



# How to Manage Data for Successful Business Intelligence

---

**Kyungwon Kim**

Assistant Professor  
Department of International Trade  
College of Global Political Science and Economics  
Incheon National University

September 1, 2021

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 목표

- 1) 빅데이터의 도전과제는?
- 2) 비즈니스 인텔리전스 인프라?
- 3) 데이터의 관계성, 패턴, 추세를 파악하는 방향은?
- 4) 고객, 웹사이트, 디지털플랫폼과의 실시간 연결성은?
- 5) 성공적 데이터관리를 위한 정책과 품질은?

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 빅데이터의 도전과제는?

- 공급자들의 지불, 주문처리, 고객 모니터링, 직원 급여 지급 등 기본적 트랜잭션 파악 위해 **DB 사용**
- 업무의 효율적 수행 및 관리자/직원들의 더 나은 의사결정을 위한 정보를 제공하기 위해 **DB 필요**
- 어떤 제품이 가장 인기인지, 어떤 고객의 수익성이 가장 높은지 알기 위해선  
그 대답은 바로 **데이터에서 찾을 수 있음**

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

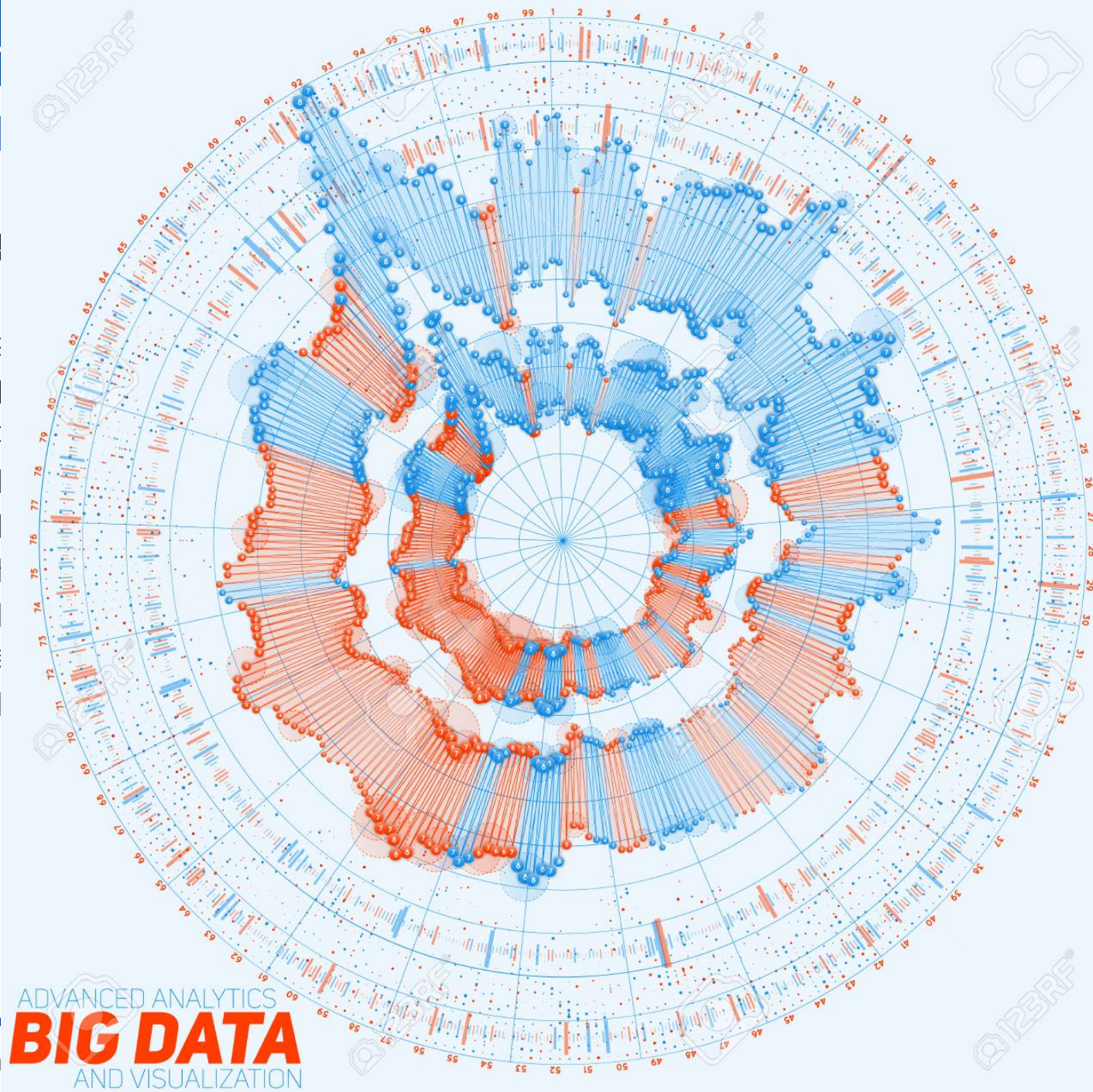
## ➤ 빅데이터의 도전과제는?

- 수집된 대부분의 데이터는 관계형 데이터베이스관리시스템의 행과 열에 쉽게 들어맞는 거래처리 데이터
- 이제 웹 트래픽, 이메일, 소셜 미디어 콘텐츠, 상태 메시지, 센서(스마트 측정기, 제조 장비, 전자 측정기 등에서 추출)로부터 자동으로 생성된 데이터나 전자적 거래 데이터들의 폭발적인 증가를 목격
  - > 비구조적이거나 반구조적일 수 있는데, 이런 경우 행과 열로 구성된 관계형 데이터베이스로 부적합
  - > **빅데이터(Big Data)**: 엄청난 양 때문에 일반적인 DBMS로는 포착/저장/분석하기 어려운 데이터 세트
    - 3V: 데이터의 엄청난 양(Volume), 광범위한(Variety), 데이터 원천 및 처리에 요구되는 속도(Velocity)
    - 빅데이터가 특정 수량을 의미하진 않지만, 보통은 페타바이트(Petabyte)나 엑사바이트(Exabyte) 수준
    - 빅데이터는 전통적인 데이터들에 비해 훨씬 많은 양과 훨씬 빠른 속도로 생성
      - . 어떤 제트 엔진은 단 30분 만에 10테라바이트의 데이터를 생성
      - . 비행기들의 비행 횟수가 하루에 2만 5,000회 이상
      - . 트위터를 통해 생성되는 데이터들은 하루에 8테라바이트
    - 국제데이터센터(International Data Center)에 따르면, 사용할 수 있는 데이터의 양은 2년마다 2배 이상

## ➤ 빅데이

- 수집된
- 이제 웹
- 등에서
- > 비구조
- > 빅데이
- 3V: V
- 빅데
- 빅데
- . 어
- . 비
- . 트
- 국제

데이터  
정기



ADVANCED ANALYTICS  
**BIG DATA**  
AND VISUALIZATION

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 빅데이터의 도전과제는?

- 기업들은 빅데이터에 많은 관심

- 빅데이터가 고객 행위, 기후 패턴, 금융시장에서의 행위, 또는 여타의 현상들에 대한 통찰을 제공해주는 잠재력
- 작은 데이터 세트에 비해 더 많은 패턴들과 더 흥미로운 이상 현상들을 파악할 수 있음
  - . 셔터스톡(Shutterstock)은 2,400만 개의 이미지를 저장하고 있는데, 매일 이미지를 10,000개씩 추가
  - . 구매 경험을 최적화 위해, 빅데이터 분석하여 웹사이트 방문자의 커서 위치와 구매 전 이미지 주변의 머문 시간 파악
  - . 정부들은 빅데이터를 활용하여 교통 흐름을 관리하고 범죄와 싸우고 있음

→ 데이터들로부터 비즈니스 가치를 도출하기 위해 전통적인 전사적 데이터들은 물론

비전통적인 데이터들까지도 관리하고 분석할 수 있는 새로운 기술들과 도구들이 필요

→ 조직들은 데이터에게 물어볼 질문과 빅데이터의 한계를 알아야 함

→ 빅데이터를 확보, 저장, 분석하는 데 많은 비용이 들 수 있고, 추출 정보다 반드시 도움이 되는 것은 아님

→ 빅데이터가 비즈니스를 위해 해결할 문제에 대해 명확하게 이해하는 것이 중요



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 비즈니스 인텔리전스 인프라?

- 비즈니스 인텔리전스를 위한 기업들의 최신 인프라는 엄청난 양의 반구조적·비구조적 빅데이터를 포함한 모든 다양한 유형의 데이터들로부터 유용한 정보를 생성해낼 수 있는 다수의 도구 보유
  - . 기업 전반에 걸친 운영 상태, 동향, 변화 등에 관해 간결하고 신뢰할 수 있는 정보가 필요하다고 가정
  - . 판매, 제조, 회계와 관련된 데이터들과 외부 인구통계학적 데이터나 경쟁사 데이터들이 필요할 수
  - . 갈수록 빅데이터를 사용할 필요성이 점점 더 커질 것
  - . 1) 데이터웨어하우스와 데이터마트, 2) 하둡, 3) 인메모리 컴퓨팅, 4) 분석 플랫폼 등
- 일부는 클라우드 서비스를 통해서도 사용 가능

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 비즈니스 인텔리전스 인프라?

### 1) 데이터웨어하우스와 데이터마트:

> **데이터웨어하우스(Data Warehouse):** 기업 전반의 현재/과거 데이터 저장하는 데이터베이스

- 20년간 기업 데이터를 분석하는 전통적인 도구로 데이터웨어하우스가 활용

(1) 우선 조직 내 다수의 운영시스템으로부터 현재 및 과거 데이터들을 추출

(2) 이 데이터를 외부 원천들로부터 나온 데이터들과 결합

(3) 부정확하거나 불완전한 데이터들을 정정

(4) 보고서와 분석을 위한 형태로 재구성

(5) 최종적으로 생성된 데이터를 데이터웨어하우스로 이전

- 데이터웨어하우스의 데이터는 필요한 누구라도 사용할 수 있지만 변경은 할 수 없음

- 특별하고 표준화된 쿼리 도구, 분석 도구, 그래피컬한 리포팅 기능들을 제공

> **데이터마트(Data Mart):** 데이터웨어하우스의 일부분이거나

매우 집중화된 조직 데이터의 일정 부분이거나, 특정 사용자 집단을 위한 개별 데이터베이스에 위치

- 비용과 시간적 효율을 위해 작고 분산된 데이터웨어하우스 구축

. 고객정보를 다루기 위해 마케팅 및 판매 데이터마트를 개발

. 도서 판매업체인 반스앤노블(Barnes & Noble)은 전사적 데이터웨어하우스를 구현하기 이전에,

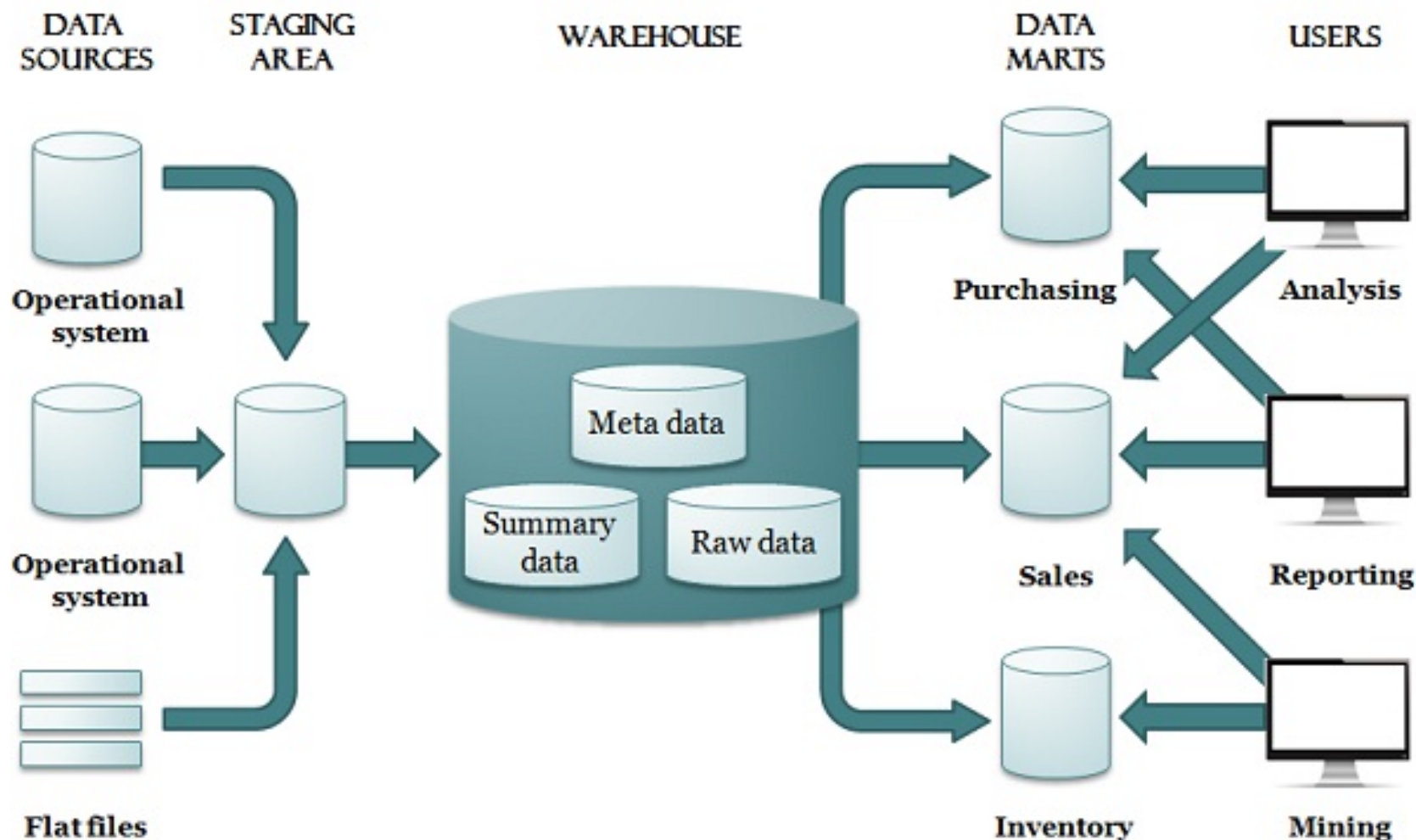
하나는 매장의 판매 시점 데이터, 다른 하나는 대학 서점들의 판매, 또 다른 하나는 온라인 판매 마트



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 비즈니스 인텔리전스 인프라?

### 1) 데이터웨어하우스와 데이터마트:



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 비즈니스 인텔리전스 인프라?

### 2) 하둡: 아파치 소프트웨어 재단이 관리하는 오픈소스 분산처리 소프트웨어 프레임워크

- 저렴한 컴퓨터들에 걸쳐 있는 방대한 양의 데이터들을 분산병행처리 방식으로 다룰 수 있게 해줌
- 대량의 비구조적이고 반구조적인 데이터뿐만 아니라 구조적 데이터까지도 포함
- 빅데이터 분석 문제를 작은 문제들로 분할, 수많은 값싼 컴퓨터 노드에 분산, 결과를 분석하기 쉬운 작은 데이터로 정리
- 핵심 서비스: 데이터 저장을 위한 하둡 분산파일시스템(Hadoop Distributed FileSystem, HDFS)과  
고성능 병행 데이터 처리를 위한 맵리듀스(MapReduce), 비관계형 데이터베이스(HBase)
  - . HDFS는 하나의 하둡 클러스터 내 수많은 파일시스템을 서로 연결하여 하나의 거대한 파일시스템으로 변환
  - . 맵리듀스는 구글의 맵리듀스 시스템에서 영감
  - . 구글의 맵리듀스는 거대한 데이터 세트의 작업을 분할하고, 그것들을 클러스터상의 여러 노드에 할당
  - . 비관계형 데이터베이스 HBase는 HDFS에 저장된 데이터들에 빠르게 접근할 수 있도록 해주고  
커다란 애플리케이션을 실시간으로 가동하기 위한 거래처리 플랫폼
- 하둡은 구조적인 거래처리 데이터, 페이스북/트위터의 글 같이 덜 구조적인 데이터, 웹서버의 로그처럼 복잡한 데이터,  
그리고 비구조적인 오디오 및 비디오 데이터 등과 같은 모든 유형의 데이터들의 방대한 양을 처리 가능
- 저렴한 서버 그룹에서 가동되며, 프로세서들은 필요에 따라 추가되거나 제거가 가능하기에 인프라 유연성 보유
- 비구조적이고 반구조적인 방대한 데이터들을 데이터웨어하우스에 이전시키기 전 준비/분석하기 위한 목적으로 하둡 사용
  - . 페이스북은 데이터를 거대 하둡 클러스터에 저장(100페타바이트 추정), 미 국회도서관 저장공간의 10,000배 이상 규모
  - . 야후는 하둡을 이용하여 사용자 행위를 추적함으로써 자사의 홈페이지를 사용자 관심사에 맞도록 수정
  - . 생명과학 연구회사 넥스트바이오(NextBio)는 하둡과 HBase를 이용하여 게놈 연구 수행 제약회사들을 위해 데이터 처리
  - . IBM, 휴렛팩커드, 오라클, MS 등 최고 데이터베이스 벤더들은 나름대로의 하둡 소프트웨어 유통망 기반 개발

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 비즈니스 인텔리전스 인프라?

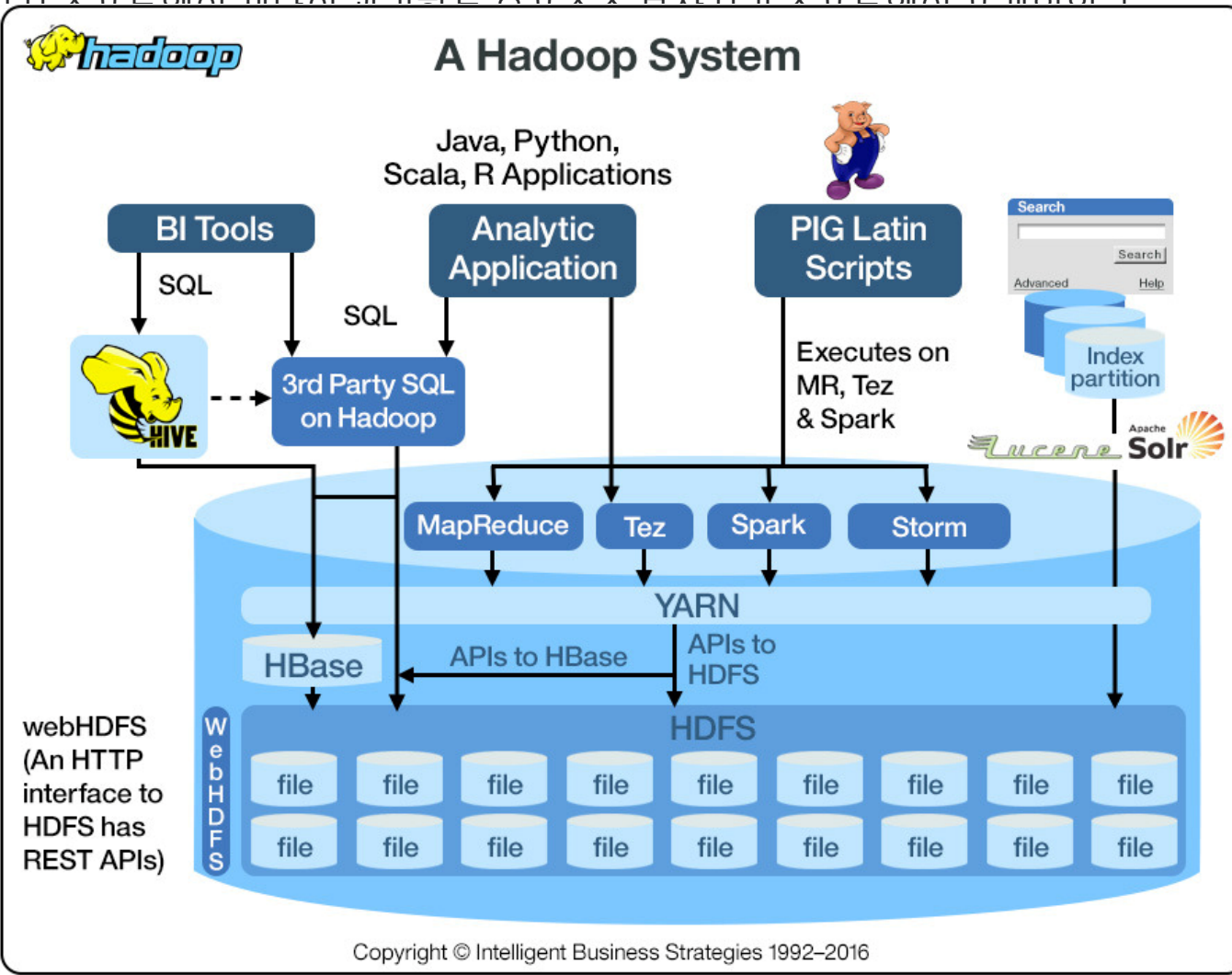
2) 하둡: 아파치 소프트웨어 재단의 관리하는 오픈 소스 분산 컴퓨팅 소프트웨어 프로젝트이다

- 저렴한 컴
- 대량의 비
- 빅데이터
- 핵심 서비

- . HDFS는
- . 매퍼듀
- . 구글의
- . 비관계
- 커다란

- 하둡은 구
- 그리고 비
- 저렴한 서
- 비구조적

- . 페이스
- . 야후는
- . 생명과
- . IBM, 휴



터로 정리

한 데이터,

유  
으로 하둡 사용  
배 이상 규모

데이터 처리

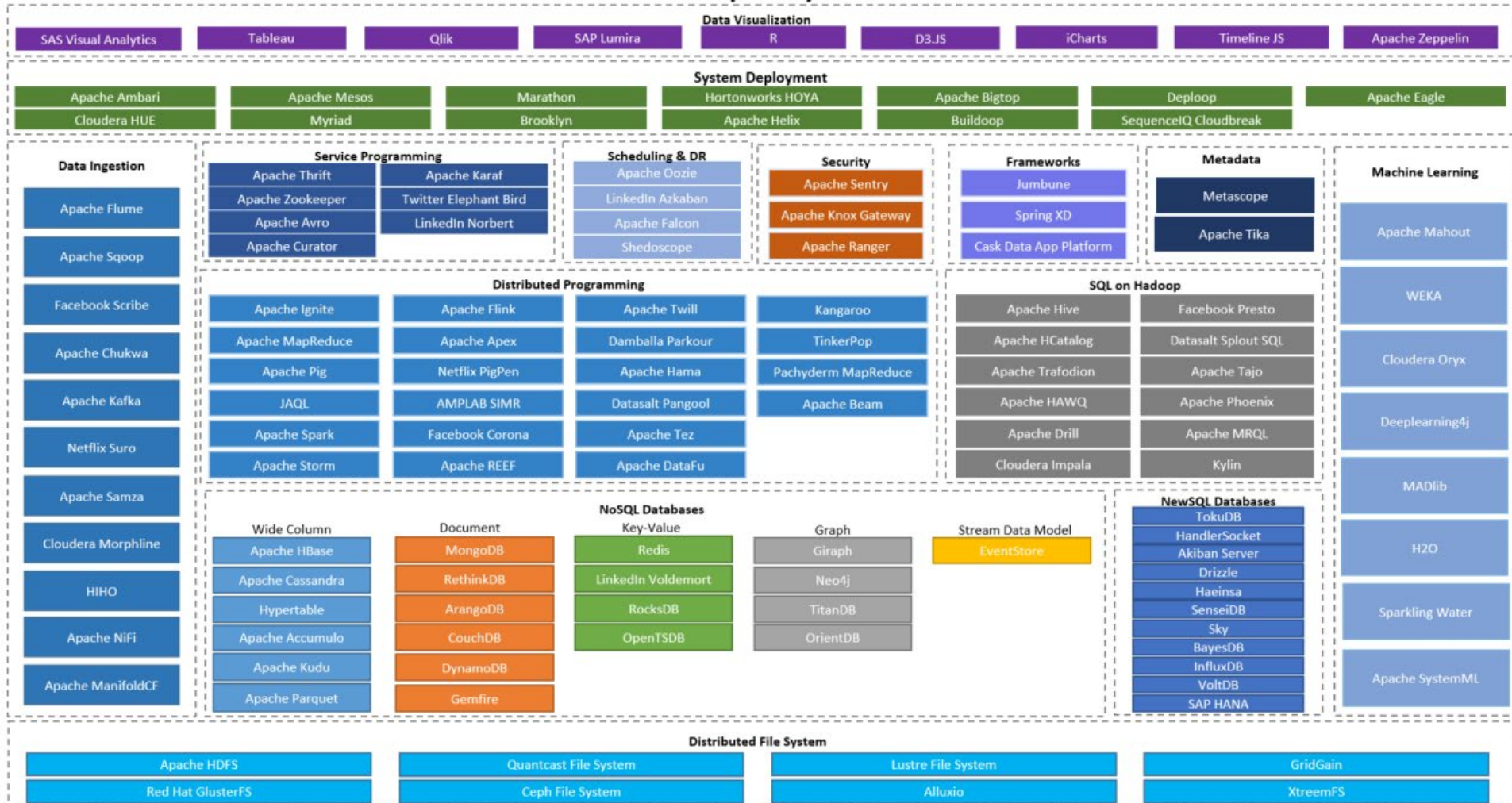
발

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 비즈니스 인텔리전스 인프라?

2) 하둡: 아파치 소프트웨어 재단이 관리하는 오픈소스 분산처리 소프트웨어 프레임워크

### Hadoop Ecosystem

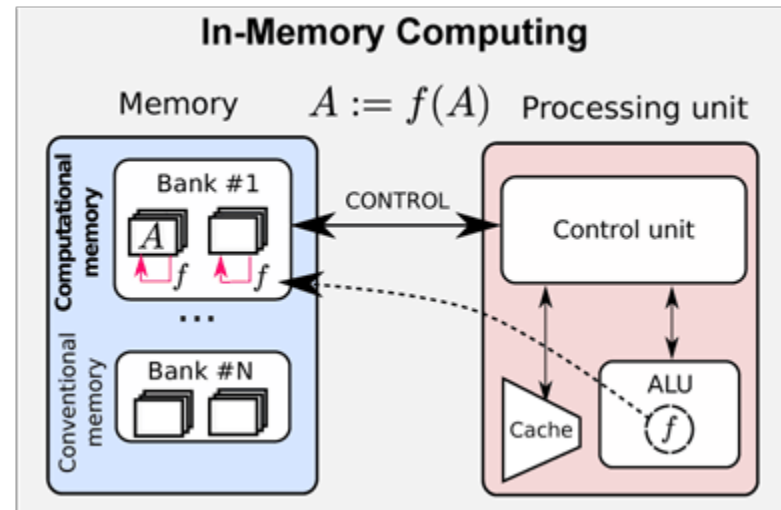
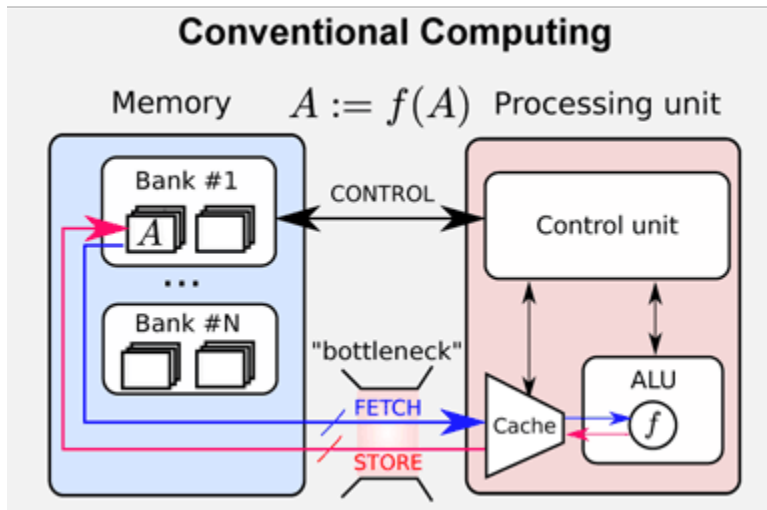


# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 비즈니스 인텔리전스 인프라?

### 3) 인메모리 컴퓨팅(In-Memory Computing):

- 기존의 DBMS는 디스크 저장시스템을 사용하지만, 인메모리 컴퓨팅은 주기억장치(RAM)에서 데이터 처리
- 주기억장치에 저장된 데이터에 접속인 전통적 디스크 기반 데이터 조회/처리에 걸리는 병목현상을 제거하고, 쿼리에 대한 응답 시간을 급격하게 줄임
- 데이터마트나 데이터웨어하우스의 규모에 이르는 매우 방대한 양의 데이터가 모두 메모리에 상주하는 것을 가능케 함
- 많은 시간 소요되는 복잡한 비즈니스 계산들이 단 몇 초 만에 처리, 심지어 휴대용 기기에서도 처리 가능
- 기업이 메모리의 활용을 최적화하고 비용은 떨어뜨리면서도 처리 성능을 향상
  - . SAP HANA, 오라클 데이터베이스 인메모리, 테라데이터 인텔리전트 메모리(Teradata Intelligent Memory)





# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 비즈니스 인텔리전스 인프라?

### 4) 분석 플랫폼(Analytic Platform):

- 관계형 기술과 대용량 데이터 특화 비관계형 기술 활용 전문화된 초고속 하드웨어-소프트웨어 융합 시스템
  - . IBM 퓨어데이터 시스템(PureDataSystem)은 데이터베이스, 서버, 저장 등의 요소들이 긴밀하게 통합
  - . 기존의 시스템에 비해 복잡한 분석 쿼리들을 10~100배 정도 더 빠르게 처리
  - . 인메모리 시스템들과 NoSQL 비관계형 데이터베이스관리시스템들도 포함하며 클라우드에서도 사용 가능

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 비즈니스 인텔리전스 인프라?

### 4) 분석 플랫폼(Analytic Platform):

- 관계형 기술과

**그림 6.13** 최신 비즈니스 인텔리전스 인프라

:템

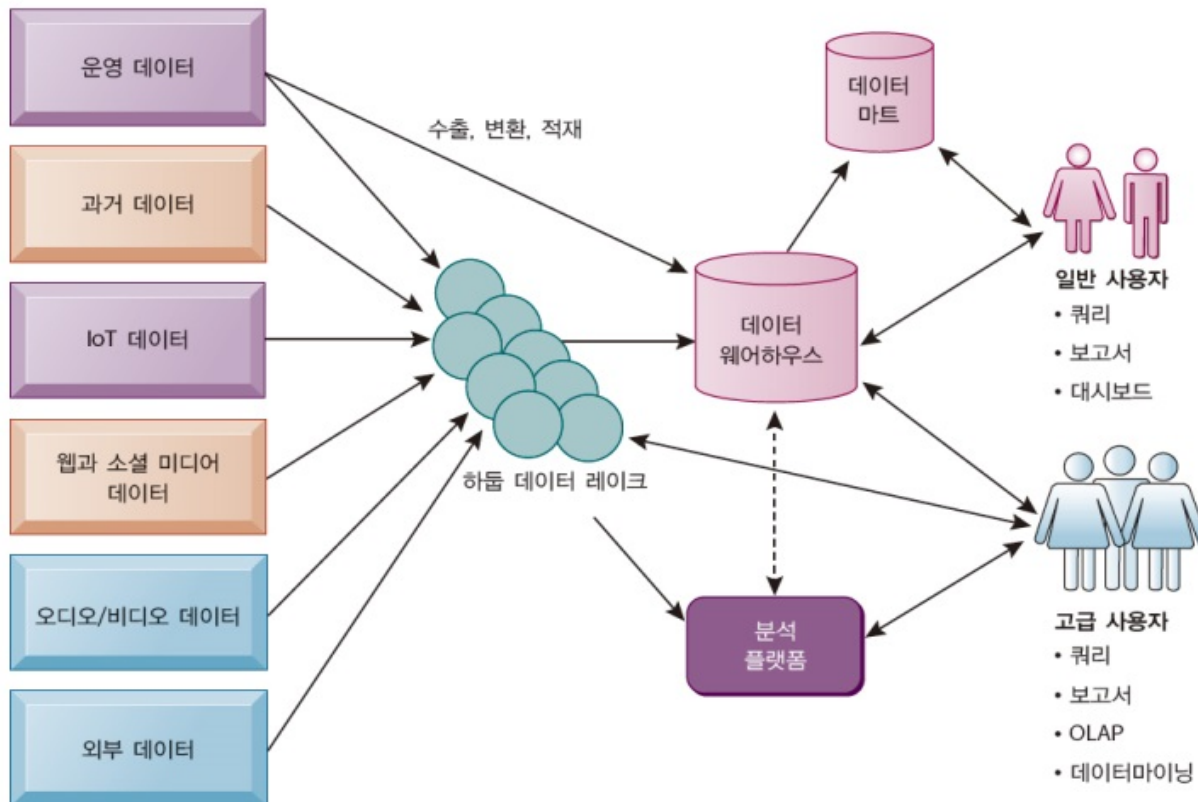
. IBM 퓨어데

. 기존의 시스

. 인메모리 시

최신 비즈니스 인텔리전스 인프라의 특징은 다양한 원천으로부터 제공되는 다양한 유형의 대용량 데이터들을 관리하고 분석할 수 있는 기능과 도구들을 가지고 있다는 점이다. 이러한 인프라는 일반 사용자들이 쉽게 사용할 수 있는 쿼리 및 보고서 생성 도구와 고급 사용자들을 위한 더 복잡하고 분석적인 분석 도구들을 포함하고 있다.

가능





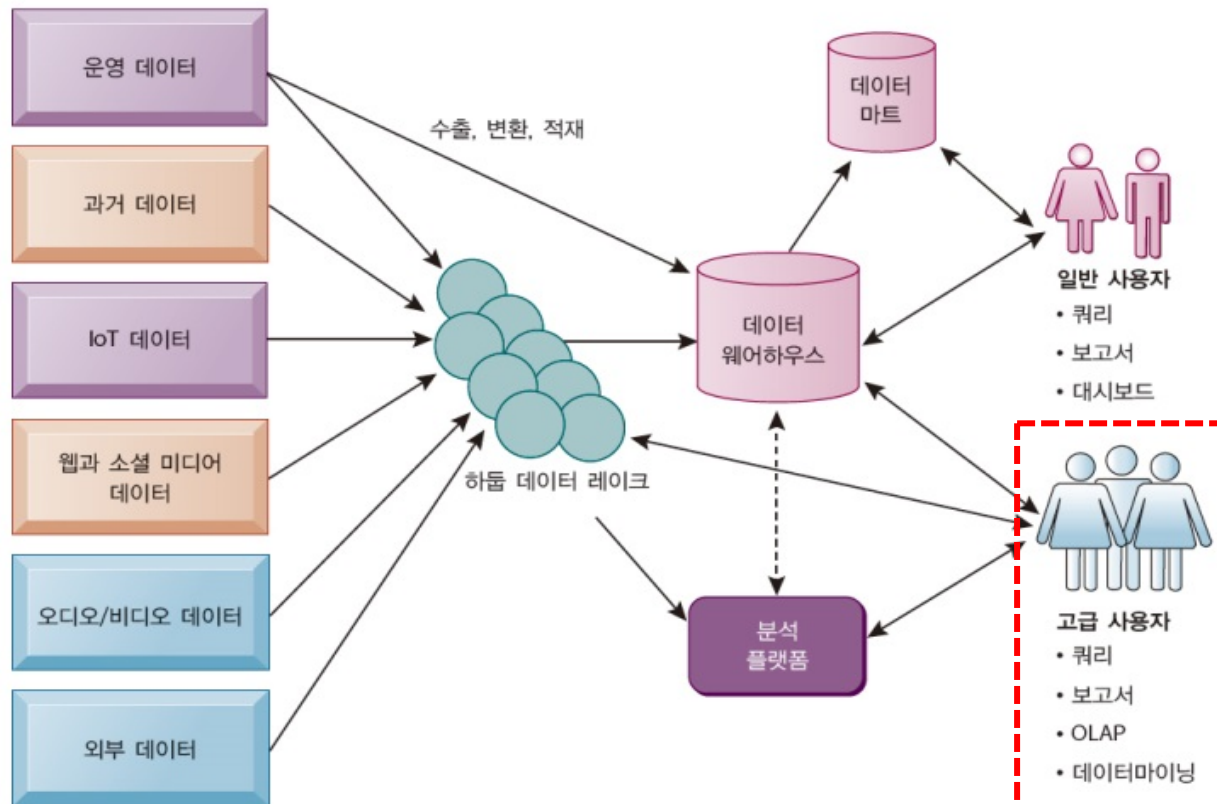
# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

1) 온라인분석처리, 2) 데이터마이닝, 3) 텍스트마이닝과 웹마이닝

**그림 6.13** 최신 비즈니스 인텔리전스 인프라

최신 비즈니스 인텔리전스 인프라의 특징은 다양한 원천으로부터 제공되는 다양한 유형의 대용량 데이터들을 관리하고 분석할 수 있는 기능과 도구들을 가지고 있다는 점이다. 이러한 인프라는 일반 사용자들이 쉽게 사용할 수 있는 쿼리 및 보고서 생성 도구와 고급 사용자들을 위한 더 복잡하고 분석적인 분석 도구들을 포함하고 있다.



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

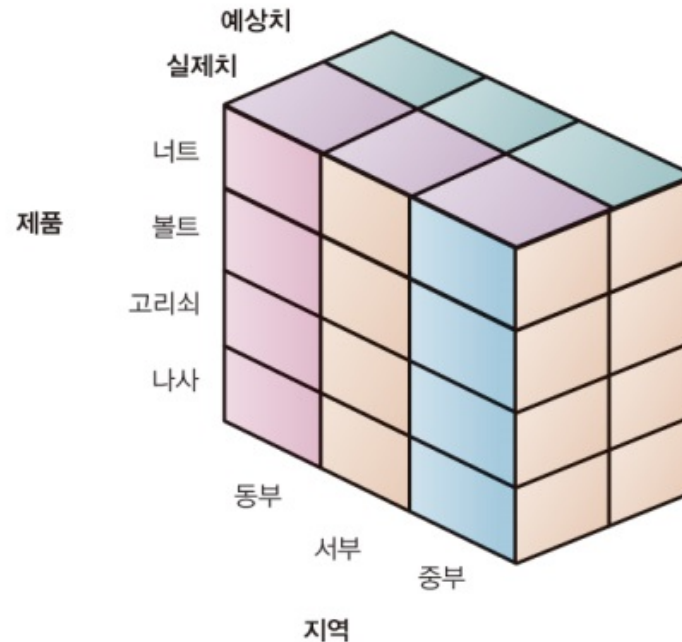
## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

### 1) 온라인분석처리(Online Analytical Processing, OLAP):

- 사용자들이 동일한 데이터를 여러 기준에서 다양하게 이용하는 다차원(Multidimensional) 데이터 분석방식 지원
- 정보에 대한 각 속성(제품, 가격, 비용, 지역, 기간) 별로 다른 상이한 차원의 정보를 쉽게 추출

**그림 6.14** 다차원 데이터 모델

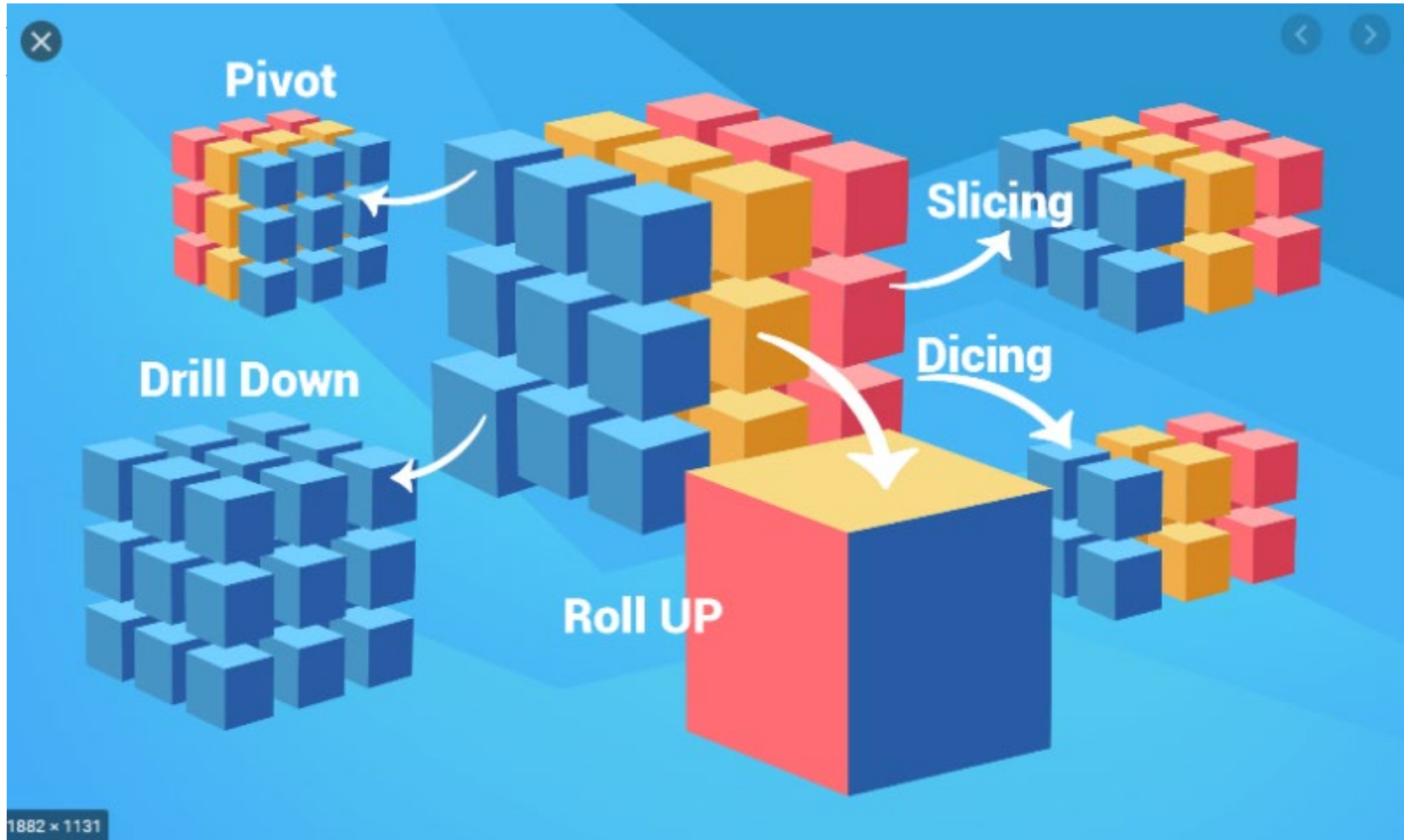
보이는 관점은 제품 대 지역이다. 여러분이 이 입방체를 90도 회전시키면 제품 대 실제 및 예상 판매량이 보일 것이다. 여러분이 이 입방체를 90도 다시 회전시키면, 지역 대 실제 및 예상 판매량을 볼 수 있다. 다른 관점들도 가능하다.



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

### 1) 온라인분석처리(Online Analytical Processing, OLAP):



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

### 2) 데이터마이닝(Data Mining):

- > 전통적인 “쿼리”들은 “제품 번호 403이 2018년 2월에 얼마나 배송되었는가?”와 같은 질문들에 대한 답 제공
- > “OLAP, 즉 다차원 분석”은 “제품 403의 지난 2년간 분기/지역별 매출액을 계획과 비교하라.” 같이 훨씬 더 복잡한 정보 제공
- > “데이터마이닝(Data Mining)”은 기존 정보보다 새로운 관계성/패턴 발견에 더욱 중점

- 대용량 데이터베이스들에 숨겨져 있는 패턴과 관계성을 찾아내고, 이런 패턴 및 관계성을 통해 미래의 행위를 예측
- OLAP을 통해서 얻어 수 없는 통찰을 제공하고, 패턴과 규칙들은 의사결정을 도와주고 효과를 예측하는 데 사용

(1) 연관(Association)정보: 하나의 이벤트와 연결된 발생 건들을 의미

. “옥수수칩이 구매될 때마다 콜라의 구매가 65% 발생하지만 판촉 활동이 있는 경우 콜라의 동반 판매는 85%”

(2) 순차(Sequence)정보: 이벤트들의 시간 흐름과 관련

. “집을 구매한 사람이 2주 내에 새로운 냉장고를 구매하는 경우는 65%, 1개월 내에 오븐을 구매하는 경우는 45%”

(3) 분류(Classification)정보: 기존의 분류 체계에 속한 아이템들을 조사하고 일련의 규칙을 추론함으로써

어떤 아이템이 속한 그룹을 설명해줄 수 있는 패턴

. “이탈할 것 같은 고객들의 특성을 발견하고 관리자들이 이런 고객들을 유지할 수 있는 특별한 캠페인을 고안하기 위해 이런 고객들을 예측할 수 있는 모델을 제공”

(4) 군집(Clustering)정보: 아직 한 번도 정의되지 않은 그룹들을 분류하는 것

. “은행카드에 관한 유사 그룹을 찾아내거나, 인구통계 및 개인 투자 유형 기반 고객 그룹들로 분할하는 것과 같이 데이터를 통해 상이한 그룹들을 발견”

(0) 예측(Forecasting): 어떤 다른 값이 가능할지 예측하기 위해 일련의 기존 값을 이용

. “관리자들이 매출액과 같은 연속적인 변수의 미래 값을 추정하는 데 도움이 되는 패턴들을 데이터에서 찾아냄”

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

### 2) 데이터마이닝(Data Mining):

- > 전통적인 “쿼리”들은 “제품 번호 403이 2018년 2월에 얼마나 배송되었는가?”와 같은 질문들에 대한 답 제공
- > “OLAP, 즉 다차원 분석”은 “제품 403의 지난 2년간 분기/지역별 매출액을 계획과 비교하라.” 같이 훨씬 더 복잡한 정보 제공
- > “데이터마이닝(Data Mining)”은 기존 정보보다 새로운 관계성/패턴 발견에 더욱 중점

- 대용

- OLAP

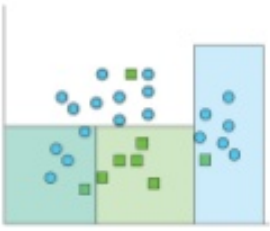
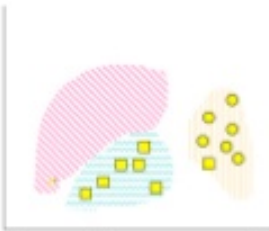
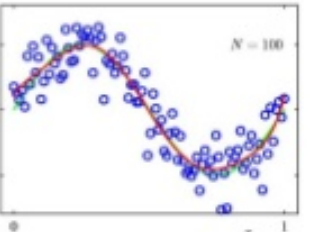

(1)

(2)

(3)

(4)

(0)

Predictive methods	Descriptive methods
<b>Classification</b>  <p>Learns a method for predicting the instance class from pre-labeled (classified) instances</p>	<b>Clustering</b>  <p>Finds “natural” grouping of instances given un-labeled data</p>
<b>Regression</b>  <p>An attempt to predict a continuous attribute</p>	<b>Association Rules</b>  <p>Method for discovering interesting relations between variables in large DBs</p>

를 예측

사용

35%”

는 45%”

과 같이

찾아냄”

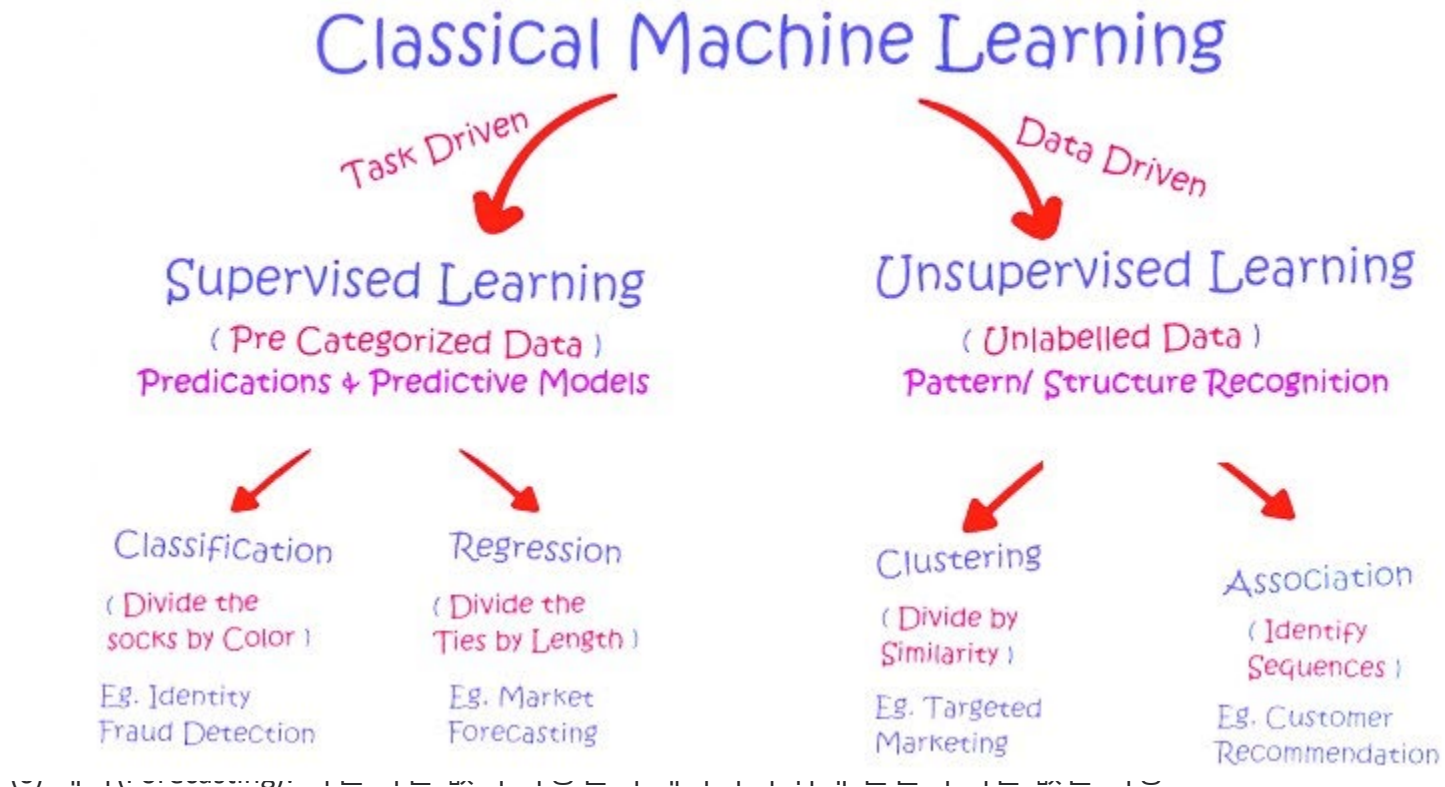


# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

### 2) 데이터마이닝(Data Mining):

- > 전통적인 “쿼리”들은 “제품 번호 403이 2018년 2월에 얼마나 배송되었는가?”와 같은 질문들에 대한 답 제공
- > “OLAP, 즉 다차원 분석”은 “제품 403의 지난 2년간 분기/지역별 매출액을 계획과 비교하라.” 같이 훨씬 더 복잡한 정보 제공
- > “데이터마이닝(Data Mining)”은 기존 정보보다 새로운 관계성/패턴 발견에 더욱 중점



“관리자들이 매출액과 같은 연속적인 변수의 미래 값을 추정하는 데 도움이 되는 패턴들을 데이터에서 찾아냄”

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

### 2) 데이터마이닝(Data Mining):

- 패턴 또는 추세에 대한 수준 높은 분석들을 수행하지만, 또한 필요한 경우 더 상세한 정보를 제공하기 위해 드릴다운 가능
- 기업의 모든 기능 영역과 정부 및 과학 분야에 대한 데이터마이닝 애플리케이션들이 존재
- 보편적 활용 사례 중, 일대일 마케팅 캠페인이나 수익성 높은 고객들을 식별 위한 고객 데이터의 구체적인 패턴 분석
  - . 시저스 엔터테인먼트는 세계에서 가장 큰 카지노업체로 고객들이 슬롯머신을 사용하거나 카지노와 호텔을 이용할 때 수집된 고객들의 데이터를 지속적으로 분석
  - . 마케팅부서는 이런 정보 활용하여 특별한 고객들의 프로파일을 구축했는데, 고객들의 회사에 대한 가치를 기반 수행
    - . “시저스는 데이터마이닝을 통해 중서부 유람선 카지노 중 하나에서 정규 고객들이 가장 좋아하는 게임이 무엇인지”
    - . “객실, 식당, 엔터테인먼트에 대한 선호도”
  - . 수익성이 매우 높은 고객들을 어떻게 늘려 나갈지, 어떻게 이런 고객들로 하여금 더 소비하게 만들지, 높은 수익성 창출 대상인 잠재 고객들을 어떻게 더 끌어들이는지에 관한 경영진의 의사결정 지원
  - . 비즈니스 인텔리전스(BI)는 시저스의 수익을 크게 증대시켰고, 이에 따라 이것은 비즈니스 전략의 핵심적인 부분



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

### 2) 데이터마이닝(Data Mining):

- 패턴 도
- 기업으
- 보편적
- . 시지
- 카지
- . 마커
- . “
- . “
- . 수으
- 높은
- . 비즈



다운 가능

분석

간 수행  
!엇인지”

분

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

### 2) 데이터마이닝(Data Mining):



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

### 3) 텍스트마이닝과 웹마이닝(Text Mining and Web Mining):

- > **텍스트마이닝(Text Mining):** 이메일, 메모, 콜센터 상담 녹취록, 설문 응답, 고소장, 특허 기술, 서비스 보고서 등의 텍스트 데이터를 기반으로 직원들의 비즈니스 의사결정에 도움될 패턴과 추세 추출 알고리즘
  - 텍스트 파일 형태로 된 비구조적 데이터들이 조직의 유용한 정보 중 80% 이상을 차지
  - 대량의 비구조적 데이터들로부터 핵심적인 요소들을 추출하고, 패턴과 관계성을 발견하고, 정보를 요약
    - . 감정 분석(Sentiment Analysis)는 이메일 메시지, 블로그, 소셜 미디어 대화, 설문지 등의 문자들을 분석하여 특정 주제들에 대해 우호적인 의견들과 비우호적인 의견들을 탐지
    - . 고객 서비스를 식별하고 이슈를 바로잡거나 또는 회사에 대한 고객의 감정을 측정하기 위해 고객 서비스 센터에 걸려온 전화 녹취록을 분석하는 데 텍스트마이닝을 활용
    - . 크래프트 푸드는 커뮤니티 인텔리전스 포털(Community Intelligence Portal)과 감정 분석을 사용하여 수많은 소셜 네트워크, 블로그, 그리고 여타의 웹사이트들에서 제품에 대한 소비자 대화 분석
- 브랜드 언급만 추적하기보다는 관련 의견들을 이해하려고 노력하는데, 고객이 바비큐하는 방식과 소스, 향신료에 대해 대화할 때 그들의 감정과 느낌을 파악

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

### 3) 텍스트마이닝과 웹마이닝(Text Mining and Web Mining):

> 웹마이닝(Web Mining): 웹으로부터 유용한 패턴과 정보들을 발견해내고 분석하는 알고리즘

- 고객 행위에 대한 패턴, 추세, 통찰을 드러내줄 수 있는 비구조적인 빅데이터에 대한 또 하나의 풍요로운 원천
- 고객 행위를 이해하거나 어떤 특별한 웹사이트의 성과를 평가하거나 또는 마케팅 캠페인의 성공을 측정

. 마케터들은 Google Trends와 Google Insights for Search 서비스를 통해 구글 검색 쿼리에서 많이 쓰이는 다양한 단어와 구문들을 파악함으로써 사람들이 관심을 가지고 있는 것과 사고 싶어 하는 것을 미리 파악

- (1) 웹콘텐츠마이닝(Web Content Mining)은 웹페이지들의 콘텐츠들로부터 지식을 추출하는 프로세스인데, 웹페이지에는 텍스트, 이미지, 오디오, 비디오 데이터들이 포함
- (2) 웹구조마이닝(Web Structure Mining)은 웹문서에 삽입되어 있는 링크들로부터 유용한 정보들을 추출하는 프로세스  
. 어떤 하나의 문서를 가리키는 링크들은 그 문서의 인기, 어떤 문서들에서 나오는 링크들은 그 문서의 풍부함/다양성
- (3) 웹사용마이닝(Web Usage Mining)은 웹사이트 자원들 요청이 인지될 때 웹서버에 기록된 사용자 상호작용 데이터 분석  
. 기업이 특정 고객들에 대한 가치, 교차 마케팅 전략, 효과적인 캠페인 등에 대한 의사결정을 할 때 도움



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 데이터의 관계성, 패턴, 추세를 파악하는 방향은?

### 3) 텍스트마이닝과 웹마이닝(Text Mining and Web Mining):

#### > 웹마이

- 고객

- 고객

. 마

다

(1) 웹

웹

(2) 웹

. 어

(3) 웹

. 기

## Text Analytics Use Cases



### Manufacturers

- Identify root causes of product issue quicker
- Identify trends in market segments
- Understand competitors products



### Government

- Identify fraud
- Understand public sentiments about unmet needs
- Find emerging concerns that can shape policy



### Financial Institutions

- Use contact center transcriptions
- Understand customers
- Identify money laundering or other fraudulent situation



### Retail

- Identify profitable customers and understand the reasons for their loyalty
- Manage the brand on social media



### Legal

- Identify topics and keywords in discovery documents
- Find patterns in defendant's communications



### Healthcare

- Find similar patterns in doctor's reports
- Use social media to detect outbreaks earlier
- Identify patterns in patient claims data



### Telecommunications

- Prevent customer churn
- Suggest up-sell/cross-sell opportunities by understanding customer comments



### Life Sciences

- Identify adverse events in medicines or vaccines
- Recommend appropriate research materials



### Insurance

- Identify fraudulent claims
- Track competitive intelligence
- Manage the brand on social media

zencos

프로세스  
함/다양성  
데이터 분석

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

SAS® Report Viewer - View Reports

Search

sasdemo

HT\_Suffolk

Close

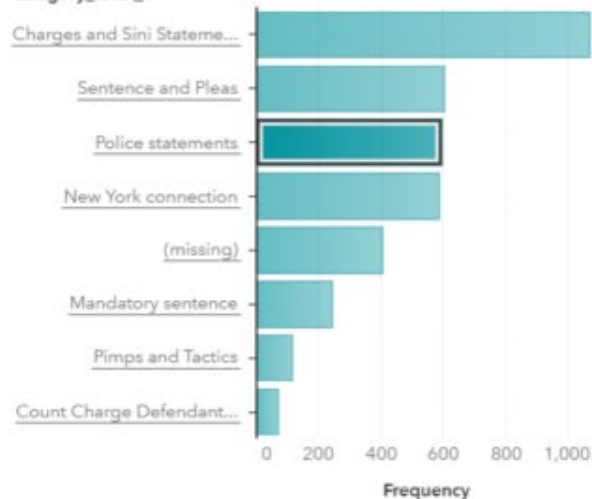
Page 1

Page 2

Frequency of category\_level\_1

All Concepts and Matches > nlpPlace

category\_level\_1



_match_text_	_sentiment_	sentences	Name
Hudson County	Negative	4 men sentenced for sex trafficking of N.J. teen runaway Posted May 27, 2016 TRENTON -- Four men who ran ...	NJTeenRunaway.txt
N.J.	Negative	4 men sentenced for sex trafficking of N.J. teen runaway Posted May 27, 2016 TRENTON -- Four men who ran ...	NJTeenRunaway.txt
Deer Park	Negative	A Bloods gang member from Deer Park ran a sex-trafficking ring in which he used the lure of drugs and fear o...	Bloods gang member from Deer Park.txt
Suffolk County	Negative	A Bloods gang member from Deer Park ran a sex-trafficking ring in which he used the lure of drugs and fear o...	Bloods gang member from Deer Park.txt
Suffolk County	Negative	Abiodun Adeleke, 32, was arraigned Saturday after police arrested him at a Plainview Holiday Inn following a...	Bloods gang member from Deer Park.txt
Suffolk County	Negative	At the request of prosecutors, Suffolk County Judge James Malone issued orders of protection for 19 women...	Bloods gang member from Deer Park.txt
Corona, Queens	Negative	Dozens of women were smuggled to New York City and forced to sell themselves for a group of men who ar...	Emilio Rojas-Romero.txt
Hudson Valley	Negative	Dozens of women were smuggled to New York City and forced to sell themselves for a group of men who ar...	Emilio Rojas-Romero.txt
Long Island	Negative	Dozens of women were smuggled to New York City and forced to sell themselves for a group of men who ar...	Emilio Rojas-Romero.txt

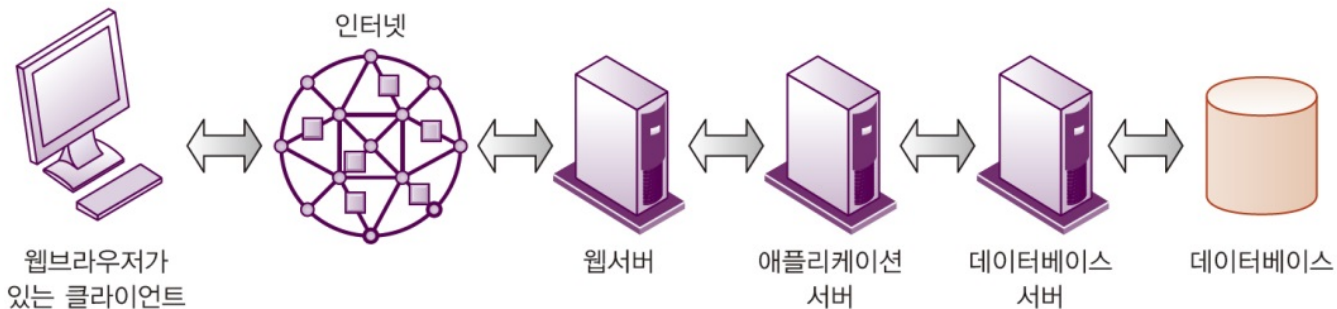
# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 고객, 웹사이트, 디지털플랫폼과의 실시간 연결성은?

- 많은 기업들은 현재 고객 및 비즈니스 파트너들이 내부 데이터베이스 일부를 사용할 수 있도록 웹을 제공

**그림 6.15** 웹에 대한 내부 데이터베이스의 연결

사용자들은 데스크톱 PC와 웹브라우저 소프트웨어를 사용함으로써 웹을 통해 조직의 내부 데이터베이스에 접근할 수 있다.



- 웹서버는 데이터에 대한 이런 요청들을 소프트웨어로 넘길 것
- 애플리케이션 서버는 브라우저 기반의 컴퓨터들과 후방 비즈니스 애플리케이션 또는 데이터베이스들 사이에서 일어나는 거래처리 및 데이터 접근을 포함한 모든 애플리케이션의 동작 관여
- 데이터베이스 서버는 조직의 내부 데이터베이스에 있는 정보를 고객 페이지로 전달해주는 웹서버로 전달



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 성공적 데이터관리를 위한 정책과 품질은?

- 데이터베이스를 설치하는 것은 시작에 불과
- 보유 데이터를 정확하고 신뢰하며 필요한 사람들이 충분히 사용하도록 데이터 관리 정책과 절차 필요
- 그러나, 아무나 원하던 것을 뭐든지 할 수 있게 두는 것은 자유라는 리스크에 대한 책임이 뒤따름
- 유지되는 방식에 관한 규칙과 데이터를 보거나 변경할 수 있는 사람들에 관한 규칙이 필요
- **정보정책(Information Policy):** 정보를 공유/분배/획득/표준화/분류/목록화하기 위한 조직규칙들을 규정
  - 특정 절차 및 책임 소재를 규정하는데, 구체적으로 어떤 사용자나 조직단위들이 정보를 공유할 수 있는지, 정보가 어느 곳으로 유통될 수 있는지, 누구에게 정보에 대한 갱신과 유지관리 책임이 있는지 등을 규정
  - . 급여 및 인적자원 부서는 특정 직원들만이 직원 봉급, 주민등록 번호 같은 민감한 데이터들을 변경/열람할 권한을 주고, 직원들의 데이터를 정확하고 신뢰성있게 유지할 책임
  - 규모가 작은 회사에 다닌다면, 정보정책은 기업주나 관리자가 수립하고 구현할 것
  - 대기업에서는 기업 자산인 정보에 대한 관리 및 계획수립을 종종 공식적인 데이터 관리 기능으로 규정
- **데이터 관리(Data Administration):** 데이터가 조직의 자원으로 관리되는 데 필요한 정책과 절차들을 완수
  - 정보정책 개발, 데이터 계획수립, 논리적 DB 설계 및 데이터 사전 개발 감독, 그리고 정보시스템 전문가들과 최종사용자들이 데이터를 어떻게 사용하는지에 대한 감독 등과 같은 활동이 포함

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 성공적 데이터관리를 위한 정책과 품질은?

- 데이터바
- 보유 데
- 그러나,
- 유지되는
- 정보정책
  - 특정 절차
  - 정보가 C
  - 급여!
  - 직원들
  - 규모가 작
  - 대기업에
- 데이터
  - 정보정책
  - 최종사용

### < 국가데이터맵 서비스 기대효과 >

#### 범정부 데이터 플랫폼 개념도



필요

이

을 규정

한을 주고,

을 완수

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 성공적 데이터관리를 위한 정책과 품질은?

- 데이터
- 보유 데
- 그러나,
- 유지되
- 정보정
- 특정 절
- 정보가
- . 급여
- 직원
- 규모가
- 대기업
- 데이터
- 정보정
- 최종사



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 성공적 데이터관리를 위한 정책과 품질은?

- 데이터베이스
- 보유 데이터
- 그러나, 데이터가 많아질수록
- 유지되는 데이터의 양이 한정적
- 정보정책
  - 특정 절차에 따라 데이터를 수집, 가공, 분석, 배포하는 등
  - 급여! 데이터가 많아질수록
  - 직원들이 데이터를 관리할 수 있는
  - 규모가 작을수록
  - 대기업에 비해
- 데이터 관리
  - 정보정책을 수립하고
  - 최종사용자에게

### < 국가데이터맵 서비스 기대효과 >

#### 범정부 데이터 플랫폼 개념도



필요

이

을 규정

한을 주고,

을 완수

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 성공적 데이터관리를 위한 정책과 품질은?

- **데이터 거버넌스(Data Governance):** IBM에 의해 처음 등장
  - 기업에서 사용되는 데이터의 가용성(Availability), 유용성(Usability), 통합성(Integrity), 보안성(Security)을 관리하기 위한 정책과 프로세스들을 다루며 프라이버시, 보안성, 데이터 품질, 관리 규정 준수를 특히 강조
  - 정보정책 개발, 데이터 계획수립, 논리적 DB 설계 및 데이터 사전 개발 감독, 그리고 정보시스템 전문가들과 최종사용자들이 데이터를 어떻게 사용하는지에 대한 감독 등과 같은 활동이 포함
- **데이터베이스 관리(Database Administration):** 대형조직은 데이터베이스 관리와 설계 그룹 구성
  - 데이터베이스 관리 그룹은, 정보시스템 부서 내에서 데이터베이스의 구조와 콘텐츠를 정의하고 구성하며, 데이터베이스를 유지관리하는 역할을 담당
  - 데이터베이스 설계 그룹은, 사용자들과의 긴밀한 협조를 통해서 물리적 데이터베이스, 관련 요소 간의 논리적 관계 그리고 접근 규칙과 보안 절차를 수립

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 성공적 데이터관리를 위한 정책과 품질은?

- **데이터 거버넌스(Data Governance):** IBM에 의해 처음 등장
  - 기업에서 사용되는 데이터의 가용성(Availability), 유용성(Usability), 통합성(Integrity), 보안성(Security)을 관리하기 위한 정책과 프로세스들을 다루며 프라이버시, 보안성, 데이터 품질, 관리 규정 준수를 특히 강조
  - 정보정책 개발, 데이터 계획수립, 논리적 DB 설계 및 데이터 사전 개발 감독, 그리고 정보시스템 전문가들과 최종사용자들이 데이터를 어떻게 사용하는지에 대한 감독 등과 같은 활동이 포함
- **데이터베이스 관리(Database Administration):** 대형조직은 데이터베이스 관리와 설계 그룹 구성
  - 데이터베이스 관리 그룹은, 정보시스템 부서 내에서 데이터베이스의 구조와 콘텐츠를 정의하고 구성하며, 데이터베이스를 유지관리하는 역할을 담당
  - 데이터베이스 설계 그룹은, 사용자들과의 긴밀한 협조를 통해서 물리적 데이터베이스, 관련 요소 간의 논리적 관계 그리고 접근 규칙과 보안 절차를 수립



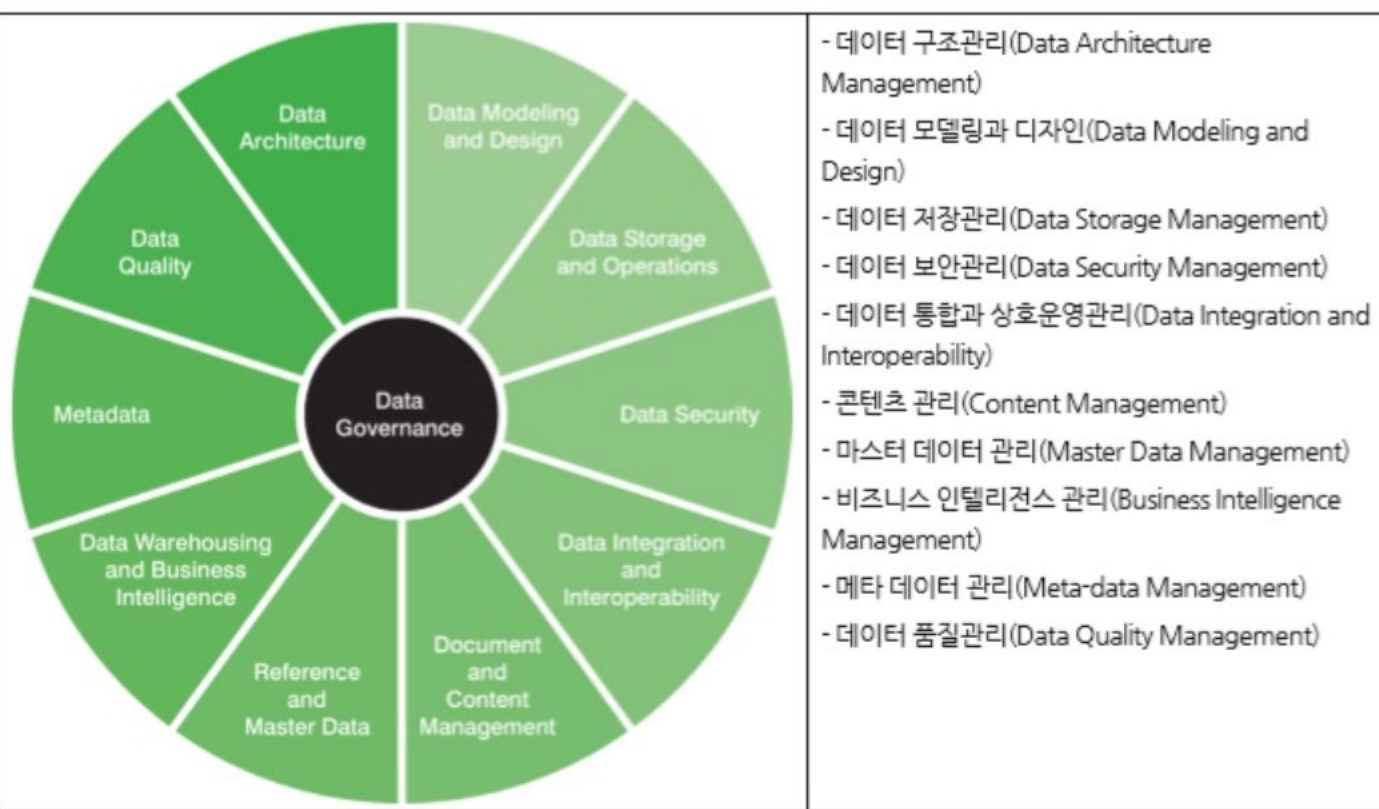
## 성공적

### 데이터

- 기업에서 정책과
- 정보정책 최종사용

### 데이터

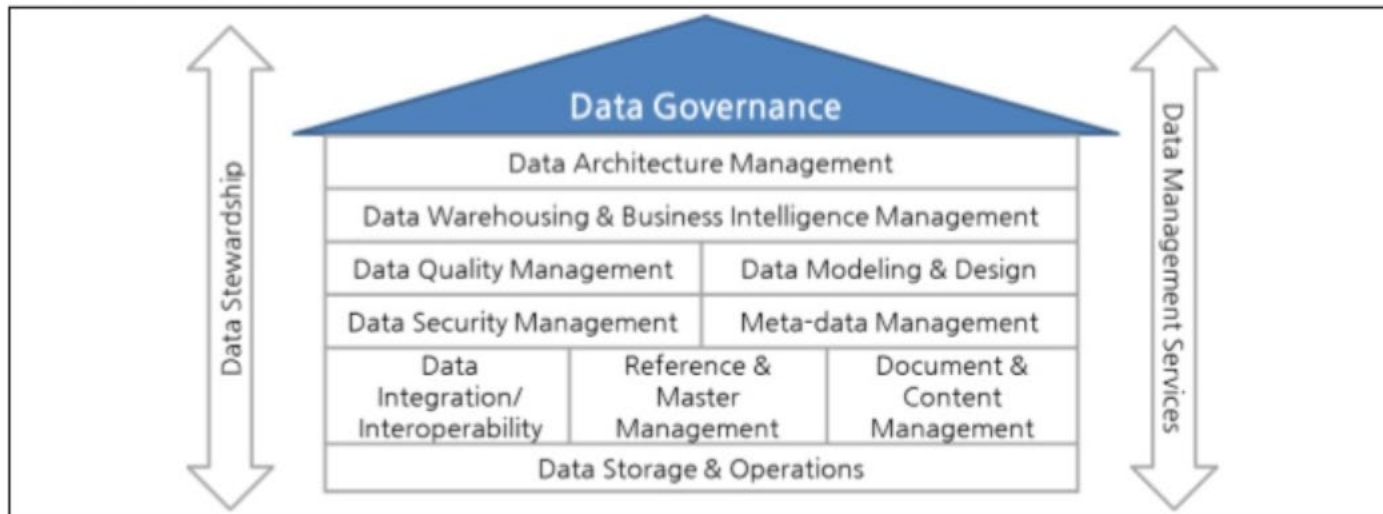
- 데이터
- 데이터
- 데이터 그리고



기 위한

관계

[그림 1] 데이터 관리 기능



[그림 2] 데이터 관리 프레임워크



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 성공적 데이터관리를 위한 정책과 품질은?

빅데이터 거버넌스 프레임워크

헬스케어	감정분석	환자 모니터링	청구 분석	유전자 검사	전자 의무기록
유틸리티		스마트미터			
소매업	Facebook 충성도	RFID 태그		얼굴인식	
통신	고객이탈 분석	위치기반 서비스	고객이탈 분석		
보험	청구조사	자동차 텔레매틱스	사기청구 분석	보험가입	
고객서비스					콜 품질보증
IT		IT 로그분석			
	웹과 소셜미디어	M2M 데이터	빅 트랜잭션 데이터	생체 정보	사람이 생성한 빅데이터

빅데이터 유형



# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 성공적 데이터관리를 위한 정책과 품질은?

- 잘 설계된 데이터베이스와 정보정책은 기업이 필요한 정보를 확실히 확보하는 데 큰 도움
- 조직 데이터베이스의 데이터들이 정확하고 신뢰할 수 있도록 유지되기 위해서는 추가적인 단계들이 필요
- 부정확하고, 시점이 다르고, 다른 소스의 정보와 불일치 데이터들은 잘못된 의사결정, 제품 리콜, 심지어는 재무적 손실까지 초래
  - 가트너(Gartner Inc.)는 포춘 1,000대 기업들의 대용량 데이터베이스의 중요 데이터 중 25% 이상이 부정확하거나 불완전
    - . 고객의 전화번호나 계좌번호가 잘못되어 있다면 어떤 일이 일어날까?
    - . 데이터베이스에 이미 판매한 제품의 가격이 잘못된 가격으로 기록되어 있거나, 판매시스템과 재고시스템이 동일 제품에 대해 서로 다른 가격을 보여주고 있다면, 이로 인해 어떤 일이 발생할까?
    - . 불량제품 코드와 제품 설명, 잘못된 재고기술, 잘못된 재무 데이터, 부정확한 공급자 정보, 부정확한 직원 데이터 등 포함
  - 데이터 품질 문제는 데이터웨어하우스에 데이터를 공급하는 여러 시스템에서 생성되는 데이터들의 중복과 불일치로 발생
    - . 판매주문시스템은 어떤 속성은 아이템 번호라 표현하는 반면, 재고관리시스템은 똑같은 속성에 제품 번호 이름을 사용
    - . 어떤 시스템은 의류 사이즈를 'medium'으로 표현하고 있는 반면, 다른 시스템은 'M'이라는 코드값을 사용
    - . 항상 같은 날 같은 다이렉트 광고 메일을 여러 개 받는다면 여러분의 이름이 데이터베이스에 여러 번 입력된 결과
    - . 처음에는 종이 양식에 기록되었지만 시스템으로 입력될 때 제대로 스캔되지 못했을 수도
    - . 이런 불일치 때문에 데이터베이스는 여러분을 서로 다른 사람으로 취급하게 될 수도

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 성공적 데이터관리를 위한 정책과 품질은?

- 대부분의 데이터 품질 문제는 데이터 입력 오류에서 발생하며, 이런 오류들의 발생률은 증가
- 기업들이 비즈니스를 웹으로 옮겨 가면서 고객들과 공급자들이 내부 시스템을 직접 갱신하는 웹사이트에 데이터를 입력할 수 있도록 하기 때문에 결국 사람의 오류가 잔존할 가능성 지속됨
- 데이터베이스가 적절히 설계되고 전사적 데이터 표준들이 수립되면 중복/불일치 데이터 요소들이 최소화
- **데이터 품질 감사(data quality audit):**
  - 정보시스템의 데이터에 대한 정확성과 더불어 완전성 수준에 대한 구조화된 조사
  - 데이터 파일 전체에 대한 조사, 데이터 파일들의 샘플 조사, 또는 데이터 품질에 대한 최종사용자들의 인식 조사 등 수행
  - 조직들은 잘못된 데이터를 식별/수정할 필요가 있고, 새로운 데이터베이스가 운영되면 더 나은 수정 절차들을 구축할 필요
- **데이터 정제(Data Cleansing/Data Scrubbing):**
  - 데이터베이스/파일데이터 중 부정확, 불완전, 부적절한 포맷, 중복된 데이터들 수정하는 활동
  - 데이터를 수정할 뿐만 아니라 개별 정보시스템에서 생성된 상이한 데이터 간의 일치성을 강화
  - 정제 소프트웨어는 데이터 파일들 조사/오류수정하며 일관된 전사적 포맷으로 데이터들을 통합하는 작업을 자동 수행
  - 불완전하고 부정확한 데이터베이스들은 형사사법제도와 공공안전에도 문제를 발생

➔ 소수 기업들에서는 개별 부서들이 자체적 데이터 품질 유지관리, 그러나 최상의 데이터 관리를 위해서는 중앙집중식의 데이터 거버넌스, 조직 데이터 표준화, 데이터 품질 유지관리, 데이터 자산에 대한 접근성 통제가 필요

# 비즈니스 인텔리전스 정보의 추출기술과 품질은?

## ➤ 성공적 데이터관리를 위한 정책과 품질은?

### 데이터 운영표준

#### 운영표준 정의 대상

고객

파트너

상품

서비스

CoA

계좌/계약

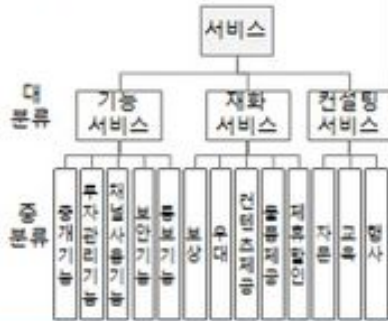
채널

조직

사원

#### 1 분류체계

- 데이터 분류, 실적집계 및 분석 기준



#### 2 속성체계

- 관리하고자 하는 단위 정보항목

Biz파트너기본	Biz파트너번호, 실명번호
주요/연락처	우편번호, 기본주소, 상세주소 유선전화국가, 지역, 국, 번 팩스번호국가, 지역, 국, 번 E-Mail URL주소
고객기본	고객번호, 내외국인유형
고객명세	직장, 직업, 성별, 관계사여부
고객등급	고객등급유형, 고객등급, VIP여부

#### 3 식별체계

- 마스터의 코드체계

상품코드(10자리)	
유의미	무의미
③④⑤⑥⑦⑧⑨	③④⑤⑥⑦⑧⑨
상품 중분류	일련번호
<ul style="list-style-type: none"> <li>상품중분류 영문1자리 + 숫자2자리</li> <li>일련번호 숫자7자리(0000001~9999999)</li> </ul>	

#### 4 표기표준

- 속성의 입력 기준

실명번호	<ul style="list-style-type: none"> <li>특수문자나 '-' 없이 입력 함</li> <li>좌측 첫 번째 자리부터 입력처리 (숫자 사이 공란을 두지 말고 연속적으로 입력)</li> <li>주민등록번호: 13자리 숫자입력</li> </ul>
유선전화번호	<ul style="list-style-type: none"> <li>숫자만 입력함. 특수문자는 입력불가함</li> <li>국가번호, 지역번호, 국번호, 전화번호 각각을 속성으로 구분하여 숫자만 입력 함</li> </ul>

#### 5 생애주기 프로세스

- 데이터 생성/변경/폐기를 위한 업무절차와 기준
- 등록시점
- R&R
- 처리기준



#### 6 품질체계

- 기준정보의 데이터 품질 및 운영수준 평가를 위한 기준

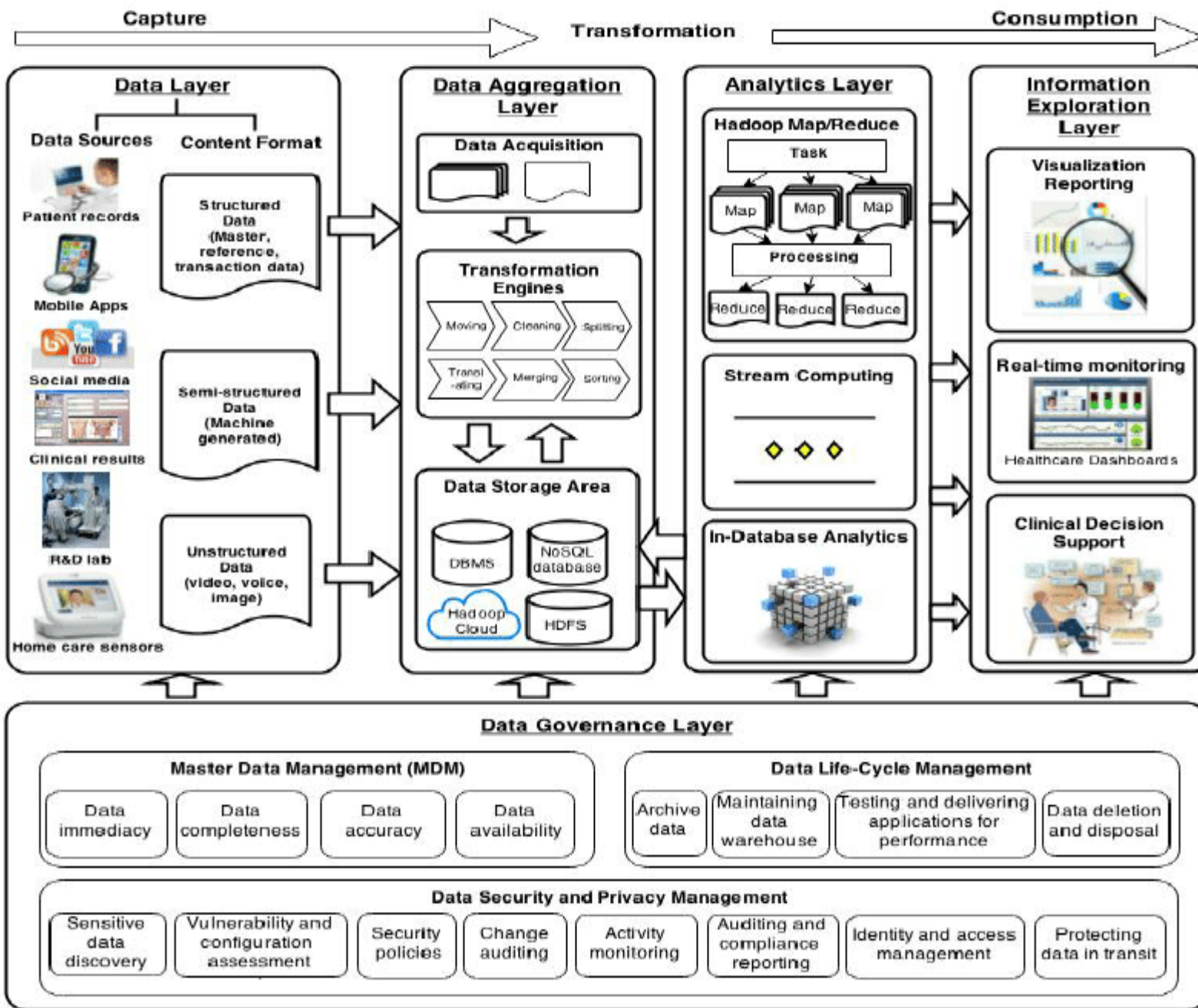
- ✓ 품질지표 : 기준정보 속성값의 품질수준 측정
  - 고객 Biz파트너 유일성 → 고객중복오류 검증
  - 고객 개인고객속성 완전성 → 값 누락 오류 검증
  - 고객 국가코드 유효성 → 값의 코드, 형식, 산식의 준수 검증
- ✓ 운영지표 : 기준정보~프로세스간, 기준정보~기준정보간 정합성 측정
  - 채널 생성리드타임 준수율
  - 서비스 단위서비스 활용율



## 성능

- 대
- 기
- 데
- 데
- 데
- 정
- 데
- 조
- 데
- 데
- 데
- 정
- 불

## 소통



트에  
소화  
행  
필요

생

는  
성



**THANK YOU**

