[2주차] 빅데이터 주요 처리과정 – 빅데이터 수집, 저장, 처리, 분석 기술 이해

김정준

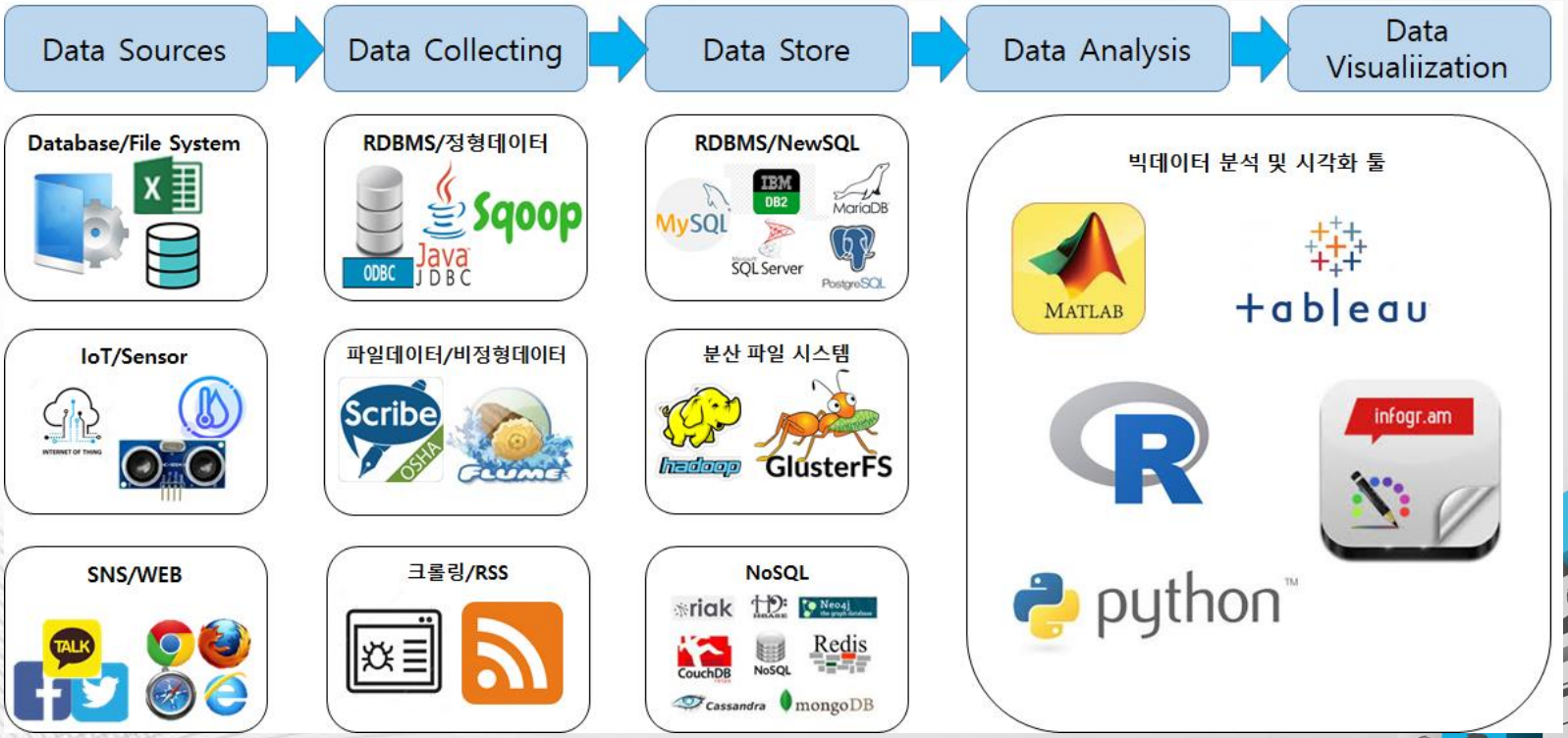
목차

I 빅데이터 처리 과정

빅데이터 처리 과정

빅데이터 처리 과정

- ✓ 빅데이터는 기존의 데이터와 속성이 다름
- ✓ 데이터 수집,저장,처리,분석,시각화 하기 위한 새로운 기술(방법) 필요

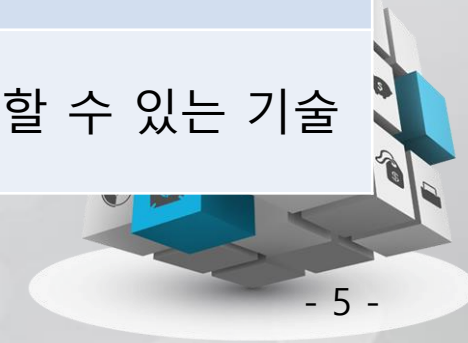


빅데이터 처리 과정

빅데이터 소스 생성과 수집 기술

- ✓ 내부 데이터 수집 : 자체적으로 보유한 내부 파일 시스템 혹은 DBMS 등에 접근 (정형 데이터)
- ✓ 외부 데이터 수집 : 인터넷으로 연결된 외부에서 수집 (비정형 데이터)
  - » 데이터 수집은 주로 툴, 프로그래밍 언어를 이용하여 자동으로 진행

방법	설명
Flume	시스템 또는 웹 서버의 로그, IOT 환경의 센싱 데이터를 수집할 수 있는 기술
SQOOP	RDBMS와 빅데이타 플랫폼(HDFS) 사이에서 데이터를 서로 공유 및 수집할 수 있는 기술
Crawling	인터넷에 분산 되어 있는 데이터를 수집하기 위해 웹(인터넷) 관련 라이브러리가 제공되는 개발 언어를 사용할 수 있는 기술
Open API	공공기관 등에서 제공하는 공공 데이터를 수집할 수 있는 기술



빅데이터 처리 과정

빅데이터 저장 기술

- ✓ 데이터에서 의미 있는 정보를 추출 하기 위해 효율적인 저장 관리 기술 필요
- ✓ '대용량, 비정형, 실시간성' 속성을 수용할 수 있는 저장 방식 필요
  - » 대량의 데이터를 파일 형태로 저장할 수 있는 기술
  - » 비정형 데이터를 정형화된 데이터 형태로 저장하는 기술

방법	설명
HDFS	대용량 데이터를 저장하기 위해서 여러 대의 PC를 클러스터링하여 분산 저장하는 파일 시스템
NoSQL	기존의 RDBMS의 특징대신 가용성과 확장성을 강화한 비정형 데이터베이스

빅데이터 처리 과정

빅데이터 처리 기술

- ✓ 방대한 양의 데이터와 데이터 생성 속도 및 종류의 다양성을 통합적으로 고려할 수 있는 기술 필요
- ✓ 대표적인 맵리듀스 기술은 일반 범용 서버로 구성된 군집화 시스템을 기반으로 한 분산 컴퓨팅 기술
  - » 기술이 확산되면서 새로운 하드웨어 시스템에 최적화된 데이터 처리기술, 반복 및 연속 처리 지원, 유연한 데이터 흐름을 표현하는 프로그래밍 모델을 개선하는 연구가 진행되고있음

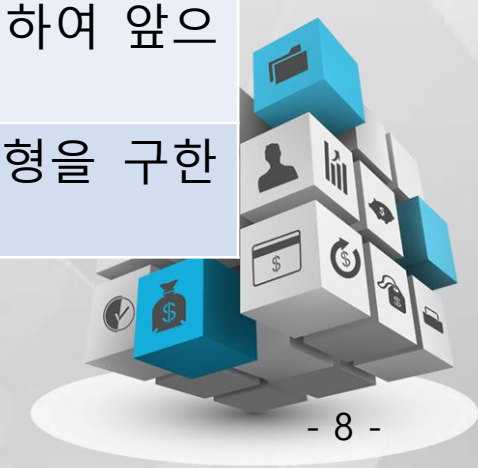
방법	설명
Pig	맵리듀스를 사용하기 위한 높은 수준의 스크립트 언어와 이를 위한 인프라로 구성되어 있음
Hive	SQL과 비슷한 HiveQL을 이용하여 데이터를 처리하는 데이터 웨어하우스 소프트웨어
Spark	디스크 기반의 처리기술이 아닌 메모리 기반의 처리기술로 기존의 처리 속도보다 10~100배 빠른 처리기술

빅데이터 처리 과정

빅데이터 분석 기술

- ✓ 분석에 사용하는 기술은 대부분 통계학과 전산학, 특히 기계 학습과 데이터 마이닝 분야에서 이미 사용한 것 들
  - » 알고리즘을 대규모 데이터 처리에 맞게 개선하여 빅데이터 처리에 적용

방법	설명
연관성 분석	데이터 간의 관계를 정의하며, 어떤 결과에 대하여 객체 간에 동시에 발생한 일 등을 분석
분류 분석	훈련 데이터를 이용하여 학습시키고 새로운 데이터가 어느 클래스에 속하게 될지 분석
군집화	특성이 비슷한 데이터를 군으로 묶어주는 기술
시계열 분석	시간의 흐름에 따라 기록된 데이터를 분석하여 앞으로의 변화를 예측
회귀 분석	연속형 변수들에 대해 두 변수 사이의 모형을 구한 후에 적합도를 측정하는 분석 방법



빅데이터 처리 과정

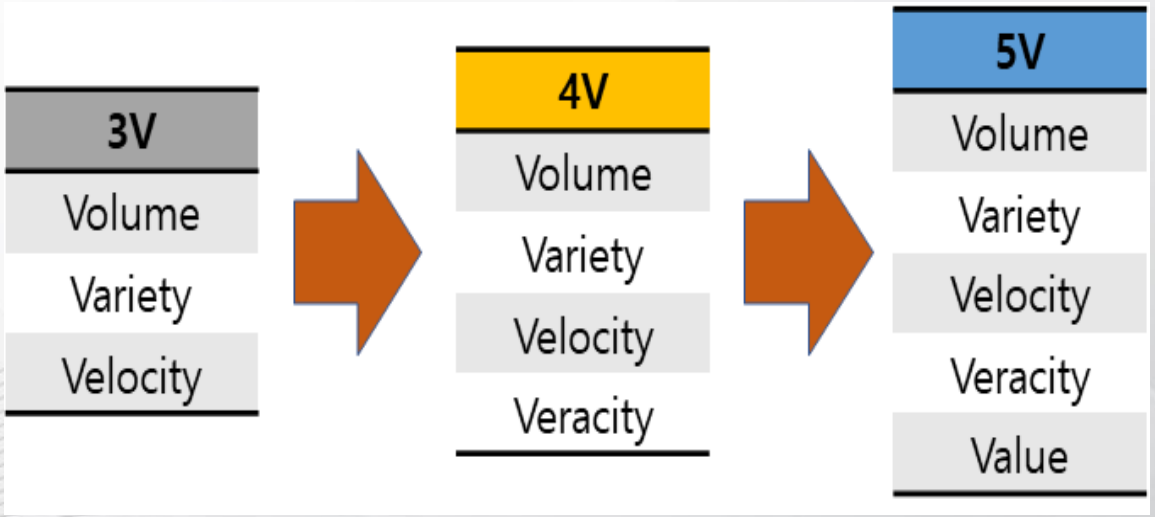
빅데이터 표현 기술

✓ 데이터 분석 결과를 효과적으로 전달하기 위해 복잡한 정보를 한눈에 쉽게 이해 할 수 있도록 간단한 도표나 3D 이미지 등으로 표현하는 기술

빅데이터 처리 과정

빅데이터 특징

- ✓ 데이터양의 증가로 기존의 데이터 저장, 관리, 분석 기법으로 데이터 처리에 한계
- ✓ 정보기술의 패러다임이 변화되면서 최근 데이터에 대한 전반적인 과정을 빅데이터라 표현함
- ✓ 3V → 5V
  - » 규모(Volume) : 데이터의 크기
  - » 속도(Velocity) : 데이터를 바르게 처리하고 분석할 수 있는 속성
  - » 다양성(Variety) : 다양한 종류의 데이터를 수용하는 속성
  - » 정확성(Veracity) : 데이터에 부여할 수 있는 신뢰 수준
  - » 가치(Value) : 빅데이터를 저장하기 위한 IT 인프라 구조 시스템 구현 비용

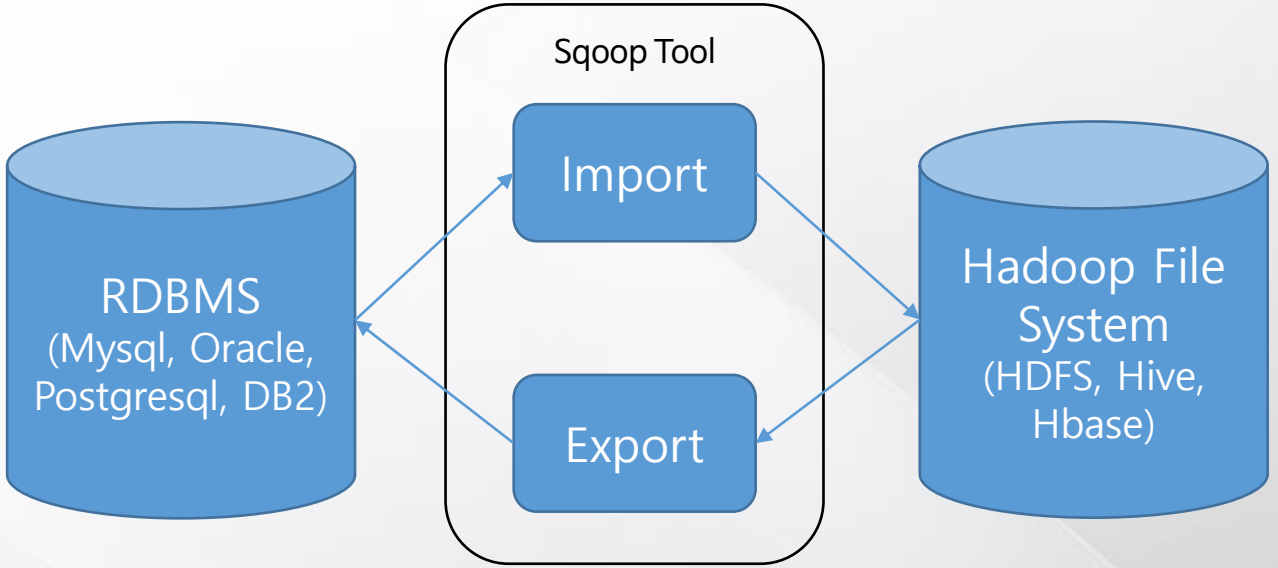


II 수집, 저장, 처리, 분석 기술

빅데이터 수집 기술

Sqoop

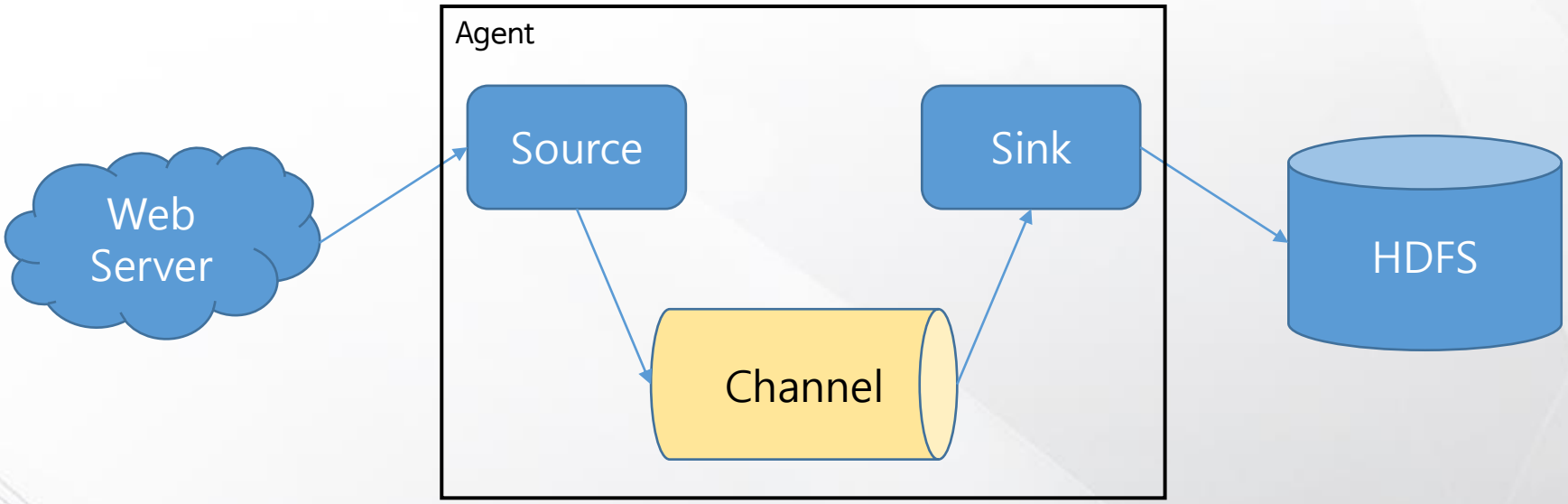
- ✓ RDBMS, Data Warehouse 등 정형 데이터를 빅데이터 저장 도구인 Hadoop에 저장 하는 데이터 수집 도구
- ✓ RDBMS의 데이터를 Hadoop에 저장하는 Import 과정, Hadoop 데이터를 RDBMS에 저장하는 Export 과정



빅데이터 수집 기술

Flume

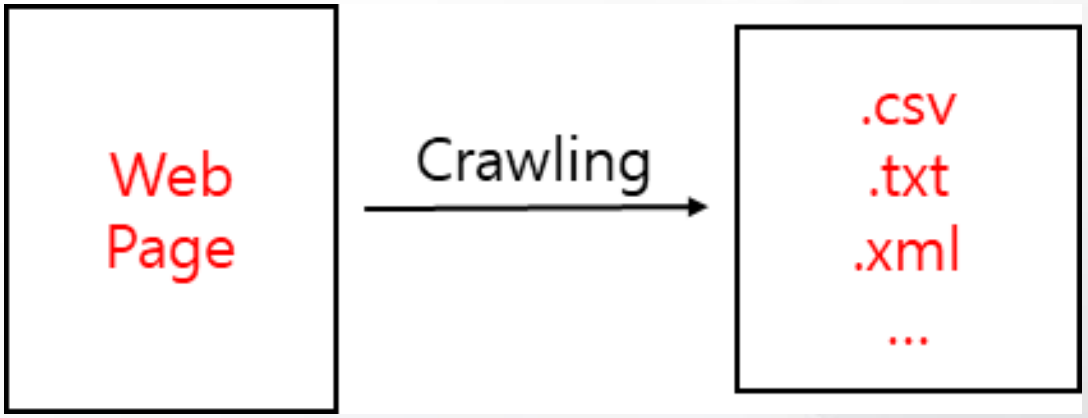
✓ 서버에서 발생하는 로그를 수집하여 빅데이터 저장도구인 Hadoop에 저장하는 데이터 수집 도구



빅데이터 수집 기술

Crawling

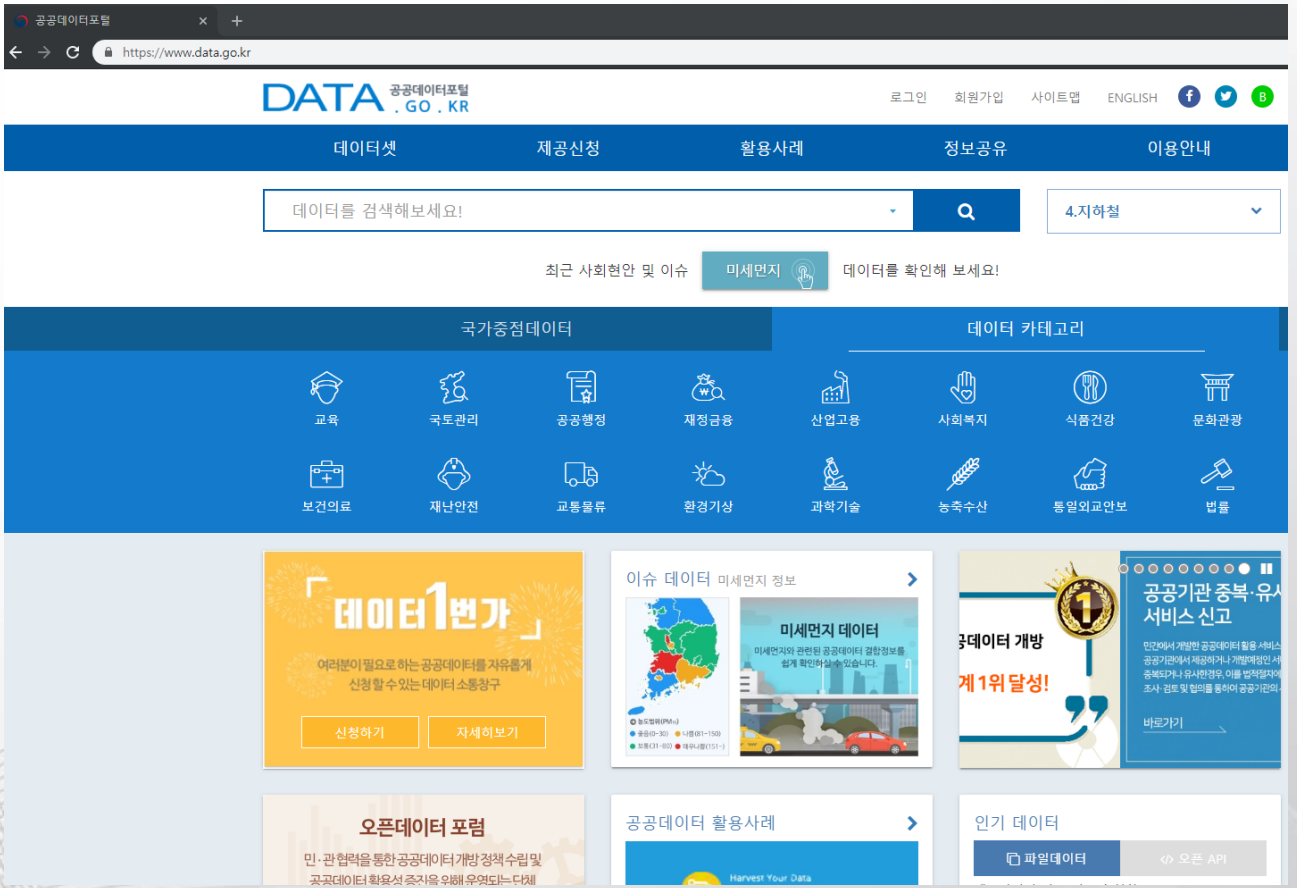
- ✓ 웹페이지의 내용을 수집하는 프로그램
- ✓ 대량의 웹 문서를 인간이 직접 구별하여 수집하는 일은 불가능에 가깝기 때문에 크롤러를 사용하여 수집을 자동화 할 수 있음



빅데이터 수집 기술

공공데이터

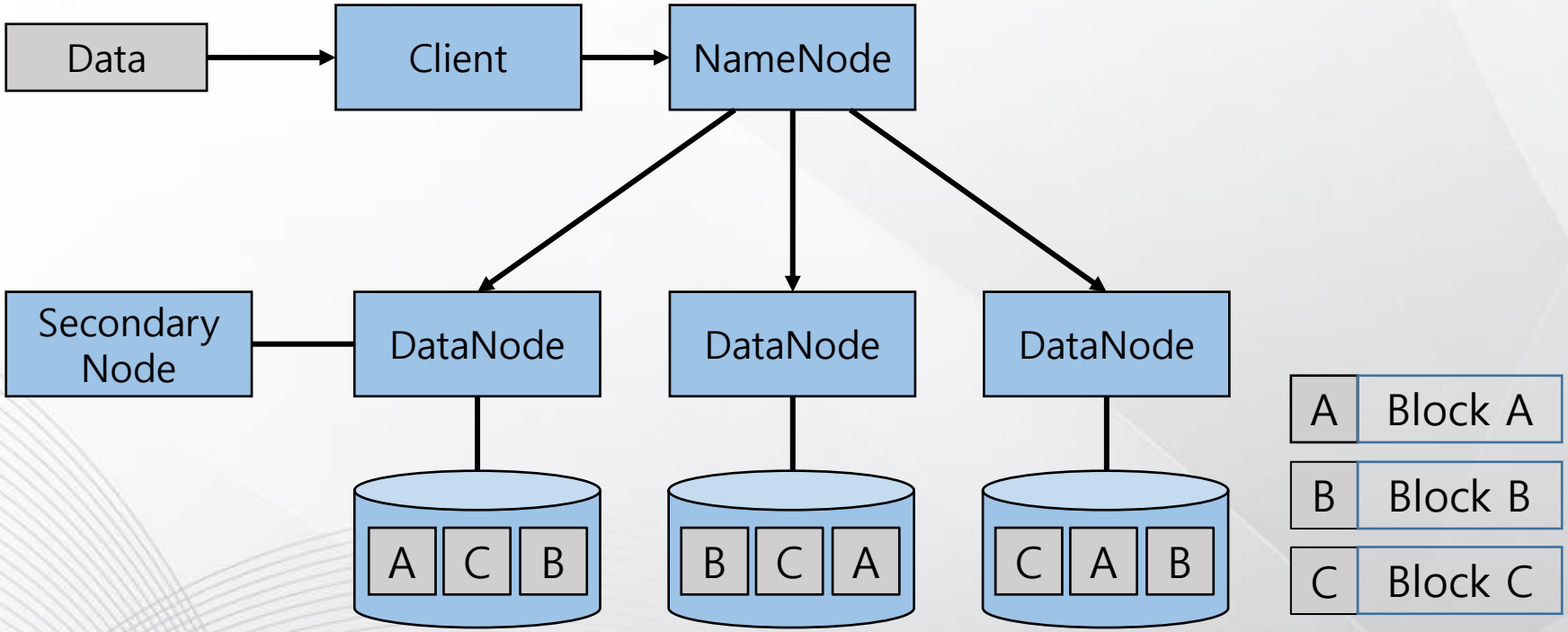
- ✓ 공공기관이 생성 또는 취득하여 관리하고 있는 데이터를 의미
- ✓ 파일데이터, 오픈API, 시각화 등 다양한 방식으로 제공



빅데이터 저장 기술

Hadoop

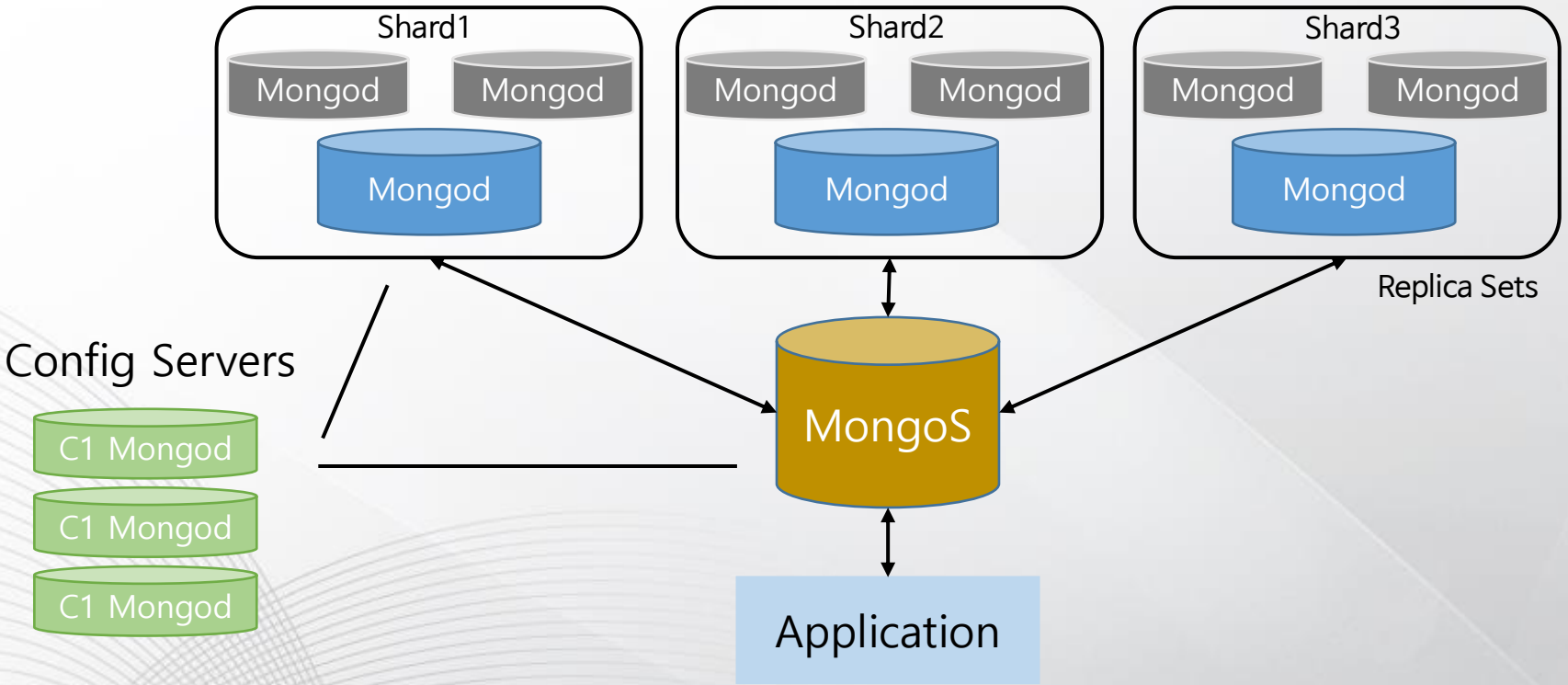
- ✓ 분산 파일 시스템(HDFS)과 Map-Reduce를 구현한 빅데이터 저장 및 처리 기술의 대표적인 프레임워크
- ✓ 분산 파일 시스템의 파일을 볼륙단위로 나눠서 복제 저장하여 노드 장애 및 데이터 손실에 빠른 복구 기술을 보유



빅데이터 저장 기술

NoSQL

- ✓ 빅데이터의 대표적인 저장 도구로 하둡과 함께 많이 사용
- ✓ Replica Set과 Sharding으로 구성된 데이터베이스로써 안정성과 가용성이 높음
- ✓ 비정형 데이터를 저장하는데 적합하며, 실시간 처리에 사용되는 대표적인 저장 도구



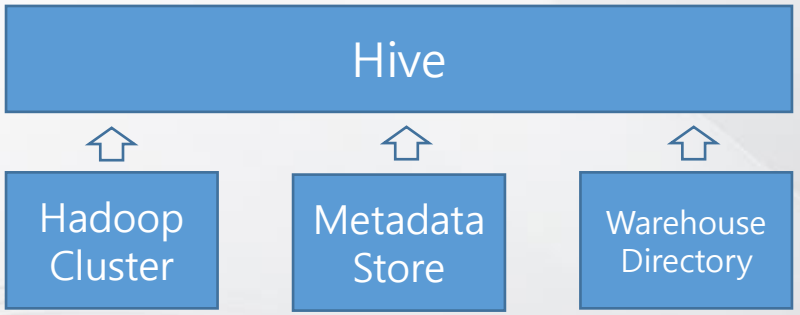
빅데이터 처리 기술

Hive

- ✓ Hive는 페이스북 주도로 개발되었다.
- ✓ Hadoop기반으로 동작하는 데이터 웨어 하우스 시스템이다.
- ✓ Hadoop에 저장된 데이터를 SQL과 유사한 HiveQL를 사용하여 처리할 수 있는 기능을 제공한다.
- ✓ 따라서 JAVA를 활용한 Map/Reduce 개발자가 아닌 RDBMS에 익숙한 개발자가 Map/Reduce를 구현하고자 할 때, 편리한 인터페이스를 제공한다.

Hive 아키텍처

- ✓ Metadata Store : HDFS 파일과 Hive 테이블을 연결한 정보가 저장되는 저장소



빅데이터 처리 기술

Pig

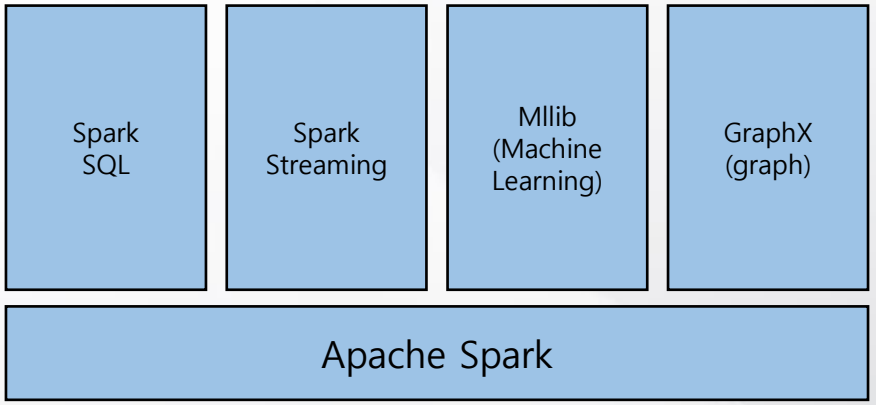
- ✓ 고수준언어로 데이터분석을 프로그래밍 할 수 있는 대용량 데이터셋 분석 플랫폼
- ✓ 맵리듀스에서 처리할 수 없는 부분들을 지원 하는데, 대표적으로 조인과 같은 연산
- ✓ 장점
  - » 데이터 구조를 자세히 검토할 수 있는 여러 명령어 제공
  - » 입력 데이터의 대표 부분 집합에 대해 표본실행 가능
  - » 확장가능 : 수행 경로상 많은 부분 UDF라는 사용자 정의 함수를 이분해 변경 가능
- ✓ 단점
  - » 대용량 데이터셋에 적합하므로 적은양의 데이터를 처리하기엔 비효율적

피그	
피그 라틴 데이터의 흐름을 표현 하기 위해 사용하는 언어 쉬운프로그래밍,최적화,효율형	**실행 환경** 피그라틴 프로그램을 수행하는 실행 환경 1. JVM에서의 로컬 실행환경 2. 하둡 클러스터 상의 분산 실행 환경

빅데이터 처리 기술

Spark

- ✓ 디스크 I/O 처리와 다르게 메모리 기반의 처리 기술로 기존의 Map-Reduce보다 10~100 빠르게 설계
  - » 데이터를 메모리에 캐시로 저장하는 모델을 사용하기 때문에 성능이 향상
- ✓ 머신러닝, 그래프 알고리즘 등 다양한 분야에서 Spark기술의 효율이 높음



빅데이터 분석 기술

분석 및 시각화 기술

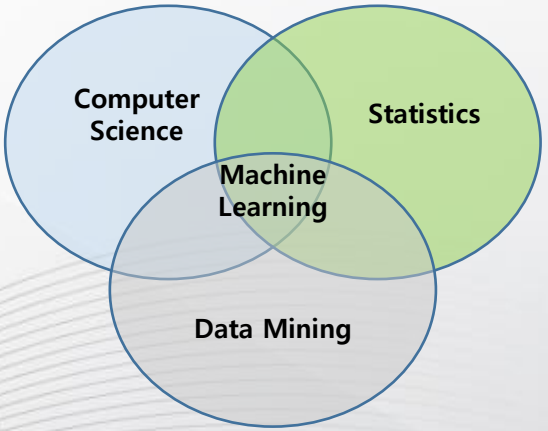
- ✓ 대량의 데이터로부터 숨겨진 패턴과 알려지지 않은 정보 간의 관계를 찾아내기 위한 과정

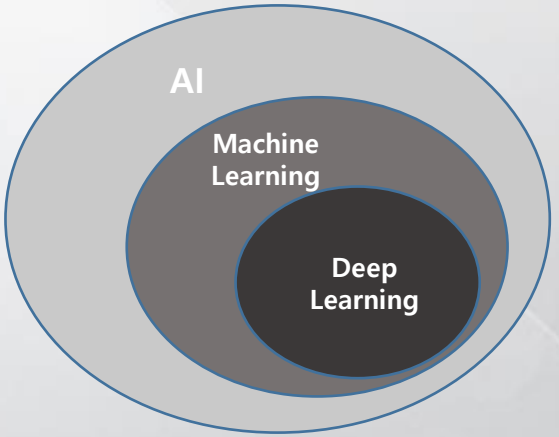
R

- ✓ 사용자가 쉽게 제공할 수 있는 R기반 다양한 패키지(라이브러리) 존재
- ✓ 다양한 시각화 방법이 있으며, 통계학자를 위한 언어이기 때문에 컴퓨터 공학이 아니더라도 접근이 쉬움

Python

- ✓ 쉬운 프로그래밍 환경을 이용할 수 있고, 통계에 좋은 라이브러리가 존재
- ✓ iPython Notebook 툴을 이용하면 다른 사람과 쉽게 코드를 공유





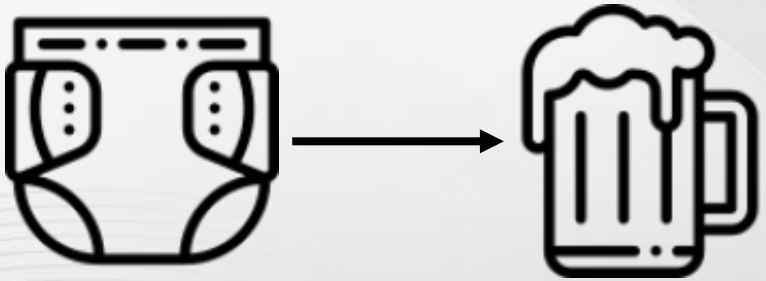
빅데이터 분석 기술

연관성 분석

- ✓ 예측이 목적이 아닌 단순한 데이터 간의 규칙을 발견하기 위한 분석
- ✓ 알고리즘에 대한 분류 기준이 필요하지않고, 규칙을 찾기 위한 규칙 학습기의 객관적 평가가 어려움

연관성 분석 예제

- ✓ 맥주와 기저귀
  - » 미국의 대형 편의점에서 일회용 아기 기저귀를 사는 사람은 맥주도 많이 산다는 연관성 규칙을 발견
  - » 실제 고객 조사결과 아내가 남편에게 기저귀를 사오라고 하면, 남편이 기저귀를 사면서 맥주도 같이 구입
  - » 맥주와 기저귀를 교차 판매하거나, 제품을 같이 진열하는 마케팅 전략에 활용 가능

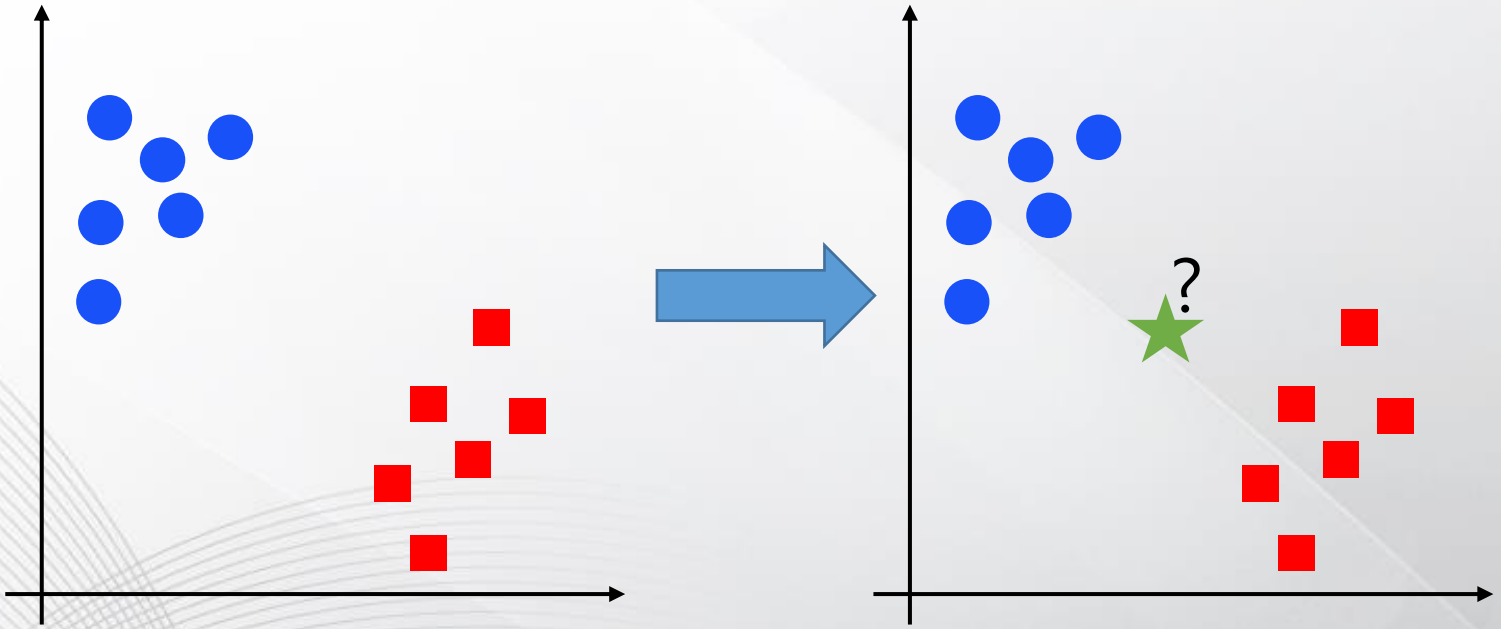


빅데이터 분석 기술

분류 분석

- ✓ 소속집단을 알고 있는 데이터를 이용하여 모형을 생성하고, 소속집단을 모르는 데이터의 집단을 결정하는 분석 방법

분류 분석 예제



빅데이터 분석 기술

군집 분석

✓ 각 객체의 유사성을 측정하여 집단을 분류하고, 군집에 속한 객체들의 유사성과 서로 다른 군집에 속한 객체간의 관계를 규명하는 통계 분석 방법

군집 분석 예제

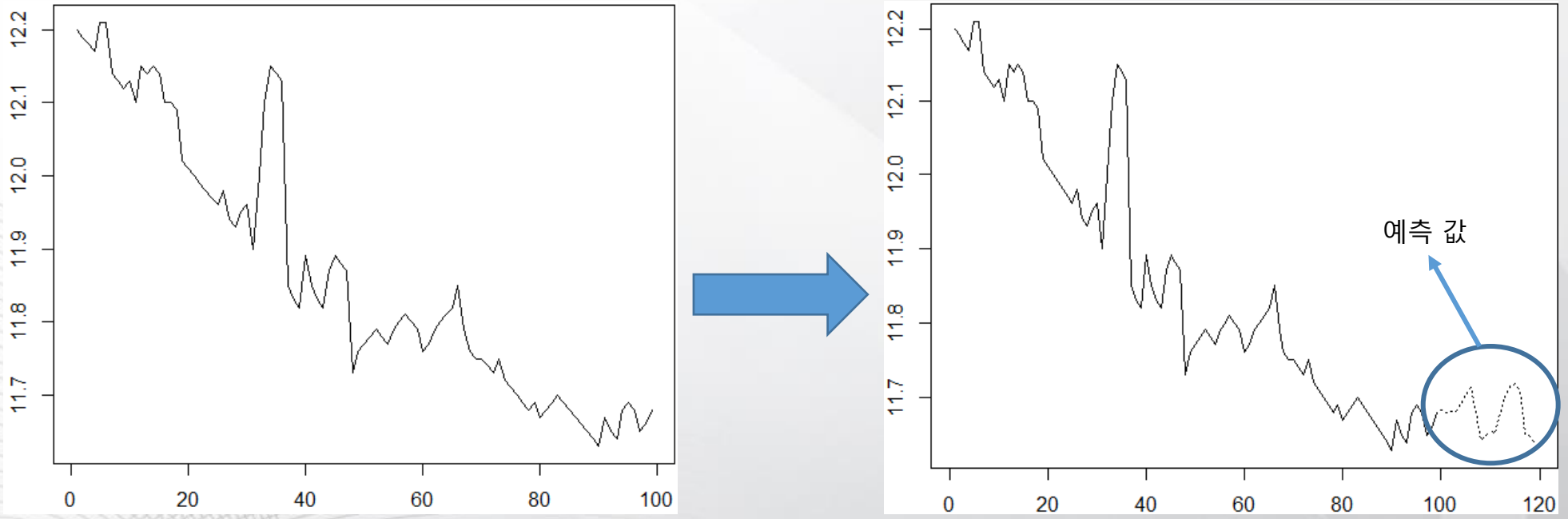
✓ [진돗개, 물소, 젖소, 포메라니안, 말티즈, 불독] 를 계층적 군집 분석

빅데이터 분석 기술

시계열 분석

- ✓ 시간의 흐름에 따라 저장되거나 기록된 데이터를 분석하고 변수들의 관계를 분석하는 방법
- ✓ 어떤 법칙에 의해 시계열이 그려지는지, 미래는 어떻게 예측 되는지 활용 가능

시계열 분석 예제

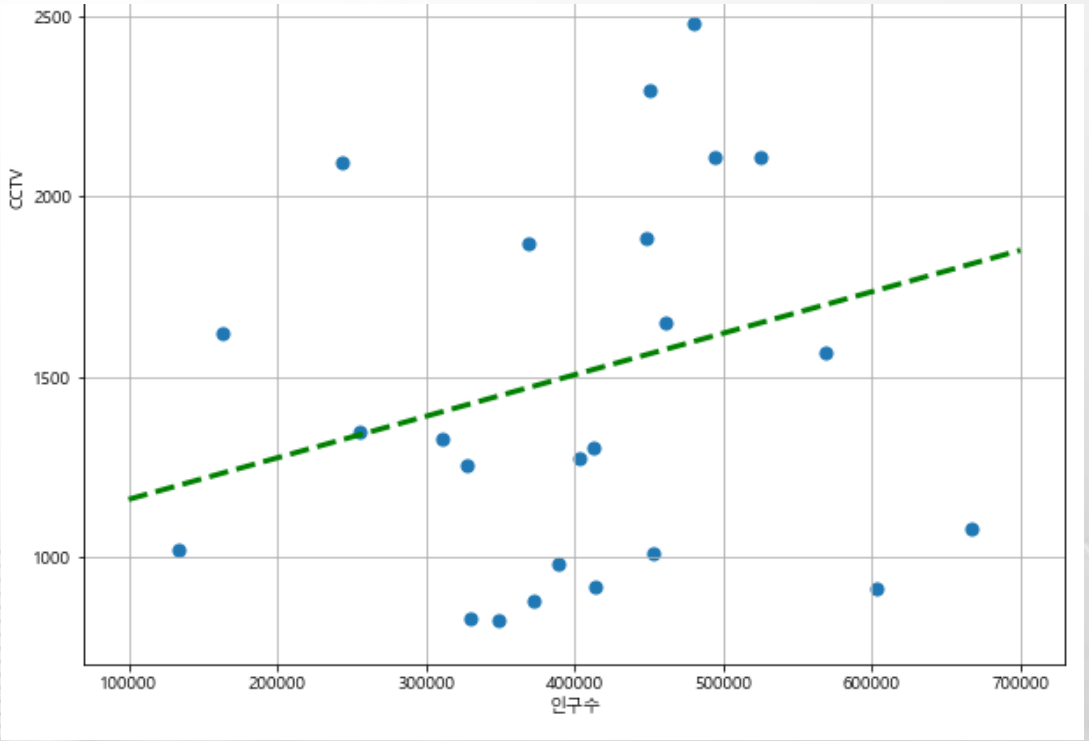


빅데이터 분석 기술

회귀 분석

- ✓ 변수들 사이의 관계를 나타내기 위해 모델(함수)을 가정하고, 모형에 측정된 변수들의 데이터로부터 추정되는 통계적인 방법
- ✓ 독립변수의 값에 의하여 종속변수의 값을 예측하기 위함

회귀 분석 예제



빅데이터 분석 기술

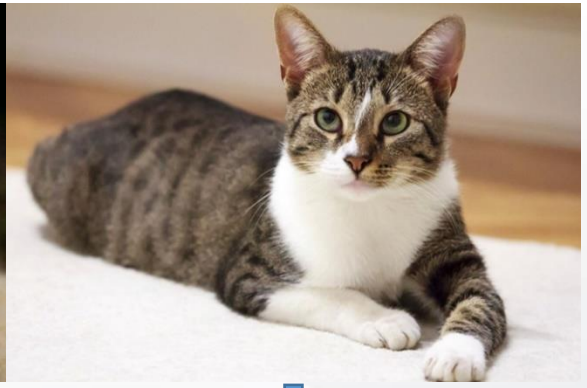
머신 러닝이란?

- ✓ 컴퓨터가 데이터를 반복적으로 학습해서 데이터에 포함된 패턴을 찾아내는 것
- ✓ 이 패턴을 모형화 한 것을 학습 모형이라 한다.
- ✓ 데이터에서 추출한 특징 값을 조합해서 패턴을 찾아내고 이를 모형으로 나타낸다.

빅데이터 분석 기술

데이터의 형식

✓ 데이터와 그에 대한 정답 데이터가 짝을 이루는 형태



개

고양이

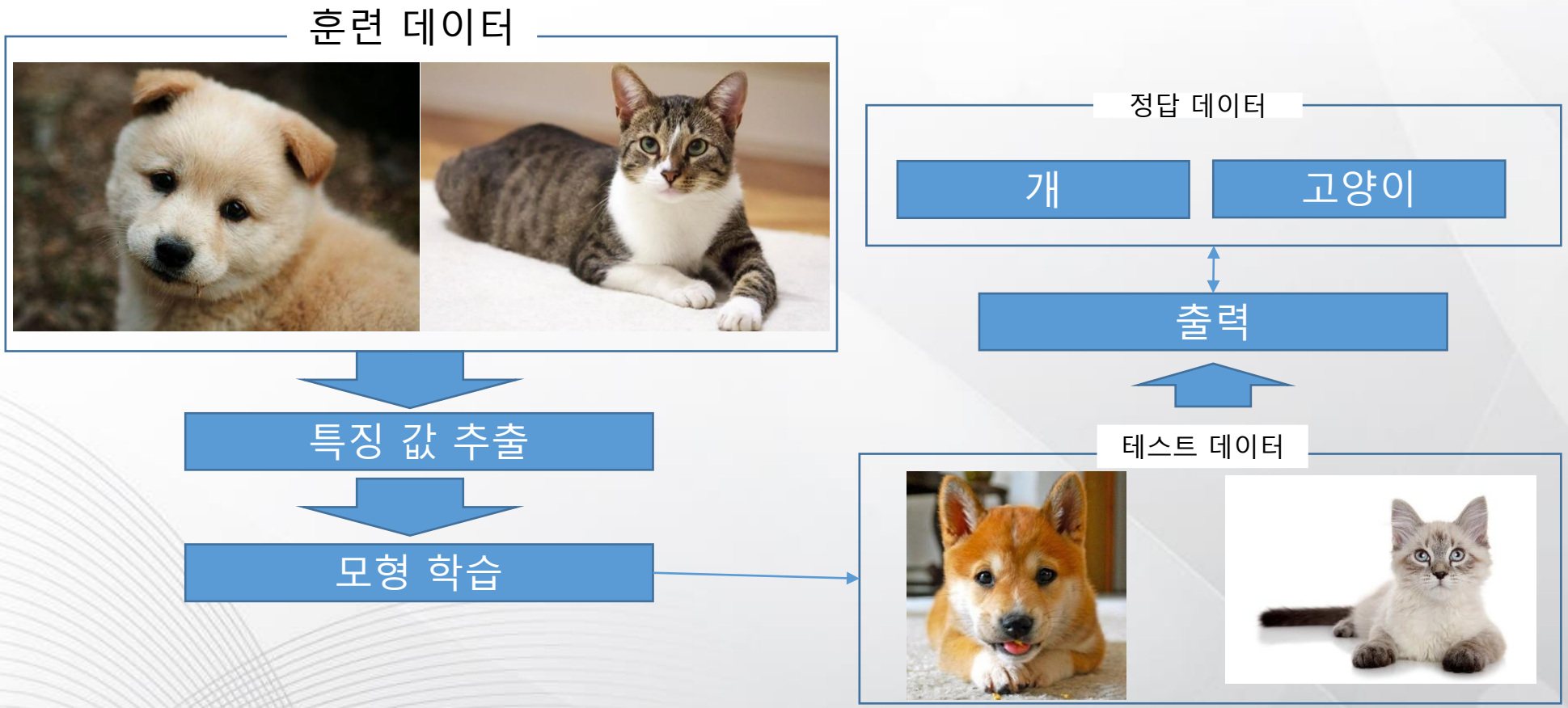
✓ 데이터 자체만을 포함하고 있는 형태

고객ID	연령	성별	…	구매금액
1	10대	여성	…	30000
2	20대	남성	…	17800
3	40대	여성	…	22100

빅데이터 분석 기술

지도학습

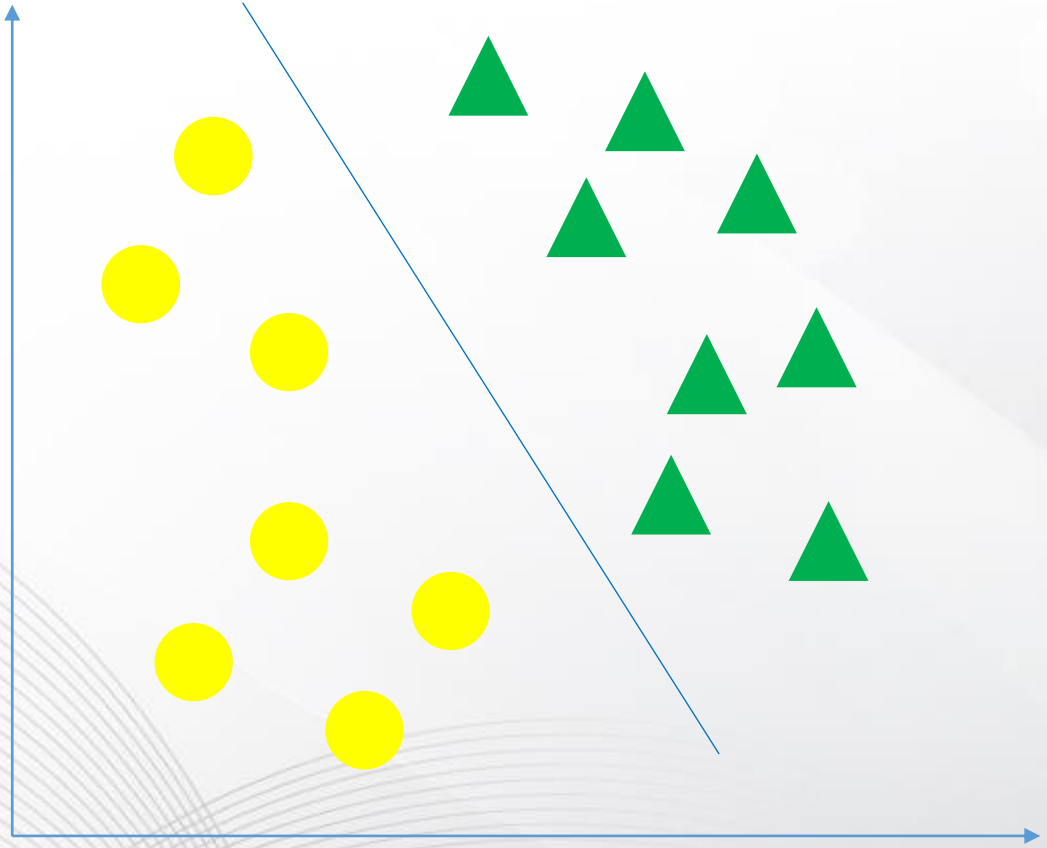
✓ 정답을 포함하는 데이터를 학습 데이터로 사용한다.



빅데이터 분석 기술

분류

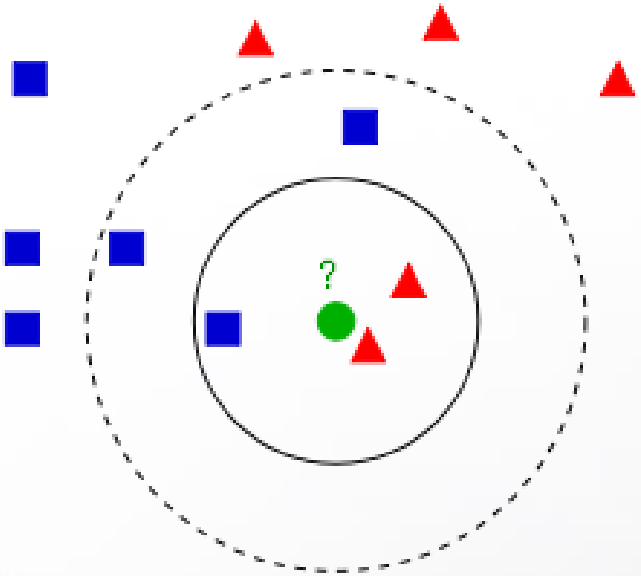
✓ 데이터가 미리 정해진 클래스 중 어느 것에 속하는지 예측



빅데이터 분석 기술

k-최근접이웃(k-nearest neighbor, KNN)

- ✓ 녹색 원이 어떤 클래스로 분류 되는 지를 결정하는 알고리즘

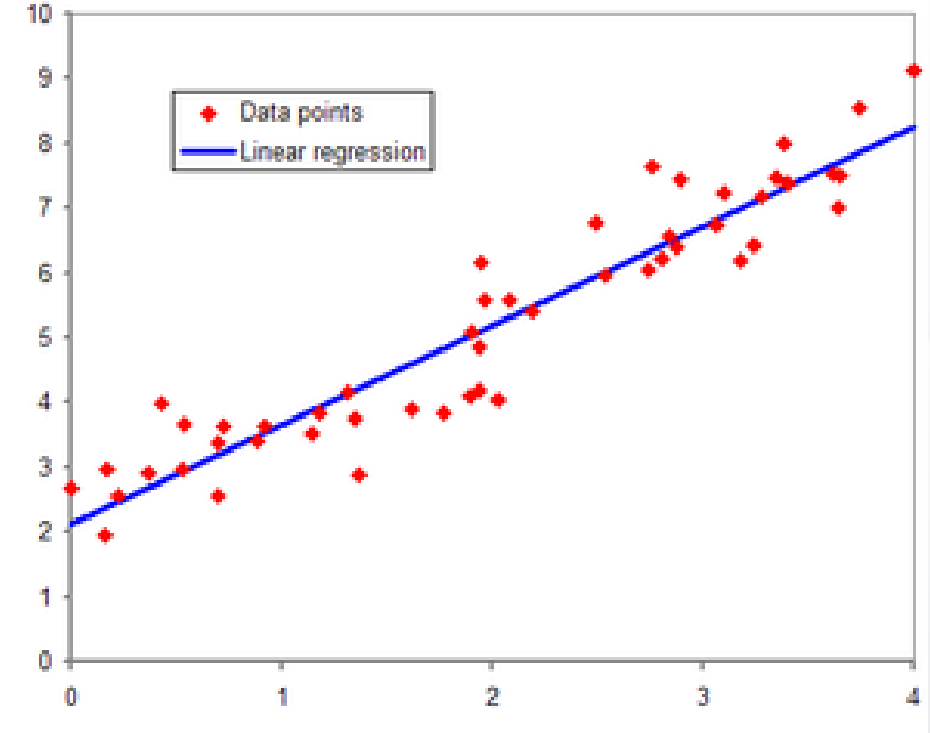


- ✓ K 값에 따라서 k=3 일 때와 k=5 일 때 다른 결과값이 나온다.
- ✓ K값이 너무 작으면 합리적이지 않은 분류가 될 수 있고, 너무 크다면 분류 자체를 못하게 되는 상황 발생
- ✓ 통상적으로 k는 전체 데이터의 제곱근 값으로 설정한다.
  - » 가장 최적의 k 값을 찾기 위한 연구가 진행중

빅데이터 분석 기술

회귀

✓ 데이터로부터 숫자를 예측



✓ Linear regression : 데이터와 부합하는 회귀 직선을 그어 예측

빅데이터 분석 기술

비지도학습

- ✓ 데이터 자체만을 포함하고 있는 형태의 데이터를 사용한다.
- ✓ 정답 데이터가 없기 때문에 정답을 맞추는 것을 목표로 패턴을 만드는 것이 불가능
- ✓ 특징 값에 기초해서 컴퓨터가 직접 패턴과 모형을 만들어야 한다.

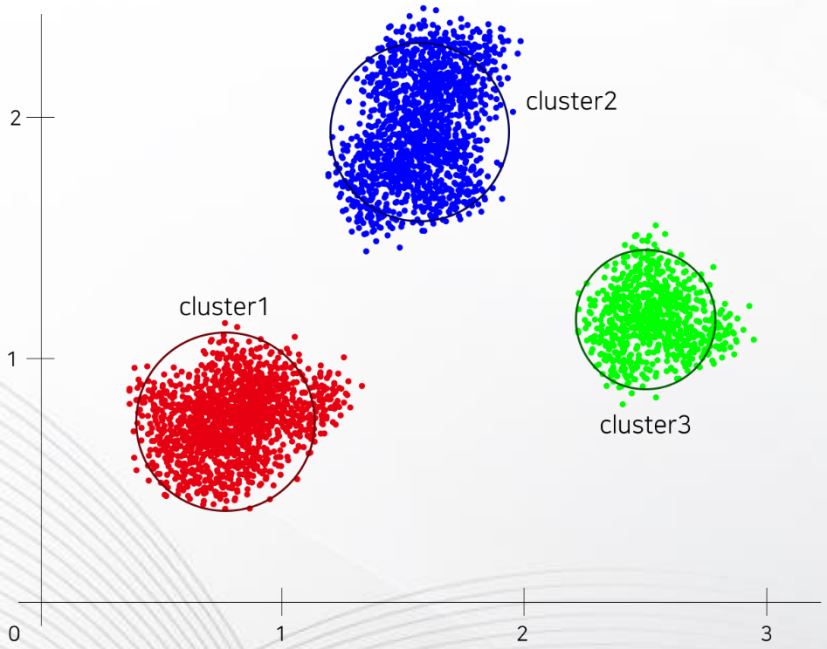
학습 데이터

고객ID	연령	성별	…	구매금액
1	10대	여성	…	30000
2	20대	남성	…	17800
3	40대	여성	…	22100

빅데이터 분석 기술

클러스터링

- ✓ 클러스터링(Clustering) : 군집화
- ✓ 데이터를 여러 그룹으로 나눔
- ✓ 대표적으로 Kmeans 알고리즘 사용

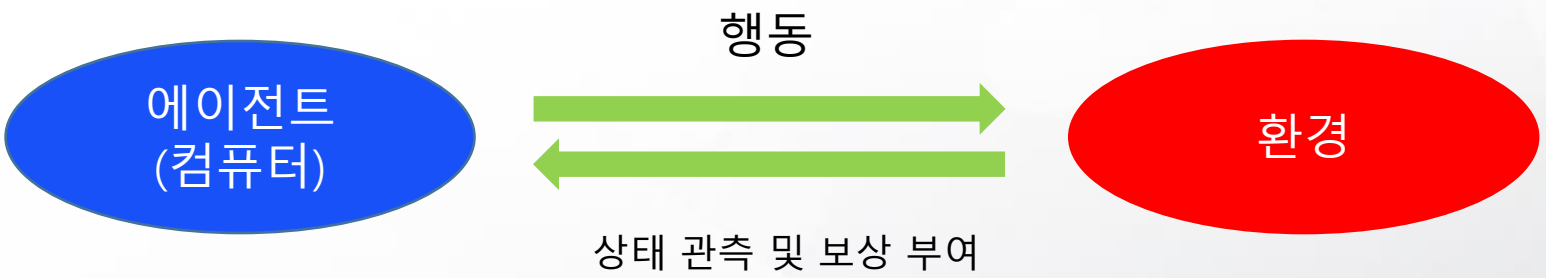


- ✓ 서로 거리가 가까운 데이터를 모아 3개의 그룹을 형성

빅데이터 분석 기술

강화 학습

✓ 컴퓨터가 스스로 학습을 통해 자신이 수행하는 처리를 최적화하고 목표를 달성하도록 하는 학습



» 에이전트는 자신이 처한 환경에서 행동을 취하고 자신의 상태를 관측한 결과와 보상을 받고 자신이 받을 보상을 최대화하는 행동을 학습하고 행동

» 이 과정을 반복하면서 행동을 최적화하고 보상을 극대화

Q & A

감사합니다