그 데이터
- 분석 데이터
- 이메일 및 문자 데이터
- SNS 데이터



# Velocity – 데이터 생산 속도

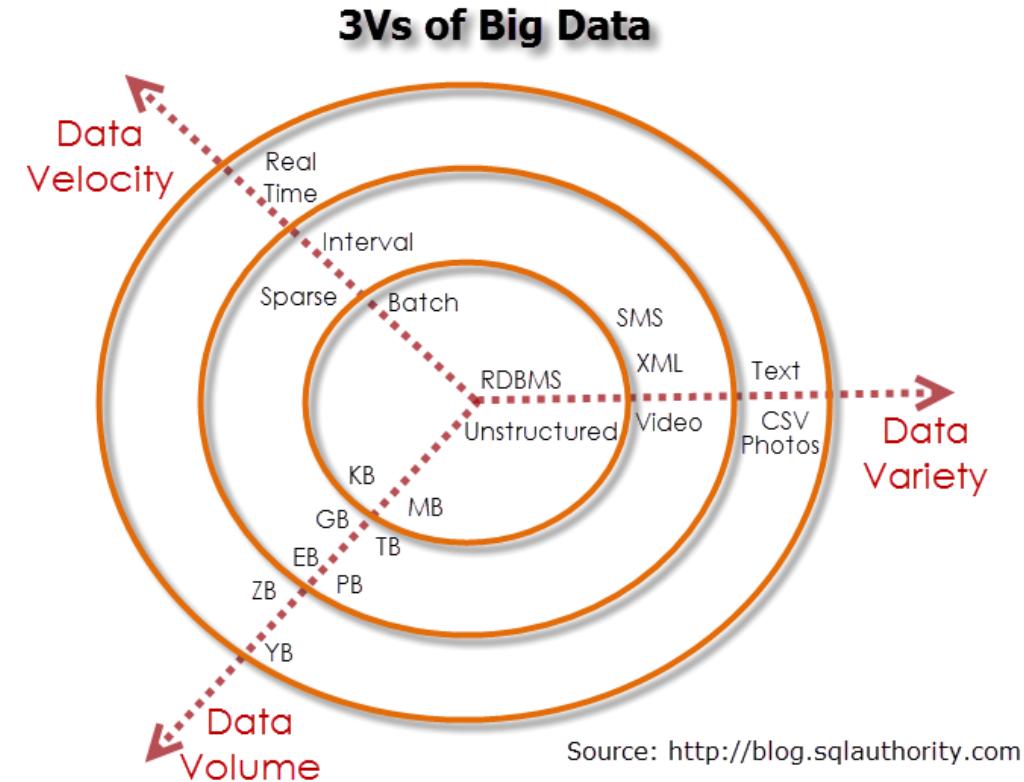
## ■ 데이터 생산 속도의 가속화

- 자동화
- 언제 어디서든 인터넷 접속
- 사용자 생성 데이터

## ■ 하루 생성되는 데이터 사례

- Twitter Message – 3.4억
- Amazon S3에 10억 개체
- Facebook 메시지와 'likes' – 27억

## ■ IoT 등장은 이 현상을 더 가속화



In 2012.

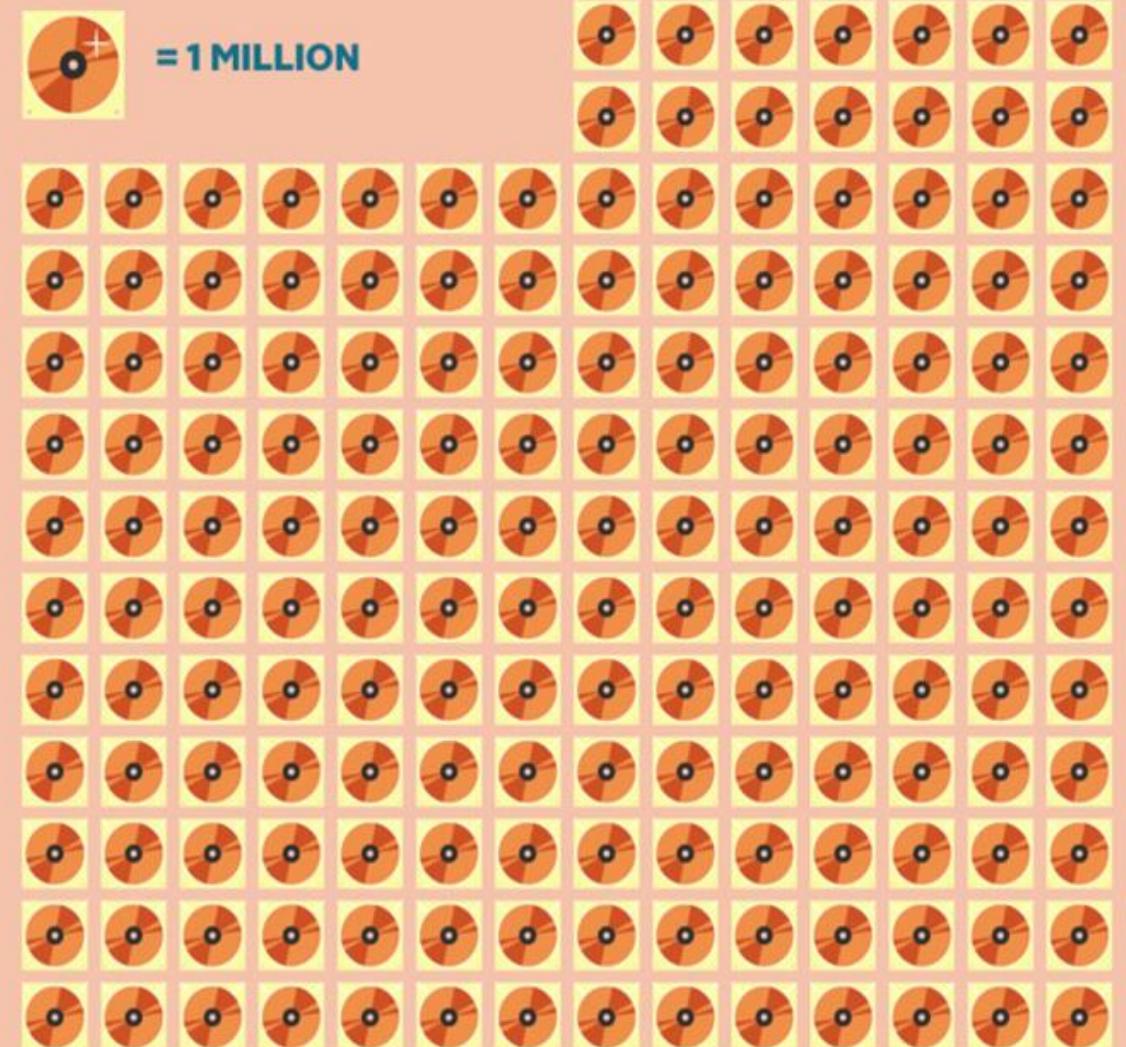


<http://www.businessinsider.com/everything-that-happens-in-one-day-on-the-internet-2012-3>

In one day, enough information is consumed by internet traffic to fill  
**168 MILLION DVDS.**



= 1 MILLION



# In 2012. (Cont.)

**294 BILLION**

emails are sent.



**2 MILLION BLOG POSTS**

are written.

Enough posts to fill  
Time Magazine for 770 years.



**172 MILLION**

different people visit Facebook.

Twitter: **40 MILLION**

LinkedIn: **22 MILLION**

Google+: **20 MILLION**

Pinterest: **17 MILLION**

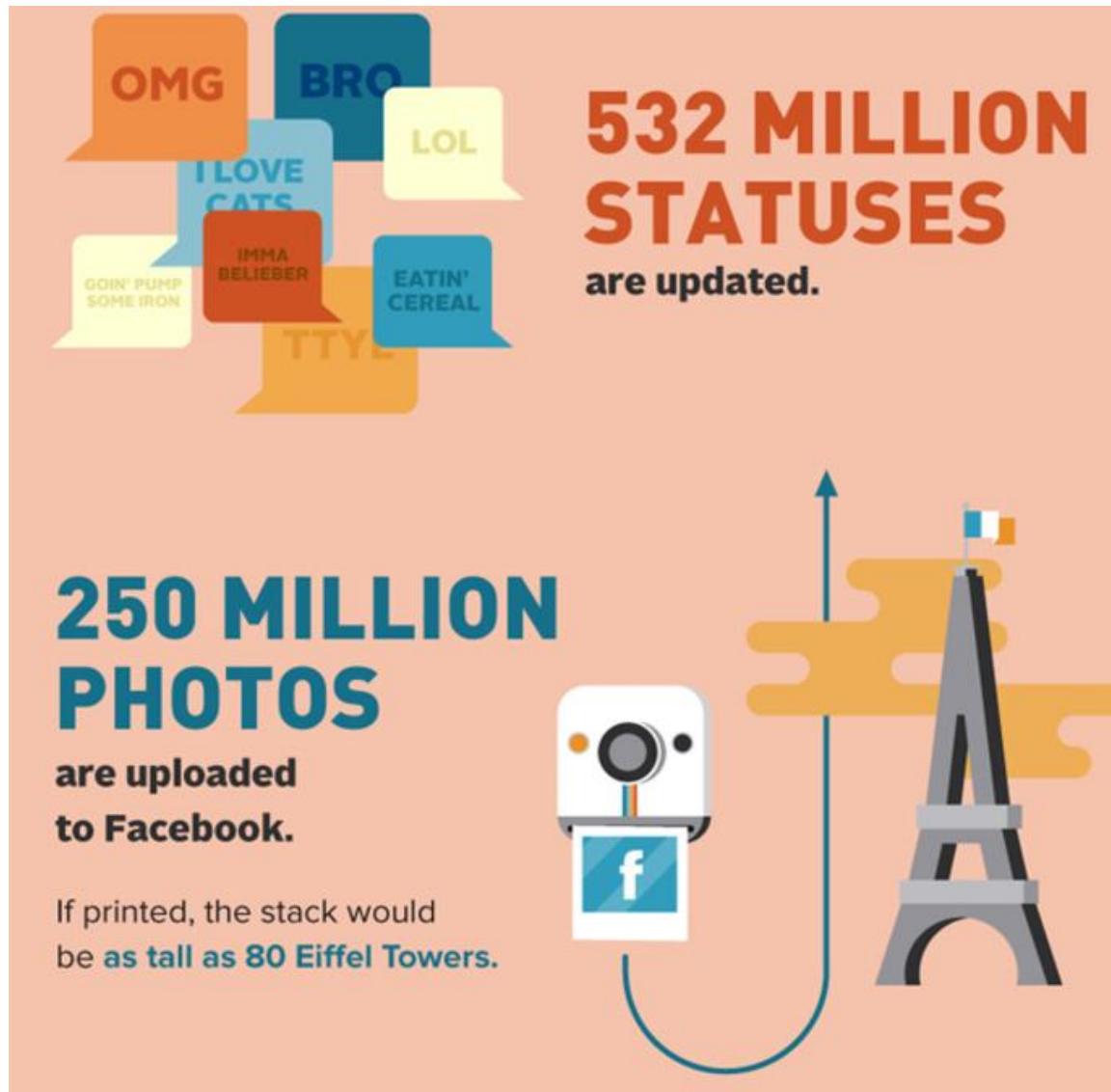


**4.7 BILLION MINUTES**

are spent on Facebook.



# In 2012. (Cont.)



# In 2012. (Cont.)



Internet users spend  
**14.6 MINUTES**  
viewing porn online.

The average fap session  
is 12 minutes.

**18.7 MILLION  
HOURS OF MUSIC**  
is streamed on Pandora.

If a computer started streaming Pandora  
in year 1 AD, it'd still be streaming now.



**1288 NEW APPS  
TO DOWNLOAD**  
And more than 35 million  
apps are downloaded.

**IPHONE SALES OUTPACE**  
the human population.



A large blue downward-pointing arrow.

**378,000**

Number of iPhones Sold



**371,000**

Number of babies born

# In 2016.

## How Much Data is Produced Every Day?



2.5 Exabytes are produced every day

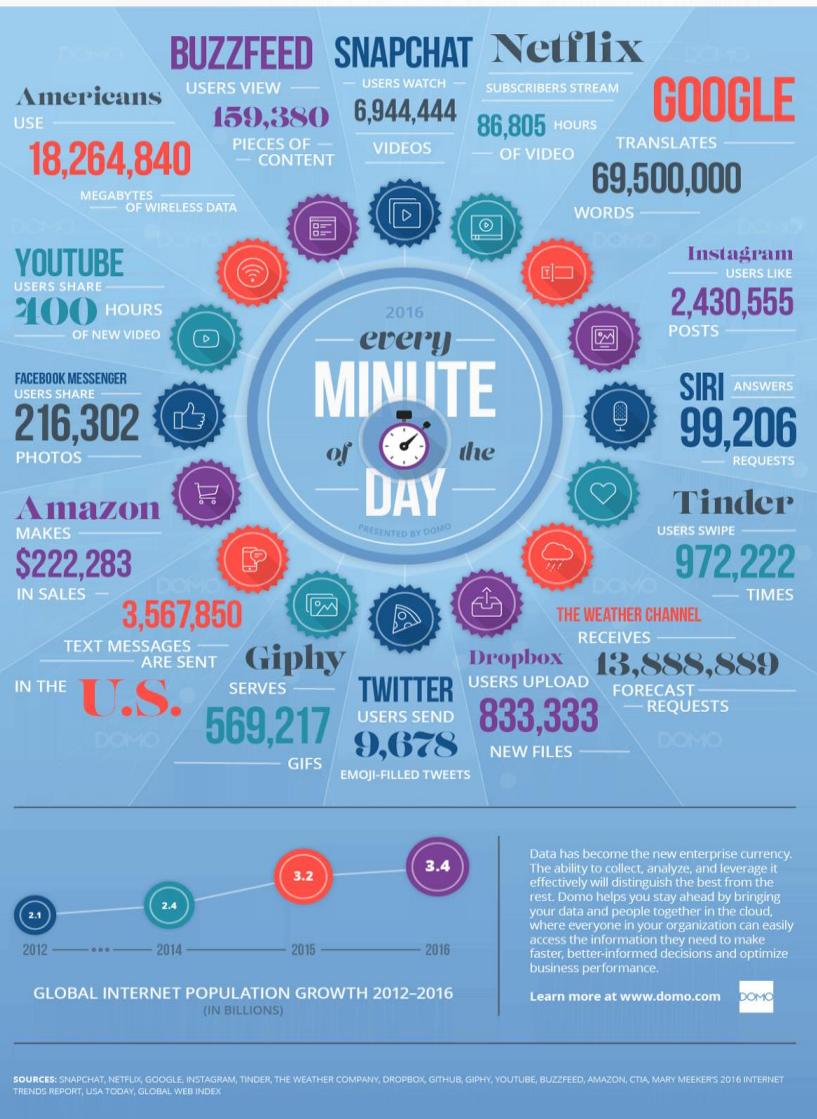
Which is equivalent to:

- ♫ 530,000,000 millions songs
- 📱 150,000,000 iPhones
- 💻 5 million laptops
- 🔖 250,000 Libraries of Congress
- ▶ 90 years of HD Video



## DATA NEVER SLEEPS 4.0

How much data is generated every minute? In the fourth annual edition of Data Never Sleeps, newcomers like Giphy and Facebook Messenger illustrate the rise of our multimedia messaging obsession, while veterans like YouTube and Snapchat highlight our insatiable appetite for video. Just how many GIFs, videos, and emoji-filled Tweets flood the Internet every minute? See for yourself below.

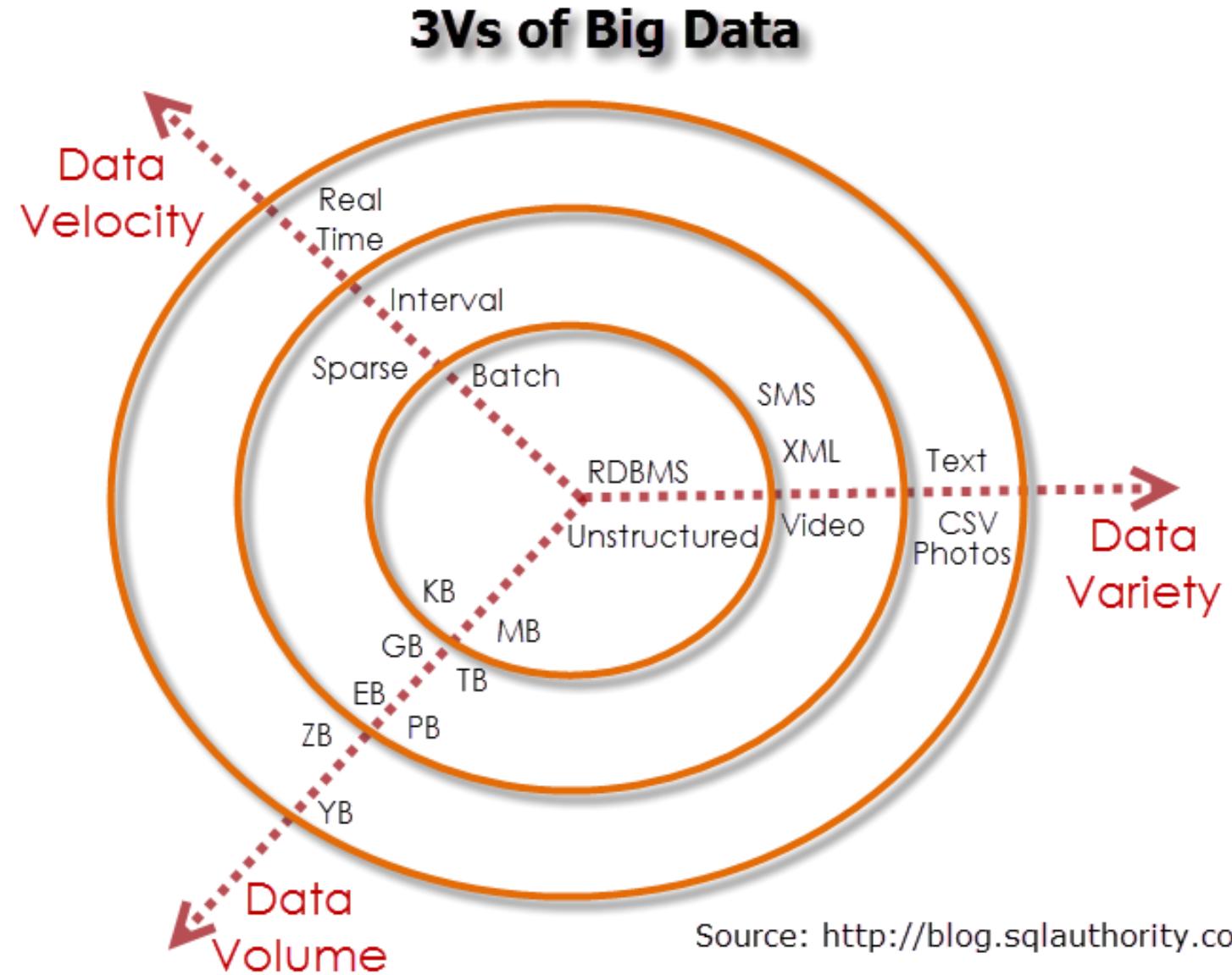


- **500 Million Tweets** sent each day!
- More than **4 Million Hours of content uploaded to Youtube** every day!
- **3.6 Billion Instagram Likes** each day.
- **4.3 BILLION Facebook messages** posted daily!
- **5.75 BILLION Facebook likes** every day.
- **40 Million Tweets** shared each day!
- **6 BILLION** daily Google Searches!

<https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/>

# Variety – 데이터 다양성

- 정형데이터
- RDBMS
- 비정형 데이터
- 비디오
- 사진
- 반정형 데이터
- 의료정보
- 시스템 로그
- XML
- PDF
- 텍스트

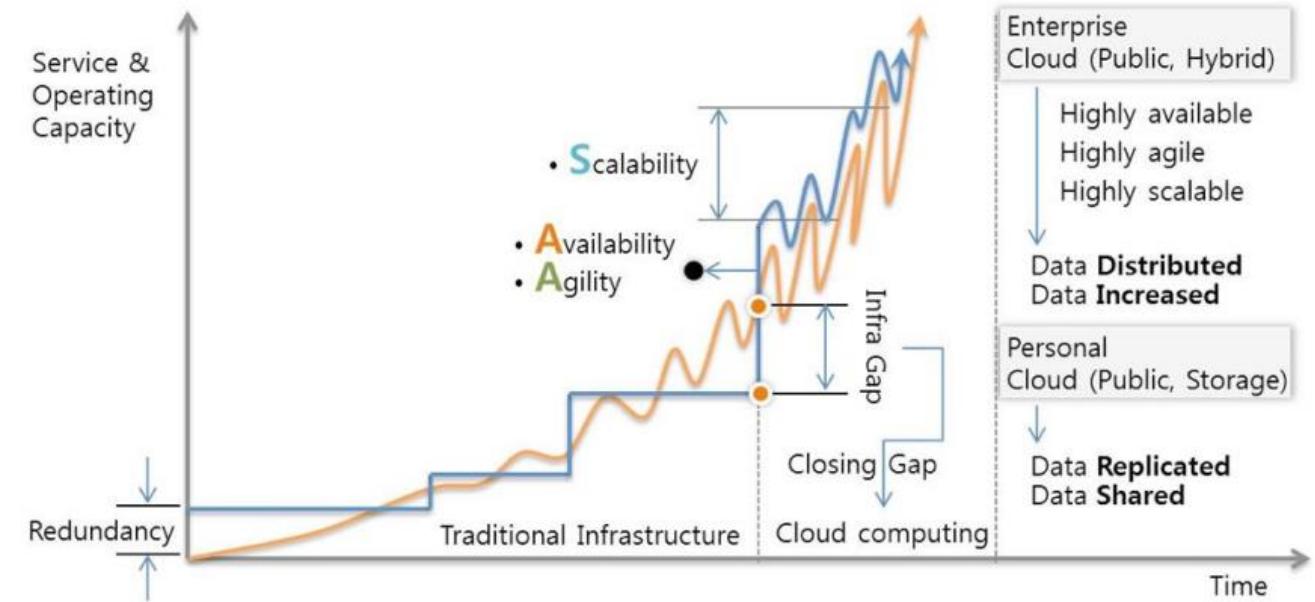


# Big Data (Cont.)

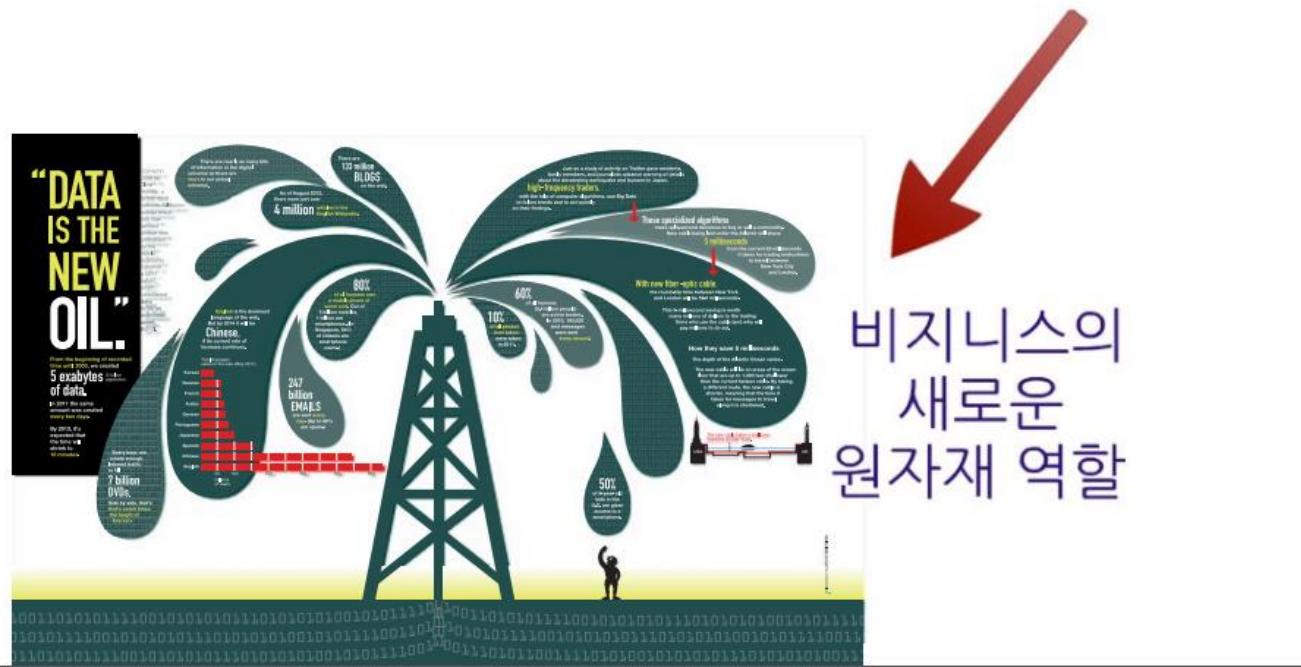
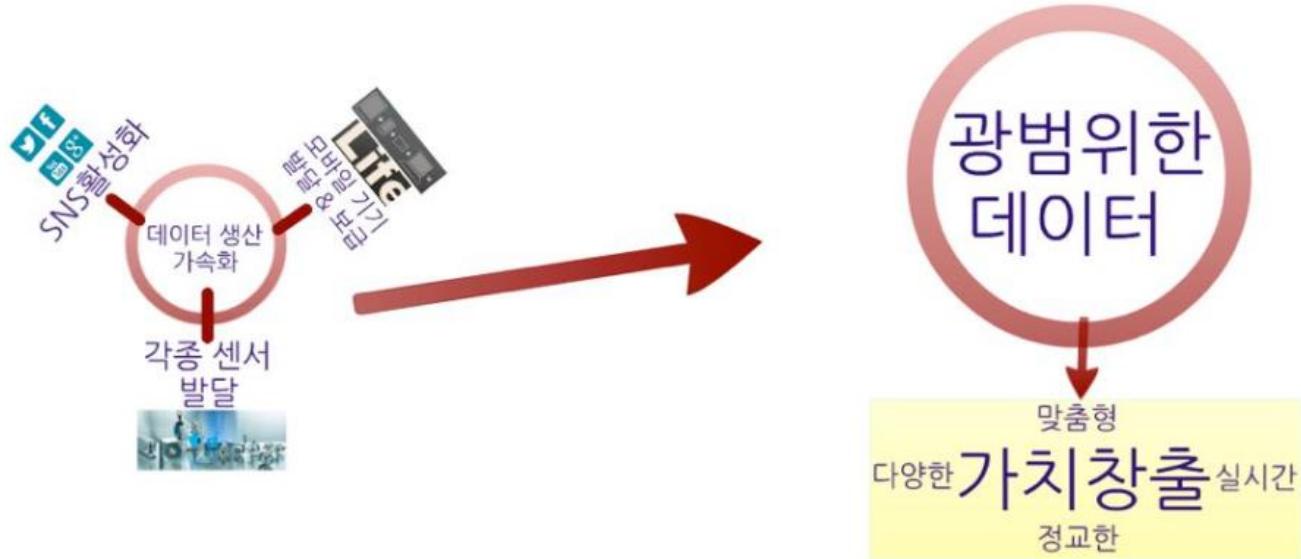
- 기존 데이터에 비해서 너무 방대해 일반적으로 사용하는 방법이나 도구로 수집, 저장, 검색, 분석, 시각화 등을 하기 어려운 정형 또는 비정형 데이터 집합을 의미
- 데이터베이스의 규모에 초점을 맞춘 정의(McKinsey)
  - 일반적인 데이터베이스 S/W가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터
- 데이터베이스가 아닌 업무 수행에 초점을 맞춘 정의(IDC)
  - 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 (데이터의)초고속, 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처

# 출현 배경

- TV, PC에 이은 스마트 폰 확산
- 기업의 고객 데이터 추적/수집행위 증가
- 멀티미디어 콘텐츠 사용 관련 정보 증가
- SNS 급격한 확산과 비정형 데이터의 폭증
  - 정보의 바다 → 정보의 홍수



# Big Data의 가치



비지니스의  
새로운  
원자재 역할

# Big Data의 가능성

- 활용할 자료의 양과 방법 등이 다양
  - 하루하루 기하급수적 늘어가는 방대한 데이터의 양
  - 생활에 많은 곳에 적용 가능성이 높음
  - 다양한 비즈니스 모델 창출 가능



# Big Data Landscape

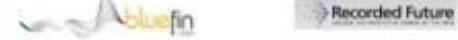
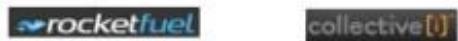
## Vertical Apps



## Log Data Apps



## Ad/Media Apps



## Business Intelligence

ORACLE | Hyperion

SAP Business Objects | RJMetrics

Microsoft | Business Intelligence

IBM COGNOS | birst

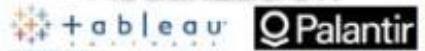
Autonomy | MicroStrategy

QlikView | bime

Chart.io | DOMO

GoodData

## Analytics and Visualization



## Data As A Service



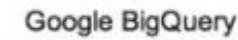
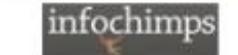
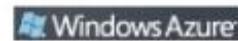
## Analytics Infrastructure



## Operational Infrastructure



## Infrastructure As A Service



## Structured Databases



## Technologies



# Marketing Technology Landscape

September 2012



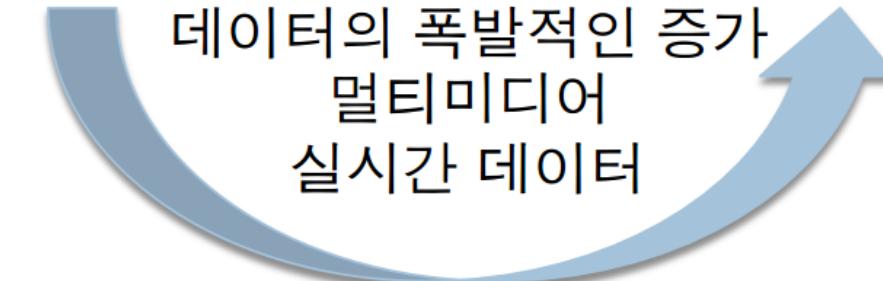
# 왜 Big Data에 관심을 가지나?

- Google
- Facebook
- Yahoo
- LinkedIn
- Twitter

- Connected Everything
  - Smartphone
  - LTE
  - Sensor Networks
  - ... ...



Cost < Value



# Big Data에서 다루는 문제들

## Volume

대용량 데이터  
(GB을 넘어서 TB,PB...)

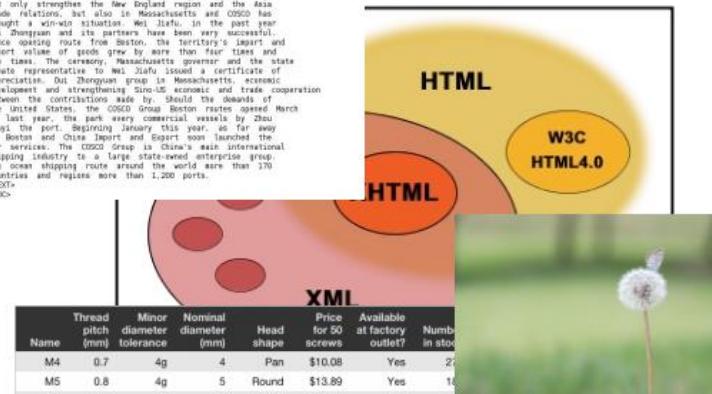


## Variety

다양한 형태의 데이터

(DB, 텍스트, XML, 이미지, 동영상...)

```
<DOC>
<DOCID> XIN200303401_0330_0001 </DOCID>
<DOCTYPE> M06 </DOCTYPE>
<TEXTTYPE> </TEXTTYPE>
<DOCINFO>
Xinhua News Agency, Boston, March 30 (expected reporter Yang
Zhi) Massachusetts Port Authority here on March held activities
to mark the celebration of the first Boston Harbor anniversary.
Zhen he round aircraft first Boston Harbor anniversary,
COSCO president Wei Jiafu, Massachusetts Port Authority secretary
Chen Zhen, COSCO chairman Wang Hui, COSCO executive director
Huang Hukang, nearly 300 people attended the celebration
ceremony. Lyric at the ceremony, far from Boston, Massachusetts,
for the opening of the round port, the opening of the regular
benefit from several other states. The opening of routes
not only strengthens the New England region and the Asia
trade, also has a positive impact on the local economy. This
brought a win-win situation. Wei Jiafu, in the past year
Wei Zhongyuan and its partners have been very successful.
Since the beginning of this year, the volume of import and
export value of goods grew by more than four times and
two times. The ceremony Massachusetts governor and the state
deputy representative to Wei Jiafu issued certificates of
appreciation. Wei Zhongyuan general manager Massachusetts economic
development and strengthening Sino-US economic and trade cooperation
and exchange, contribution. COSCO Group Boston routes opened
the United States and China COSCO Group Boston routes opened
March 22 last year, this park every commercial vessels by Zhou
Yang, Yang, Wei Jiafu, Wei Jiafu, this year, Wei Jiafu
Wei Jiafu and China Import and Export Bank launched the
air services. The COSCO Group is China's main international
shipping industry to a large state-owned enterprise group.
Its ocean shipping ports around the world are more than 370
countries and regions more than 1,200 ports.
</TEXT>
</DOC>
```



Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock
M4	0.7	4g	4	Pan	\$10.08	Yes	27
M5	0.8	4g	5	Round	\$13.89	Yes	14
M6	1	5g	6	Button	\$10.42	Yes	1043
M8	1.25	5g	8	Pan	\$11.98	No	298
M10	1.5	6g	10	Round	\$16.74	Yes	488
M12	1.75	7g	12	Pan	\$18.26	No	998
M14	2	7g	14	Round	\$21.19	No	235
M16	2	8g	16	Button	\$23.57	Yes	292
M18	2.1	8g	18	Button	\$25.87	No	664
M20	2.4	8g	20	Pan	\$29.09	Yes	486
M24	2.55	9g	24	Round	\$33.01	Yes	982
M28	2.7	10g	28	Button	\$35.66	No	1067
M36	3.2	12g	36	Pan	\$41.32	No	434
M50	4.5	15g	50	Pan	\$44.72	No	740

## Velocity

배치, 실시간, 스트리밍  
(센서, 상거래, 주식거래...)



# Big Data에서 다루는 문제들 (Cont.)

대용량 데이터를 저장, 처리하기 위해서 필요한  
클러스터 컴퓨팅, 분산컴퓨팅 인프라에 대한 이해



- 수십-수천대의 서버를 구축, 관리
- 효율적인 네트워크 구축
- 상태 모니터링 및 장애 대책
- 애플리케이션 배포
- 데이터의 저장, 백업
- 확장성, 가용성을 고려

# Big Data에서 다루는 문제들 (Cont.)

빅 데이터를 위한 다양한 오픈 소스



# Big Data에서 다루는 문제들 (Cont.)

레거시 시스템과의 연동, 마이그레이션 이슈



Legacy Data Platform



Big Data Platform

# Big Data에서 다루는 문제들 (Cont.)

데이터 마이닝 , 머신 러닝 알고리즘

Single Machine → Multiple Machines



## Big Data에서 다루는 문제들 (Cont.)

데이터 유출시 기업경영에 치명적인 영향을 준다  
어느 데이터나 활용할 수 있는 것이 아니다  
통합보다 분산이 더 안전

데이터 보안



개인 정보 보호



# Big Data에서 다루는 문제들 (Cont.)

대용량 데이터, 정형/비정형데이터, 이벤트/스트리밍 데이터

분산컴퓨팅, 클러스터 컴퓨팅

클라우드 컴퓨팅

오픈소스

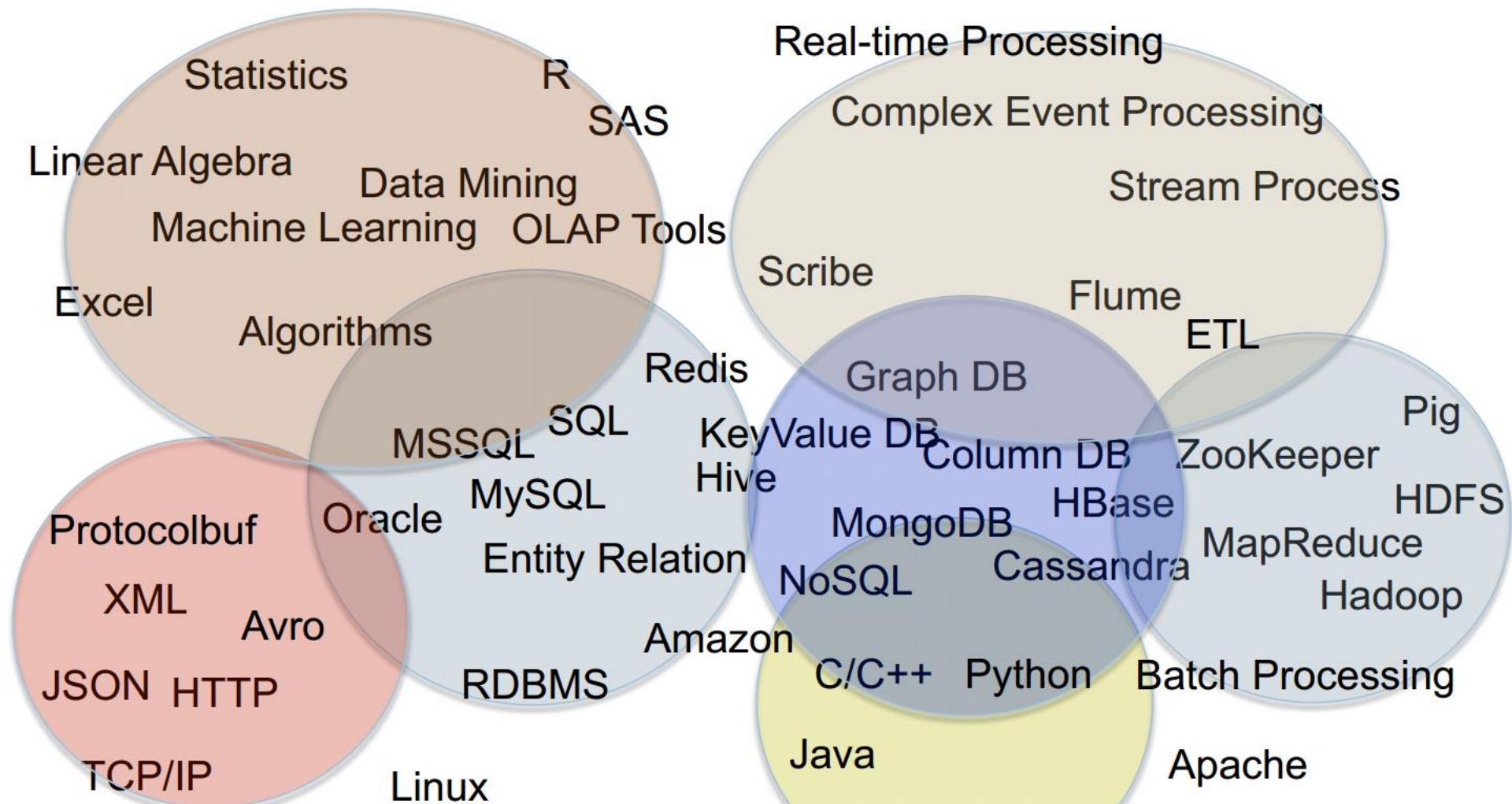
데이터 마이닝/머신러닝

레거시 시스템 연동 / 데이터 수집

보안/개인정보 보호 이슈

Parallel Processing  
Distributed Computing

# Knowledge & Technology for Big Data



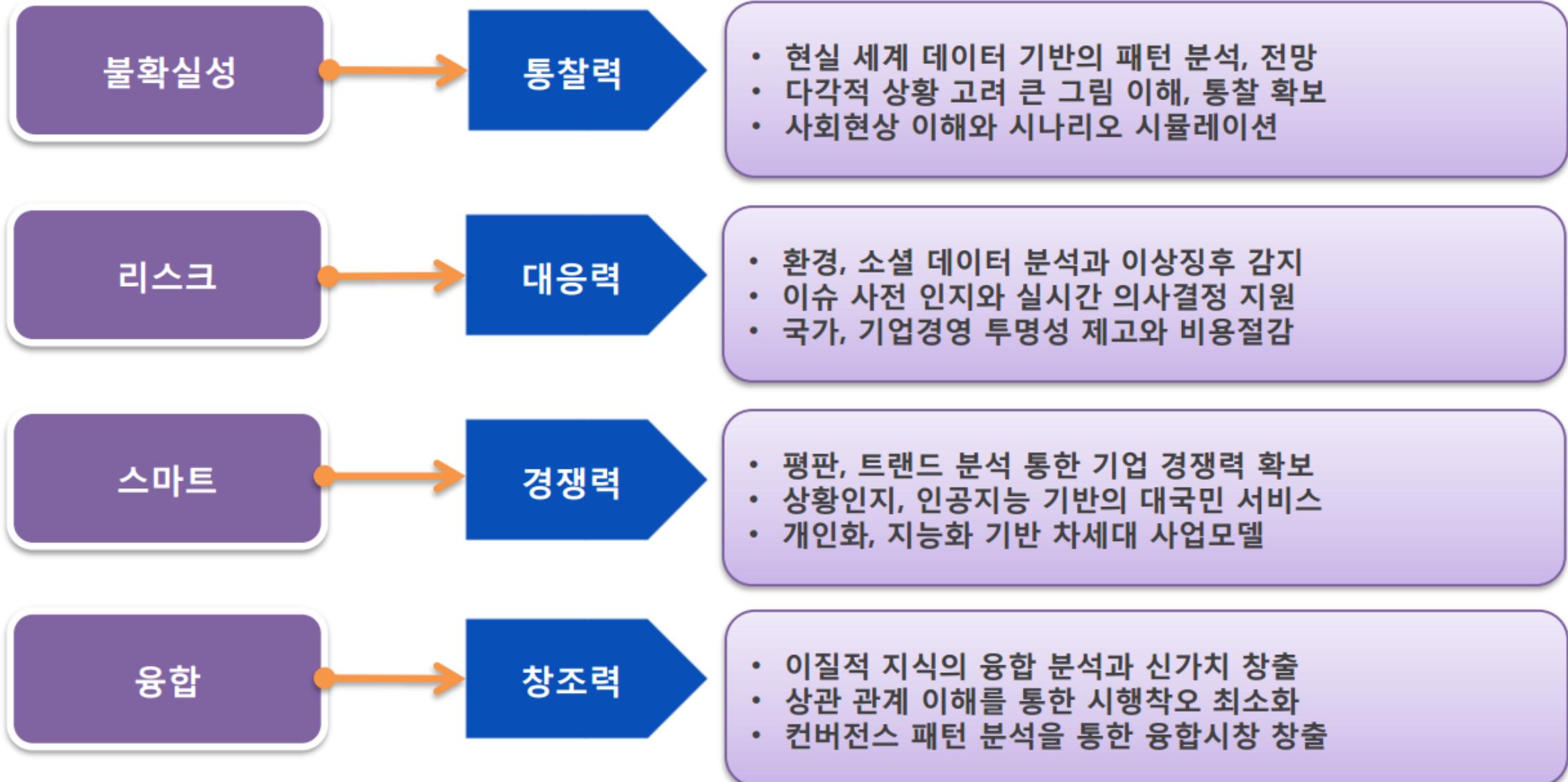
***Distributed Computing***

출처 : <http://kimws.wordpress.com>

***Cluster Computing***

***Cloud Computing***

# Big Data의 역할



# Big Data의 역할 (Cont.)

생산성과  
효율성 제고

- 실시간으로 업무 현황과 실행 성과를 파악하고 공유하여 개선을 도모

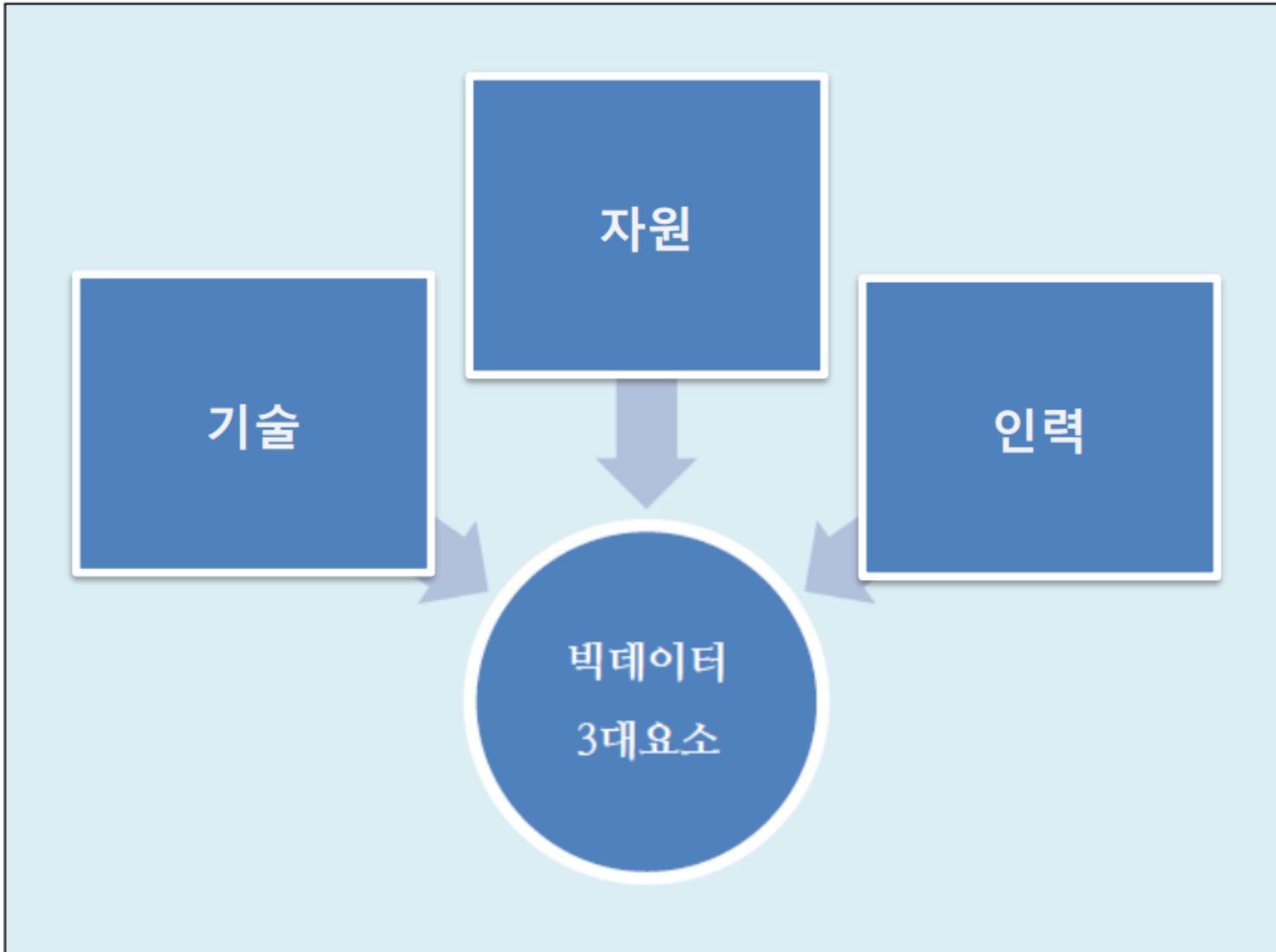
의사결정  
능력 향상

- 광범위한 기업 내외부 데이터에 근거하여 의사결정의 정확도를 향상

문제 발견 및  
해결

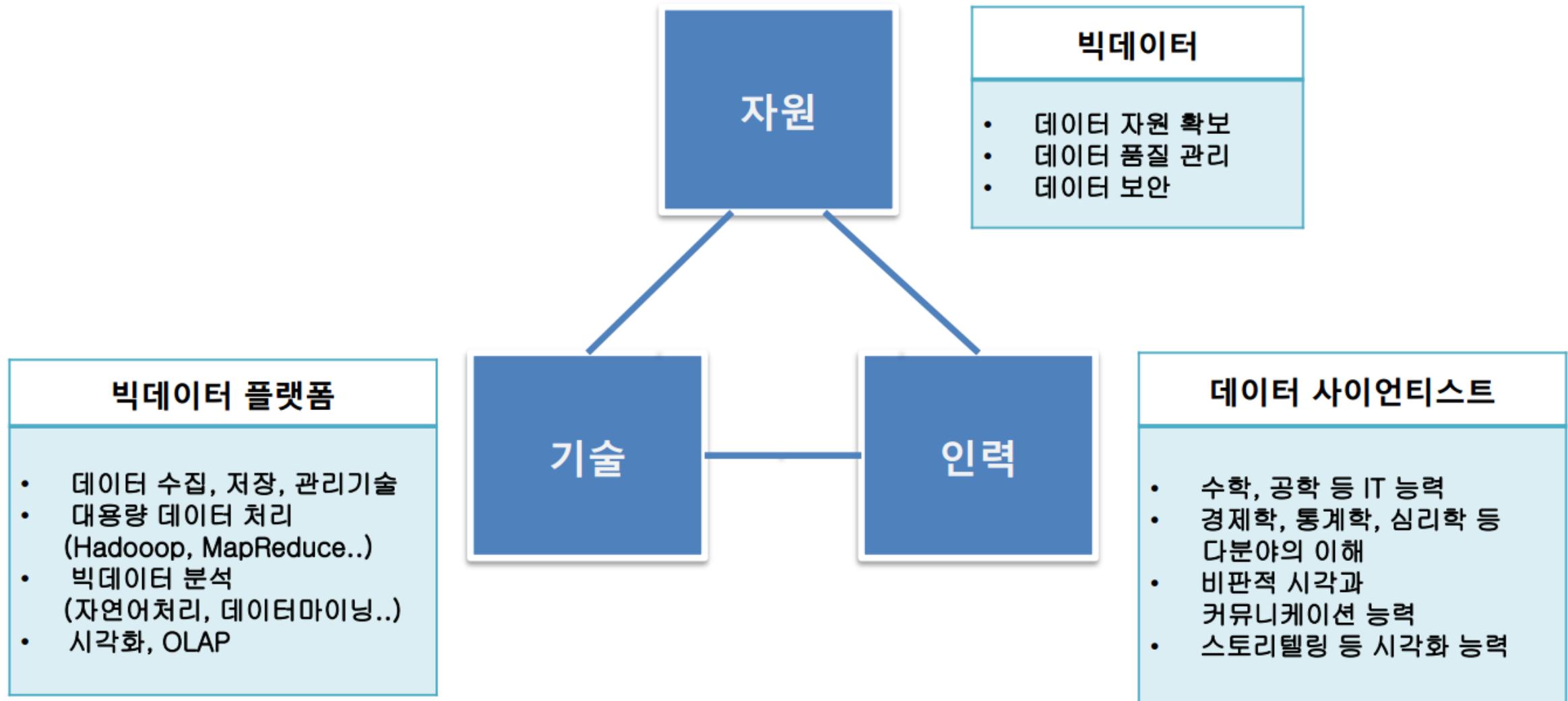
- 기존에 파악하기 어려웠던 문제점을 발견하고 그 해결점을 모색

# Big Data의 3대 요소

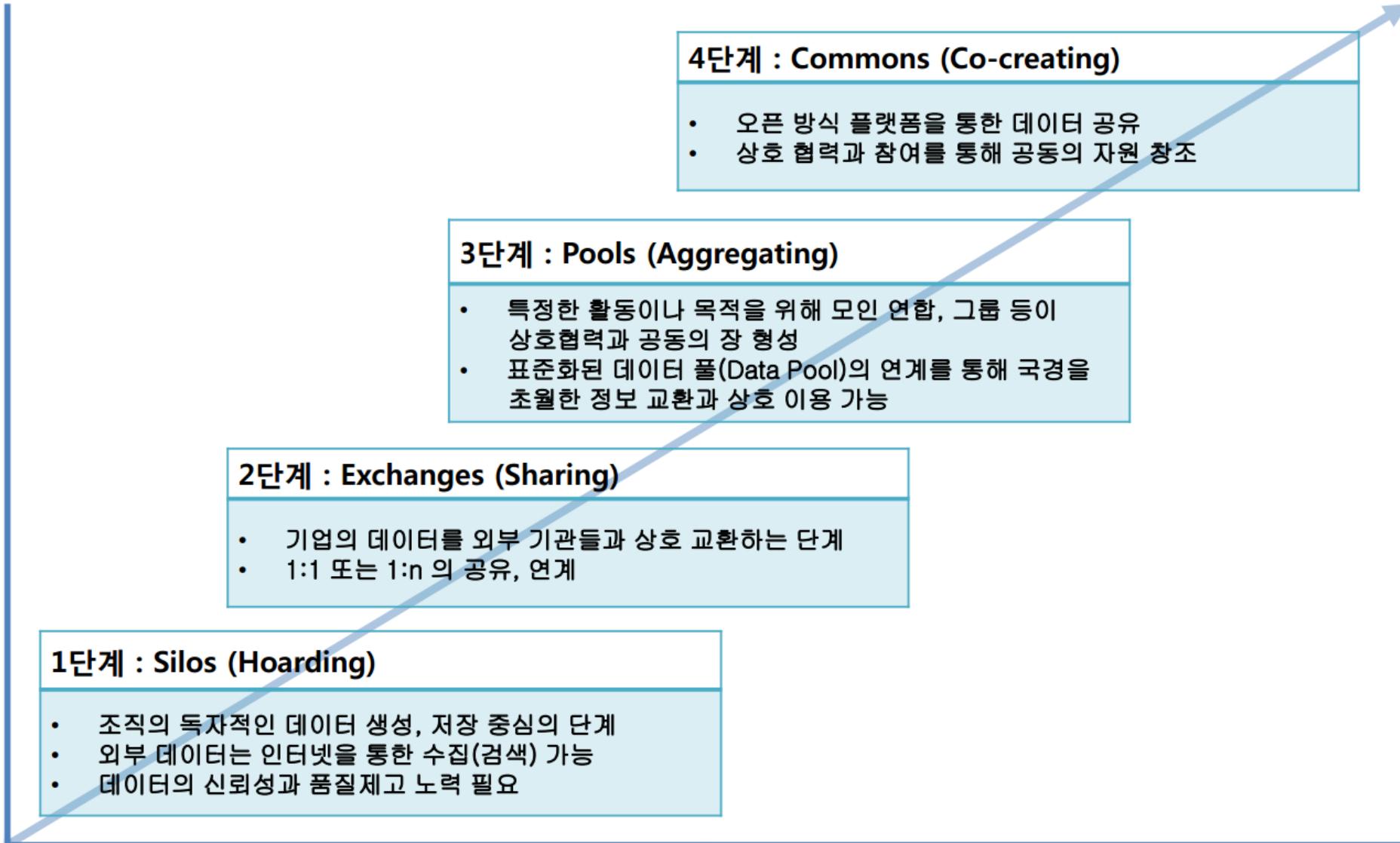


출처: 한국정보화진흥원, "성공적인 빅데이터 활용을 위한 3대요소".

# Big Data의 3대 요소 (Cont.)



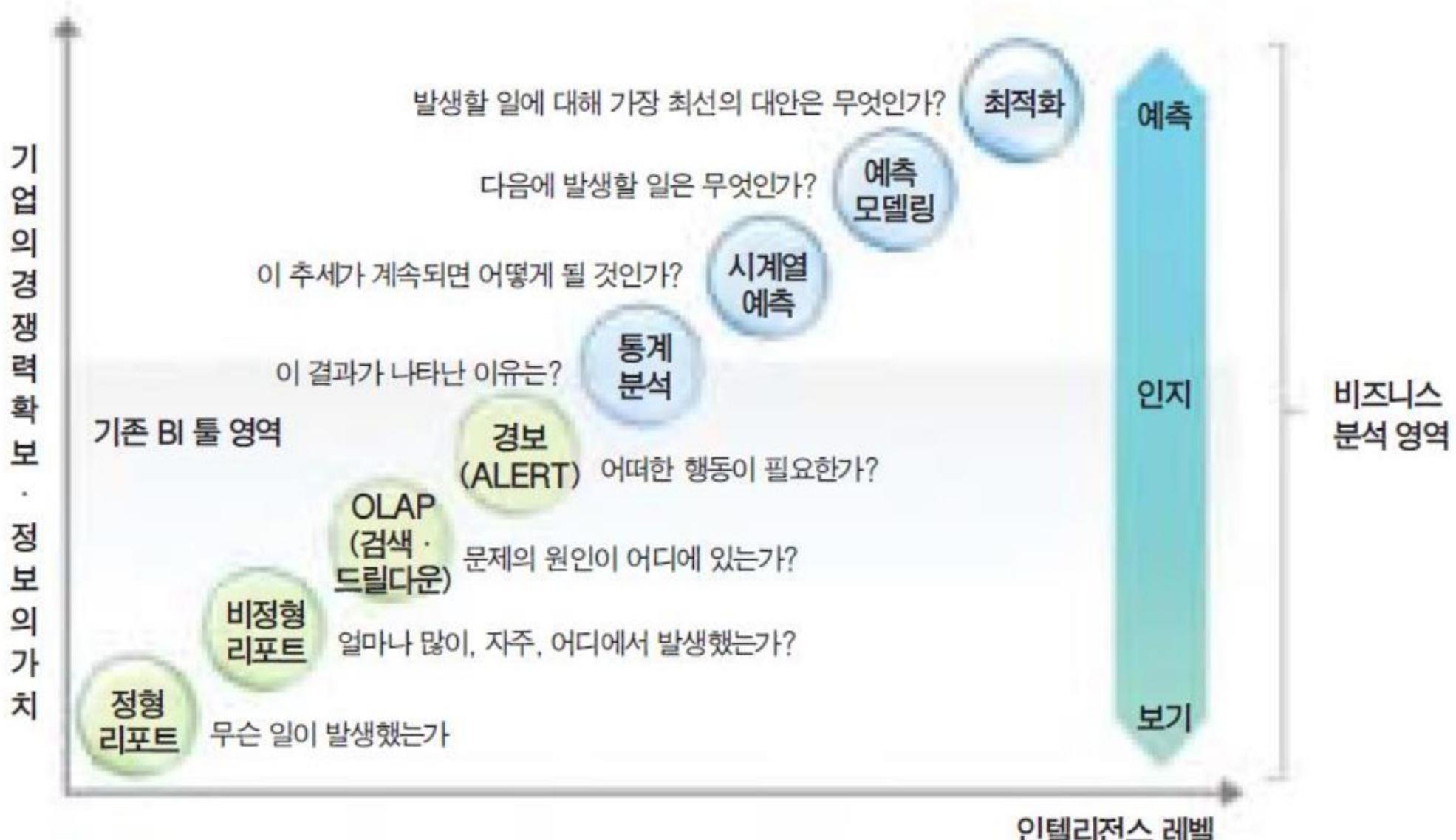
# Big Data의 3대 요소 (Cont.) – 자원 확보



# Big Data의 3대 요소 (Cont.) – 분석 기술

- BI(Business Intelligence)에서 BA(Business Analytics)로
    - BI는 신속하고 정확한 비즈니스 의사결정을 위해 사용하는 데이터의 접근, 수집, 보관, 분석 등의 애플리케이션과 기술의 집합
    - 최근 BI에서 BA로 진화하고 있으며, 데이터의 생성부터 폐기까지 전사적인 범위에서 기업의 미래를 예측
    - BI가 OLAP(Online Analytical Processing) 도구라면 BA는 ETL, DI/DQ, MDM, 분석/예측/최적화 기술이 통합된 것
- # DI/DQ(Data Integration / Data Quality) : 데이터 통합 및 품질 관리  
# MDM(Master Data Management) : 마스터 데이터 관리

# Big Data의 3대 요소 (Cont.) – 분석 기술



\*출처 : 데이터넷, 급부상하는 '비즈니스분석'

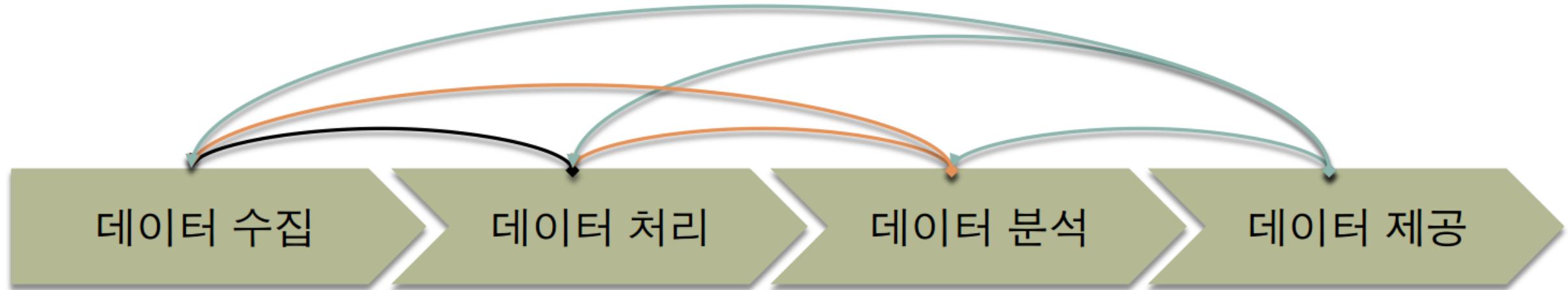
# Big Data의 3대 요소 (Cont.) – 인력

- 빅데이터 시대의 연금술사, 데이터 사이언티스트
  - 대규모 데이터 속에서 숨겨진 정보를 찾아내는 데이터 사이언티스트
    - 이베이는 고객 데이터를 분석하고 가공하는 일을 맡고 있는 직원수가 5000여명에 이를
    - EMC는 경제학, 통계학, 심리학 등을 전공한 박사급 인재들로 구성된 애널리틱스 랩을 운영중
    - IBM은 사내에 200명 이상의 수학자들이 '분석학'을 집중 연구하고 500개 이상의 특허를 취득
  - 데이터 사이언티스트 인재가 전세계적으로 부족
    - 미국에서는 2018년까지 14만~19만명의 전문가가 부족할 것으로 예측
    - 수요는 급증할 것으로 예상되며, 21세기 유망직업 중 하나로 부각됨.
  - 데이터 사이언티스트의 자질 6개
    - 기본 자질은 (1)수학 과 (2)공학 능력
    - 데이터를 분석에서 가설을 세우거나 검증하는데 필요한 (3)비판적 시각 과 이를 잘 작성할 수 있는 (4)글쓰기 능력
    - 다른 사람에게 잘 전달할 수 있는 (5) 대화 능력
    - (6)호기심과 개인의 행복 도 중요함

# Big Data의 난관



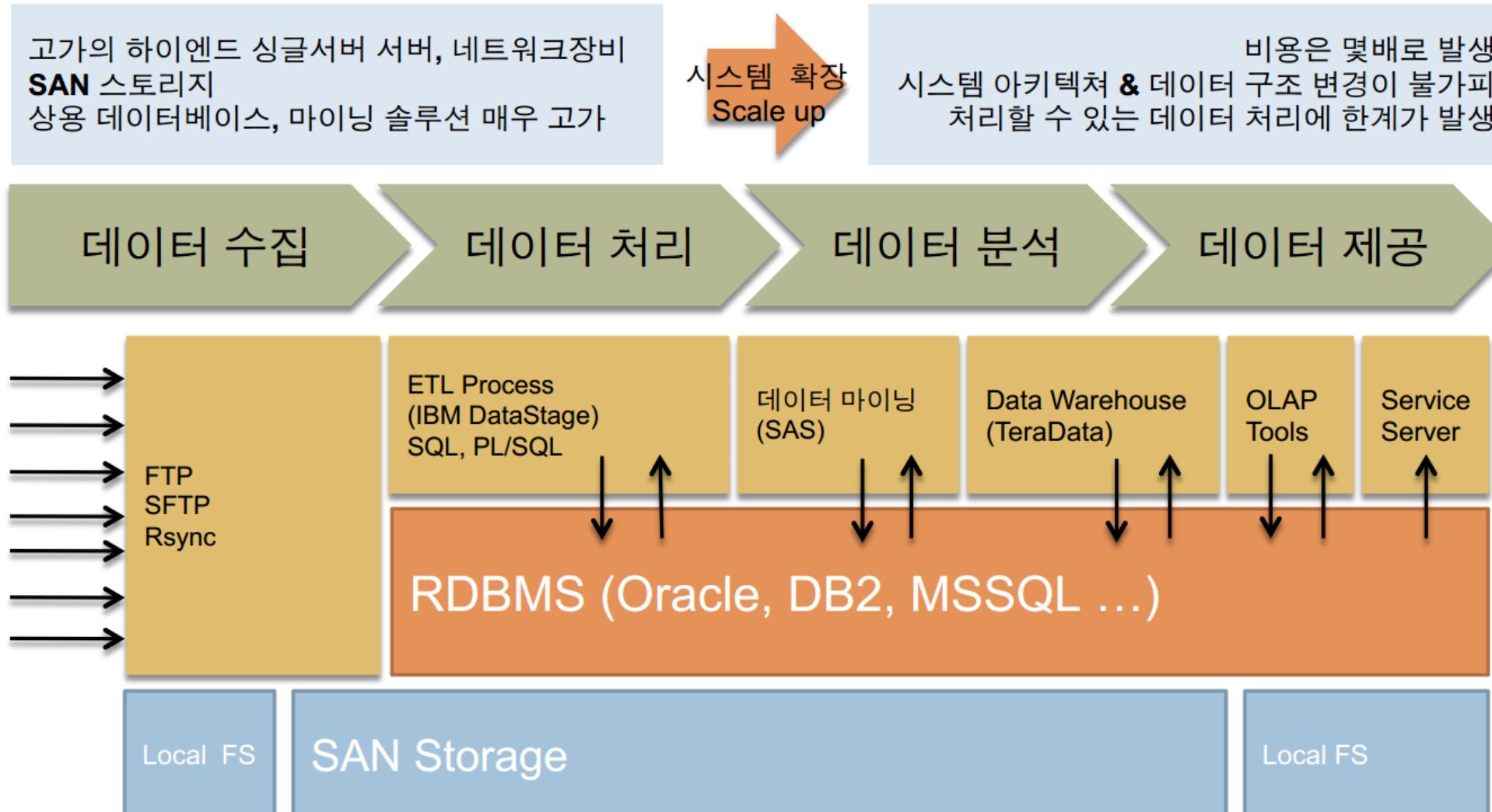
# Big Data Lifecycle



- |                            |            |           |                     |
|----------------------------|------------|-----------|---------------------|
| □ 데이터 연동                   | □ 데이터 클리닝  | □ 모델 검증   | □ 보고서               |
| □ ftp, sftp, rcp,<br>rsync | □ 데이터 요약   | □ 데이터 마이닝 | □ 데이터 시각화           |
| □ 데이터 변환                   | □ 데이터 기초통계 | □ 텍스트 마이닝 | □ 서비스 데이터           |
|                            | □ 데이터 탐색   |           | ■ 상품 추천<br>■ 유사 아이템 |

# Legacy Data Platform

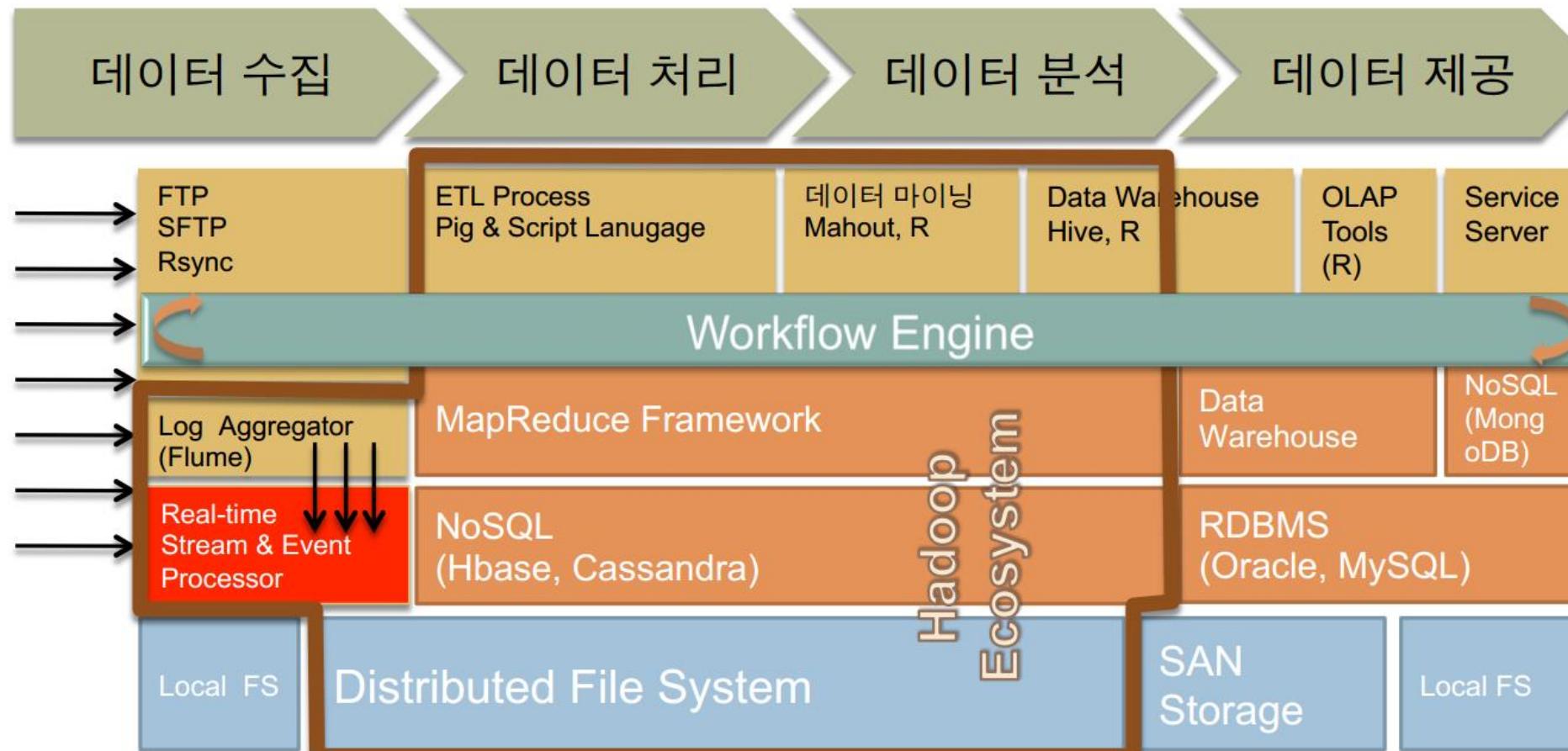
DBMS 을 기반으로하는 데이터 플로우 , Scale-up 구조의 아키텍처



# Big Data Platform

다양한 데이터 소스/데이터 프로세싱 프레임워크를 수용할 수 있는 워크플로우, 통합 관리 시스템 필요

- 대용량 데이터 저장과 대용량 데이터 분석 시스템에는 **Hadoop** 및 오픈소스기반의 **NoSQL**, **R**을 적극 활용
- 최종 분석 결과 제공을 위한 **OLTP** 기반 리포트 시스템 / 서비스 데이터들은 기존의 기술을 적극 활용
- **Hadoop** 및 오픈소스 기반의 시스템들은 저가의 범용서버 및 네트워크위치를 활용해서 클러스터를 구성



# Big Data Platform 환경 비교

## 기존 데이터 플랫폼 환경

- Single Machine
  - ▣ Multi-core (> 16 cores)
  - ▣ Scale-up
  - ▣ High Price H/W, S/W
  - ▣ SAN Storage
- MS Windows, AIX, HP-UX
- Commercial Solutions
- IBM, Oracle, Microsoft
- RDBMS
- TeraData, Exadata, Netizza
- SAS, SPSS

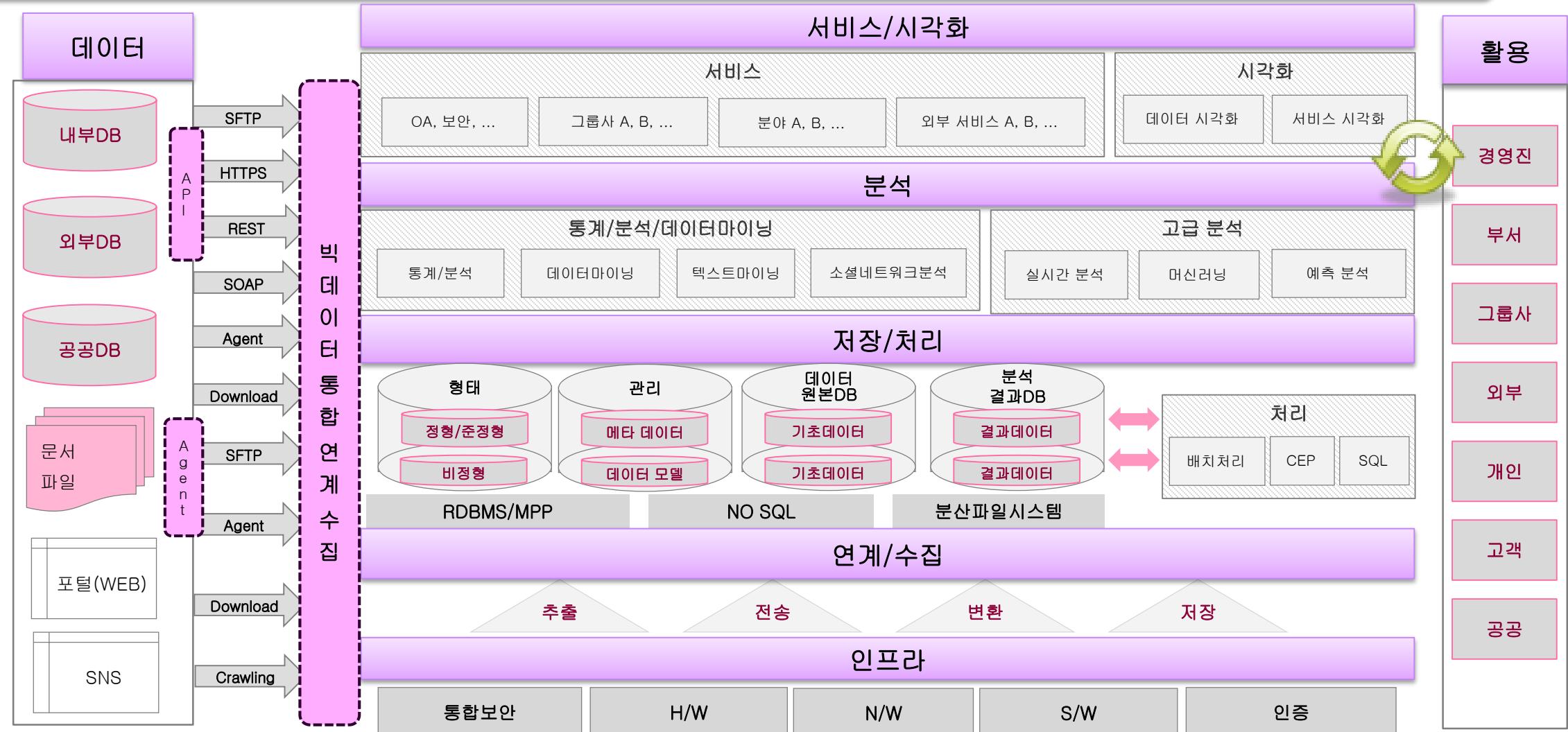
## 빅데이터 플랫폼 환경

- Multiple Machine
  - ▣ Commodity H/W, N/W
  - ▣ Scale-out
  - ▣ Low Cost
  - ▣ Distributed File System
- Linux
- Open Source
- Yahoo!, Facebook, Twitter
- NoSQL
- Hadoop, Hive, Pig
- R, Mahout

# 빅데이터 플랫폼

빅데이터 수집/분석/서비스를 위한 목표 플랫폼

## 빅데이터 플랫폼

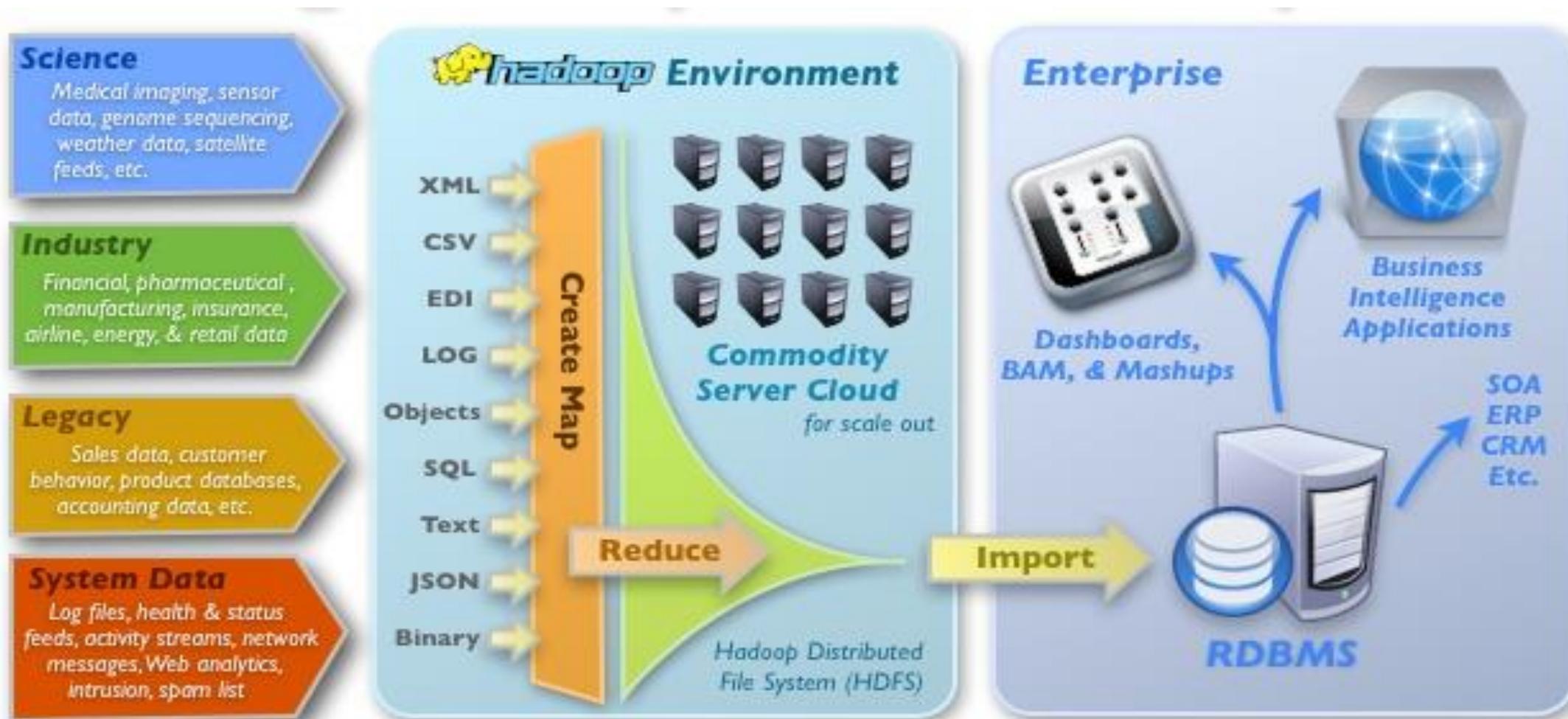


# 빅데이터 플랫폼 - 예시

국내 A 업체의 빅데이터 아키텍처 및 소프트웨어 구성도(우측의 S/W는 오픈소스 위주로만 표시함)



# 전형적인 빅데이터 처리 과정



# 필요한 SW 기술

데이터

수집

저장

분석

시각화

서비스

## 수집(ETL) → Flume

Load Balancing : Collector 분산, 노드 부하 모니터링

Fault Tolerance : 노드장애시 다른노드로 전환

Network Traffic 분산

다수의 Collector에 Traffic 분산 & local Disk 저장

→ Hadoop Network & Hacked Together To HDFS

Fire & Forget Agent 서버 실패시 재전송, End to End 전송 보장

Management

중앙관리 시스템 : N Agent & N Collect Node

## 저장 → HDFS, Hive, HBASE

HDFS : 분산파일시스템

Hive : HDFS 기반의 Database, Data는 HDFS에 저장, 스키마정보는 Mysql에 저장, HiveQL ->SQL Query로 분석 (select,groupby,join,union..)

HBASE : NoSQL for Hadoop, Data는 HDFS에 저장, 스키마정보는 Hbase Region Server에서 관리, 노드당 초당 수만개 Rows Insert, 수십만 Rows Read & Scan

## 분석 → MapReduce, HiveQL, Pig, R

MapReduce : 분산 병렬 처리 시스템

HiveQL : SQL Query Language & Dataware House Mysql Query와 유사, HDFS와 HBASE에 저장된 Data 분석언어

Pig : Pig Latin Script Language

Data & Flow 기반의 분산병렬 처리 언어, Hive와 동일하게 HDFS와 HBASE에 저장된 Data 분석

R & RHadoop Package

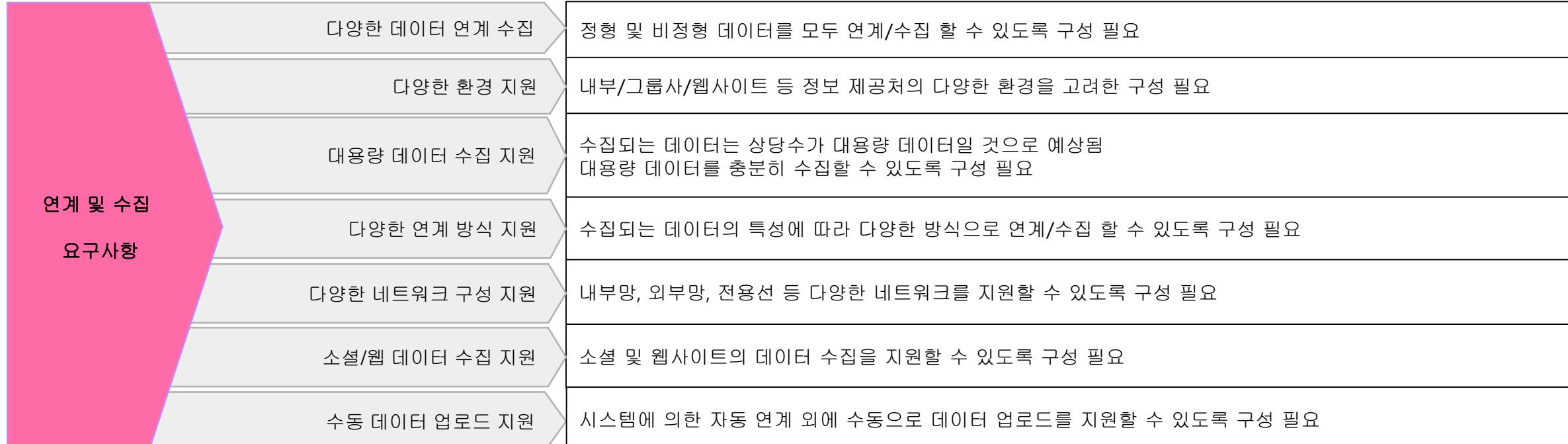
RHadoop 패키지에서 HDFS에 저장된 데이터를 Input으로 해서 통계분석 처리



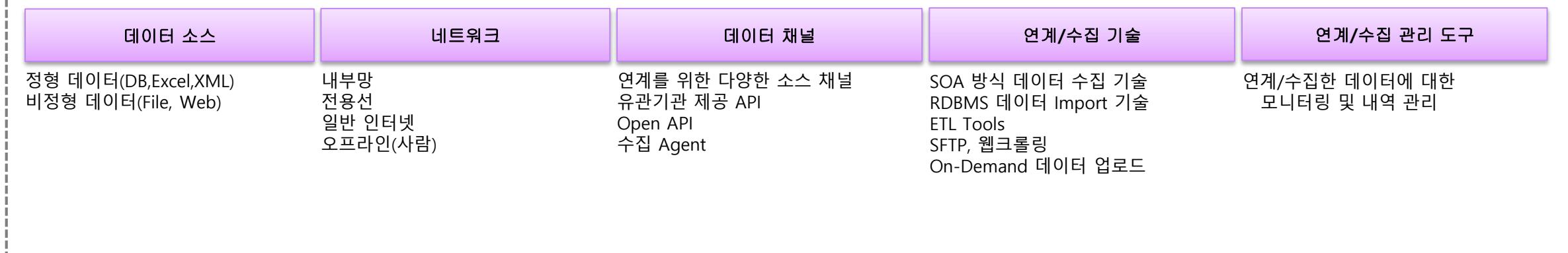
다양한 분석솔루션,  
시각화 도구를 통한  
대시보드 구성 및 보고서  
기능 구현



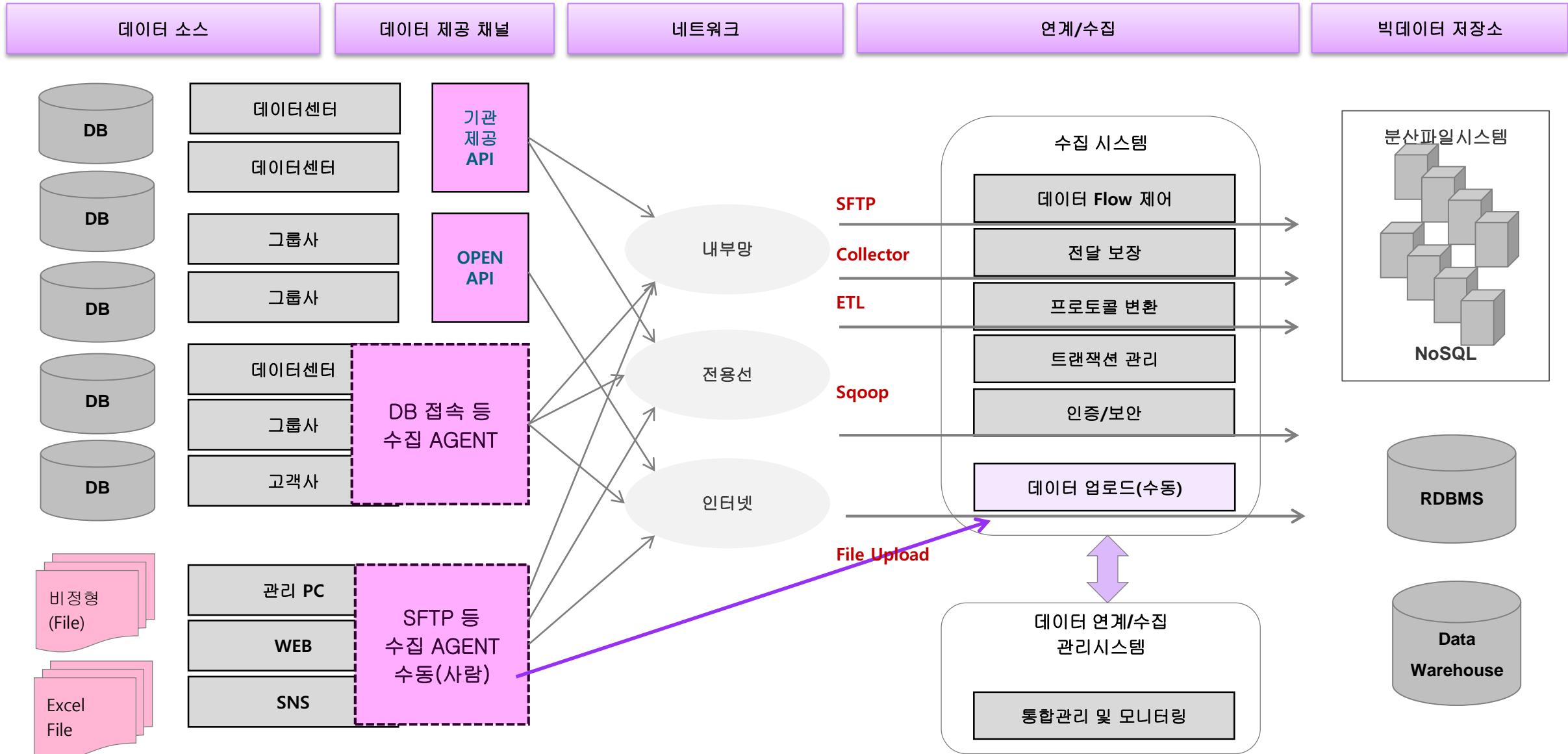
# 연계 및 수집 Layer



## 연계 및 수집 구성 요소

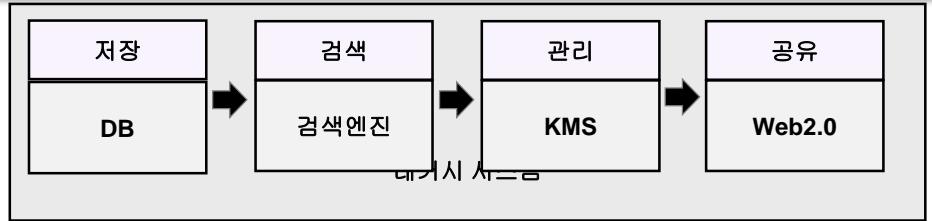


# 데이터 연계 및 수집 관련 구성 요소 및 Flow



# 분석 및 시각화 Layer

## 빅데이터 분석을 통한 새로운 가치창출



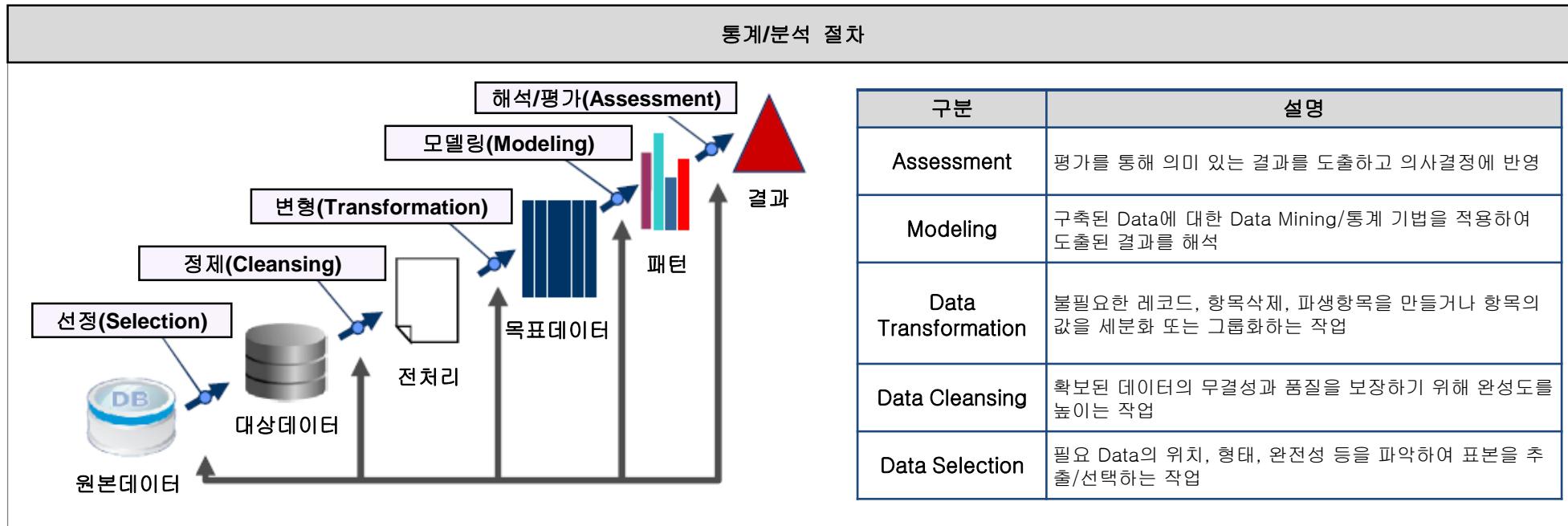
## 시각화 특징

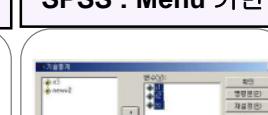
컴퓨터 스크린을 통해 제공되며, 이것은 곧 수작업이 아닌 컴퓨팅 시스템을 활용해야 함  
컴퓨팅 시스템을 활용하여 데이터의 필터링, 드릴다운, 세부 내용의 심화 분석 등 인터랙티브한 기능을 통한 자유로운 데이터 조작이 가능  
시각적 표현에 사용된 요소의 물리적인 위치, 길이, 모양, 색상, 크기 등을 제공함으로써 시각화 이외의 방법으로는 발견하기 어려운 인사이트를 제공  
일정한 형태, 색상을 활용하여 추상적인 데이터의 인지적인 의미를 쉽게 파악  
인터랙티브한 정보를 통해 인간의 기억을 지원하여 데이터를 더 쉽게 이해할 수 있는 길을 제시

## 주요 분석 기법

텍스트 마이닝 Text Mining	자연어 처리 Natural Language Processing 기술을 사용해 인간의 언어로 쓰인 비정형 텍스트에서 유용한 정보를 추출하거나 다른 데이터와의 연계성을 파악하며, 분류나 군집화 등 빅데이터에 숨겨진 의미 있는 정보를 발견하는 것
웹 마이닝 Web Mining	인터넷에서 수집한 정보를 데이터 마이닝 기법으로 분석하는 것
오피니언 마이닝 Opinion Mining: 평판 분석	<ul style="list-style-type: none"><li>다양한 온라인 뉴스와 소셜 미디어 코멘트, 사용자가 만든 콘텐츠에서 표현된 의견을 추출 · 분류 · 이해하고 자산화하는 컴퓨팅 기술</li><li>텍스트 속의 감성과 감동, 여러 가지 감정 상태를 식별하려고 감성 분석 사용</li><li>마케팅에서는 버즈 Buzz: 일소문 분석이라고도 함</li></ul>
리얼리티 마이닝 Reality Mining	<ul style="list-style-type: none"><li>휴대폰 등 기기를 사용하여 인간관계와 행동 양태 등을 추론하는 것</li><li>통화량, 통화 위치, 통화 상태, 대상, 내용 등을 분석하여 사용자의 인간관계, 행동 특성 등 정보를 찾아냄</li></ul>
소셜 네트워크 분석 Social Network Analysis	수학의 그래프 이론 Graph Theory를 바탕으로 소셜 네트워크 서비스에서 소셜 네트워크 연결 구조와 연결 강도를 분석하여 사용자의 명성 및 영향력을 측정하는 것
분류 Classification	<ul style="list-style-type: none"><li>미리 알려진 클래스들로 구분되는 훈련 데이터군 Group을 학습시켜 새로 추가되는 데이터가 속할 만한 데이터군을 찾는 지도 학습 Supervised Learning 방법</li><li>가장 대표적인 방법으로 KNN K-Nearest Neighbor이 있음</li></ul>
군집화 Clustering	<ul style="list-style-type: none"><li>특성이 비슷한 데이터를 합쳐 군 Group으로 분류하는 학습 방법</li><li>분류와 달리 훈련 데이터군을 이용하지 않기 때문에 비지도 학습 Unsupervised Learning 방법</li><li>트위터에서 주로 사진/카메라를 논의하는 사용자군과 게임에 관심 있는 사용자군 등 관심 사나 취미에 따라 분류</li></ul>
기계 학습 Machine Learning	<ul style="list-style-type: none"><li>인공지능 분야에서 인간의 학습을 모델링한 것</li><li>컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하여 수신한 이메일의 스팸 여부를 판단할 수 있도록 훈련</li><li>결정 트리 Decision Tree 등 기호적 학습, 신경망이나 유전자 알고리즘 등 비기호적 학습, 베이지안 Bayesian이나 은닉 마코프 Hidden Markov 등 확률적 학습 등 다양한 기법이 있음</li></ul>
감성 분석 Sentiment Analysis	문장의 의미를 파악하여 글의 내용에 긍정/부정, 좋음/나쁨을 분류하거나 만족/불만족 강도를 지수화. 그런 다음 이 지수를 이용하여 고객의 감성 트렌드를 시계열적으로 분석하고 고객 감성 변화에 기업의 신속한 대응 및 부정적인 의견의 확산을 방지하는 데 활용

# 분석 및 시각화 Layer



주요 통계 도구		
R : Interpreter Language	SAS : Procedure 기반	SPSS : Menu 기반
<pre>&gt; Tot = 0 &gt; for(l in 1:10) {   Tot = tot + 1 } &gt; print(Tot) [1] 55 &gt; sum(1:10) [1] 55</pre>	<pre>PROC FREQ OPTIONS; TABLES request/OPTIONS; WEIGHT variable; BY variable;</pre>	

주요 데이터마이닝 기법	
데이터 마이닝	<ul style="list-style-type: none"> <li>대용량의 데이터베이스에 있는 데이터로부터 패턴인식, 통계적 기법,</li> <li>다양한 알고리즘으로 숨겨져 있는 데이터간의 상호 관련성 등을 추출</li> </ul>
텍스트 마이닝	<ul style="list-style-type: none"> <li>텍스트 마이닝은 텍스트 기반의 데이터로부터 새로운 정보를 발견할 수 있도록 정 보 검색, 추출, 체계화, 분석을 모두 아우르는 <b>Text-processing</b> 기술</li> </ul>
평판분석	<ul style="list-style-type: none"> <li>텍스트 마이닝의 관련 분야로는 오피니언 마이닝, 혹은 평판분석(<b>Sentiment Analysis</b>)라고 불리는 기술로 소셜미디어 등의 정형/비정형 텍스트의 긍정, 부정, 중립의 선호도(작자의 의견이나 감정 등)를 판별하는 기술</li> </ul>
쇼셜웹 분석	<ul style="list-style-type: none"> <li>대용량 소셜미디어를 언어분석 기반 처리를 통해 이슈를 탐지하고, 시간의 경과에 따라 유통되는 이슈의 전체과정을 모니터링하고 추이를 분석하는 기술</li> </ul>
클러스터 분석	<ul style="list-style-type: none"> <li>다면하는 데이터 간의 유사도를 정의하고 각 데이터 간의 거리를 구하고 서로의 거 리가 가까운 것부터 순서대로 합쳐가는 방법</li> </ul>

# 분석 및 시각화 Layer

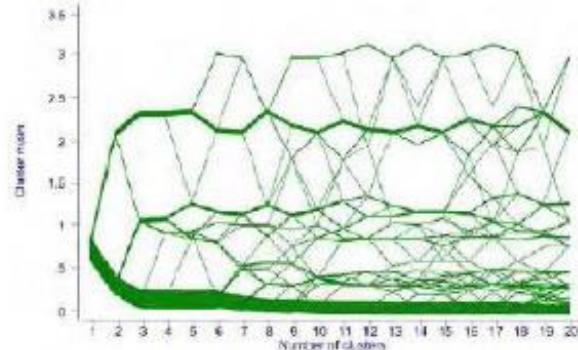
## Spatial information flow

- 특정 공간 안에서 정보의 흐름을 보여줌.
- 흐름이 많을 수록 링의 크기가 커짐



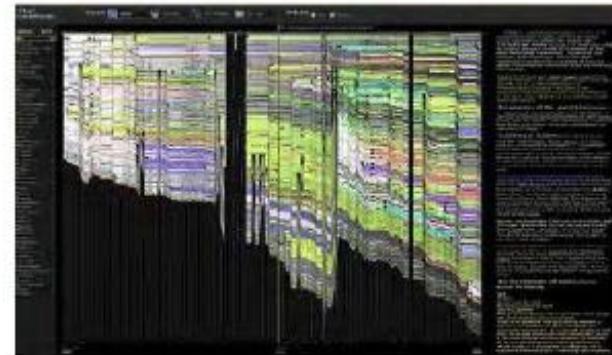
## Clustergram

- 클러스터 개수가 늘어남에 따라 각각의 데이터 셋 부분들에 클러스터가 어떻게 할당되는지를 보여주는 집약 분석 기법



## History flow

- 위키피디아의 경우 처럼 다수의 저자들이 문서를 수정하면서 문서가 변화된 양상을 기록한 것



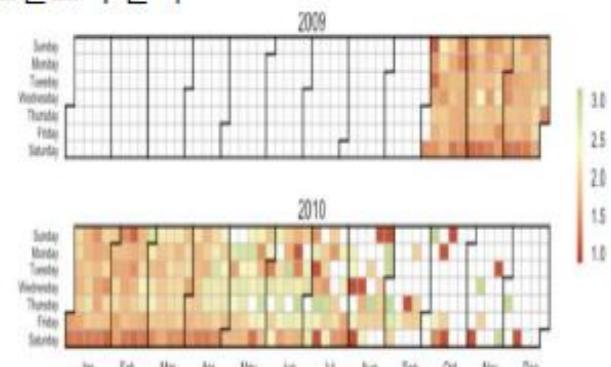
## Facebook transaction

- Facebook 사용자의 활동을 정보의 흐름과 빈도로 표시 (지역별 사용 정도를 일목요연하게 보여 줌)



## Heat Map

- Orbitz에서 분석한 월별 호텔 투숙 기간 트렌드의 분석



## R

- 오픈소스 기반의 통계 분석 및 시각화 제공

