

# 빅데이터 시스템 구축 및 딥러닝 분석

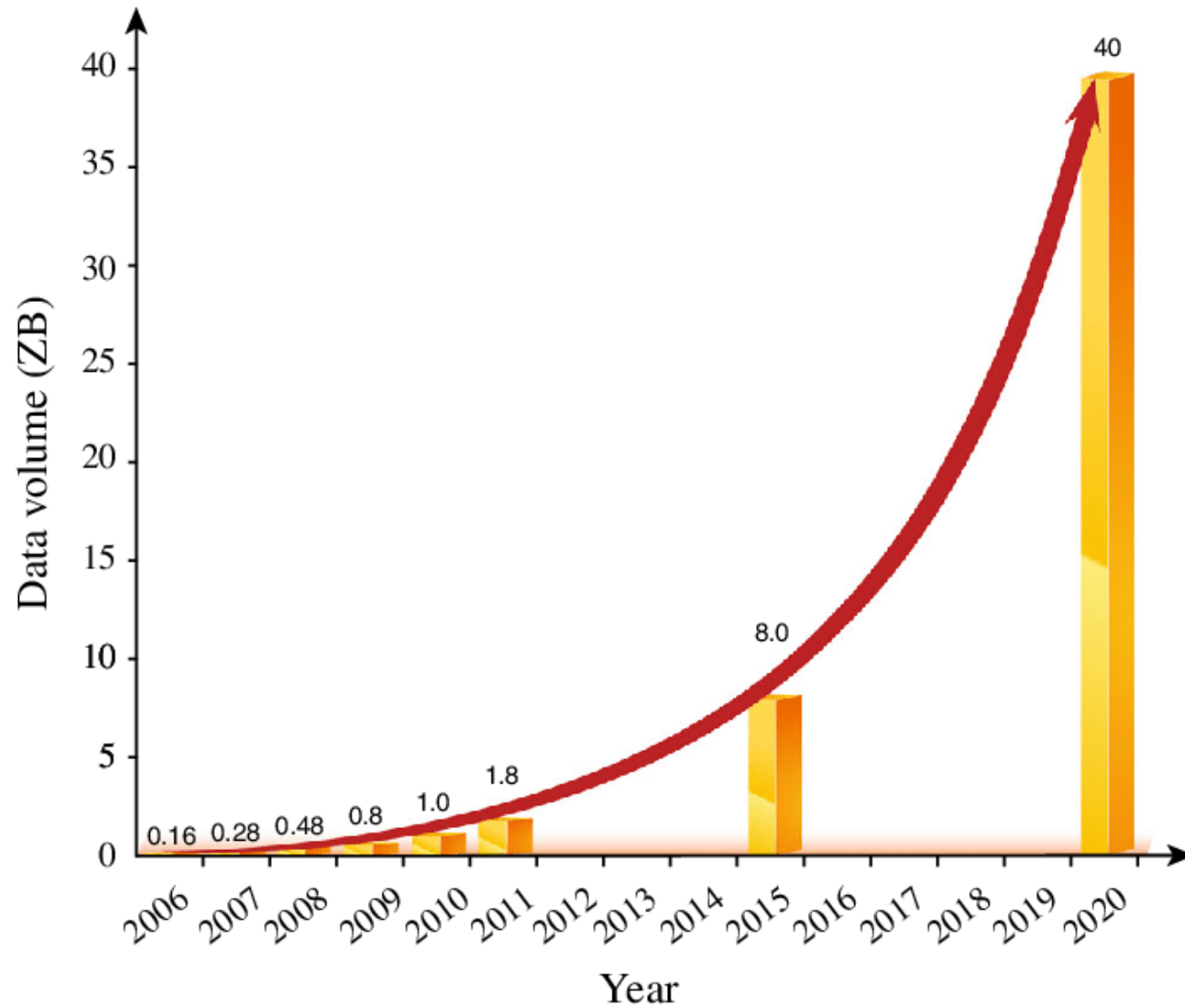
2022. 3. 29 ~ 4. 1

정 준 수 PhD

# 빅데이터 시스템 구축 및 딥러닝 분석 과정 진행 일정

일정	학습 내용	상세 내용
1일차	과정 소개	- 빅데이터 시스템 구축 및 딥러닝 분석 과정 소개
		- 빅데이터의 개념, 목적 및 활용 - 빅데이터 파일럿 프로젝트 도메인의 이해, 빅데이터 파일럿 아키텍처 이해
	점심시간	
		- 빅데이터 파일럿 프로젝트용 클러스터 환경 구축 - 빅데이터 수집 개요 및 기술, 빅데이터 수집 요구사항 및 아키텍처
2일차		- 빅데이터 수집 기능 구현 및 테스트 - 빅데이터 적재 개요 및 기술, 빅데이터 적재 요구사항 및 아키텍처
	점심시간	
		- 빅데이터 적재 기능 구현 및 테스트 - 빅데이터 실시간 적재 개요 및 기술, 빅데이터 실시간 적재 요구사항 및 아키텍처
3일차		- 빅데이터 실시간 적재 기능 구현 및 테스트 - 빅데이터 탐색 개요 및 기술, 빅데이터 탐색 요구사항 및 아키텍처
	점심시간	
		- 빅데이터 탐색 기능 구현 및 테스트 - 빅데이터 분석 개요 및 기술, 빅데이터 분석 요구사항 및 아키텍처
4일차		- 빅데이터 분석 기능 구현 및 테스트 - Python 기초 문법 - 머신러닝과 신경회로망
	점심시간	
		- 케라스를 이용한 기초 신경망 구현 - 케라스를 이용한 딥러닝 - 신경망 모델을 이용한 예측 - 분석 및 예측 결과 시각화

# 데이터의 증가 추세





# Database Evolution History

1960s

First Computerized Database Models



1970s

The Dawn of the Database

- The relational model and its language SQL emerge
- The disruptive model causes the demise of other models

1970 E.F. Codd Writes a Paper on the Relational Database Model



1980s

An Industry Develops

- SQL becomes the de-facto standard
- Commercial offerings from IBM, Oracle grow market
- Other data models enter the scene, without much traction

ORACLE  
1st commercially available RDBMS

IBM DB2

SAP Sybase  
Informix



1990s

Technology Shifts

- Data explodes with the Internet age
- Single server SQL databases run into resource problems
- Business Intelligence and Analytics move out of transactional databases



Teradata



2000s

New Players Emerge

- Data variety, velocity and volume increase
- New analytics SQL databases are introduced
- NoSQL databases fill the gap for processing unstructured data
- Hadoop gains traction for analyzing petabytes of data

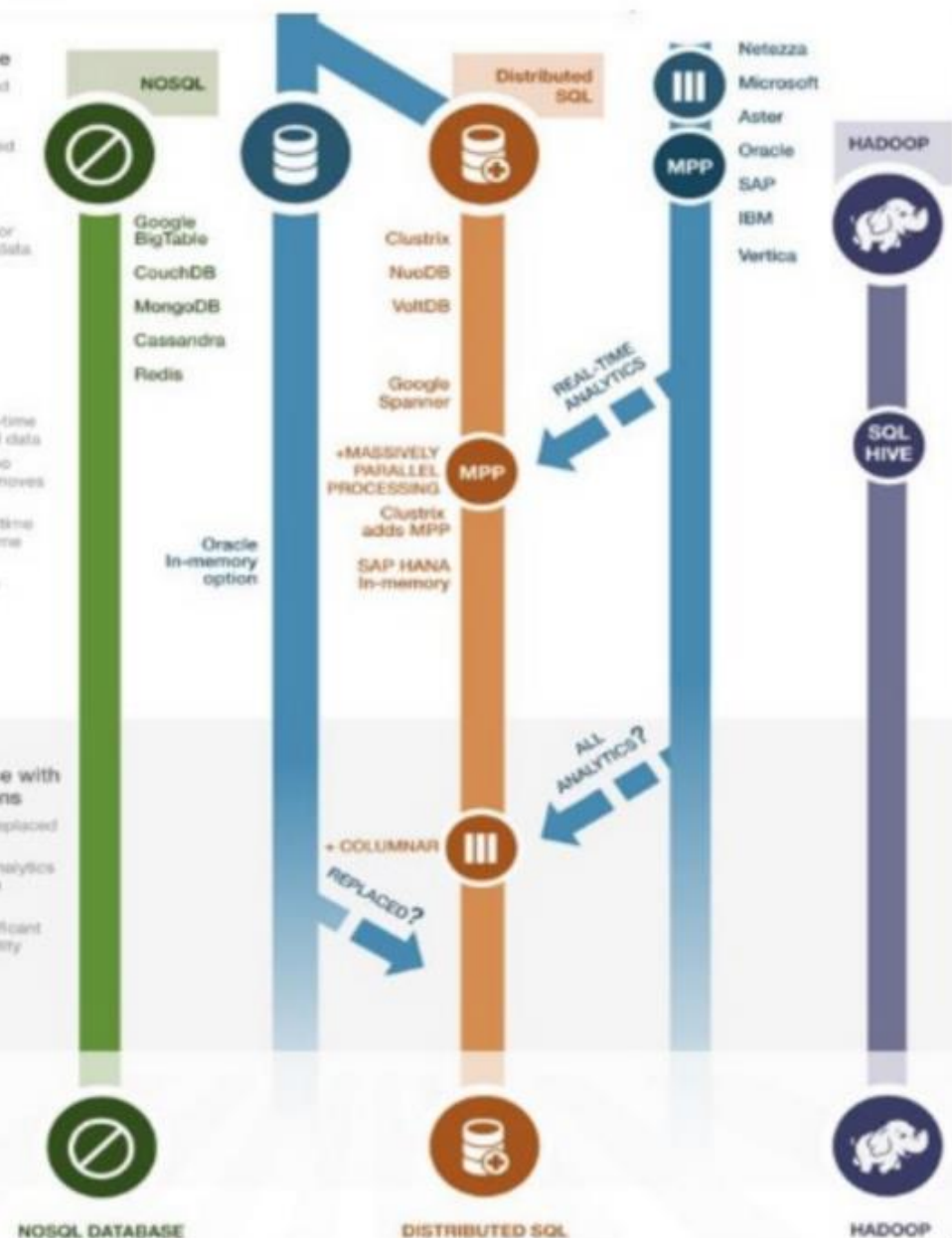
Today  
Databases Adapt and Evolve

- Businesses require real-time analytics on operational data
- Scale-up SQL proves too costly, but scale-out removes resource constraint
- Scale-out provides real time analytics with high volume transactions
- Google and Clustrix are pioneers in this space

The Future  
Businesses Advance with Database Innovations

- Single node SQL gets replaced by scale-out SQL
- Data warehouse type analytics will become available in real-time database
- Businesses gain a significant edge and increased agility

Winning Database Platforms





# Apache Hadoop Ecosystem



**Ambari**

Provisioning, Managing and Monitoring Hadoop Clusters



**Scoop**  
Data Exchange



**Zookeeper**  
Coordination



**Flume**  
Log Collector



**Oozie**  
Workflow



**Pig**  
Scripting



**Mahout**  
Machine Learning

**R Connectors**  
Statistics



**Hive**  
SQL Query



**Hbase**  
Columnar Store



**YARN Map Reduce v2**

Distributed Processing Framework

**HDFS**

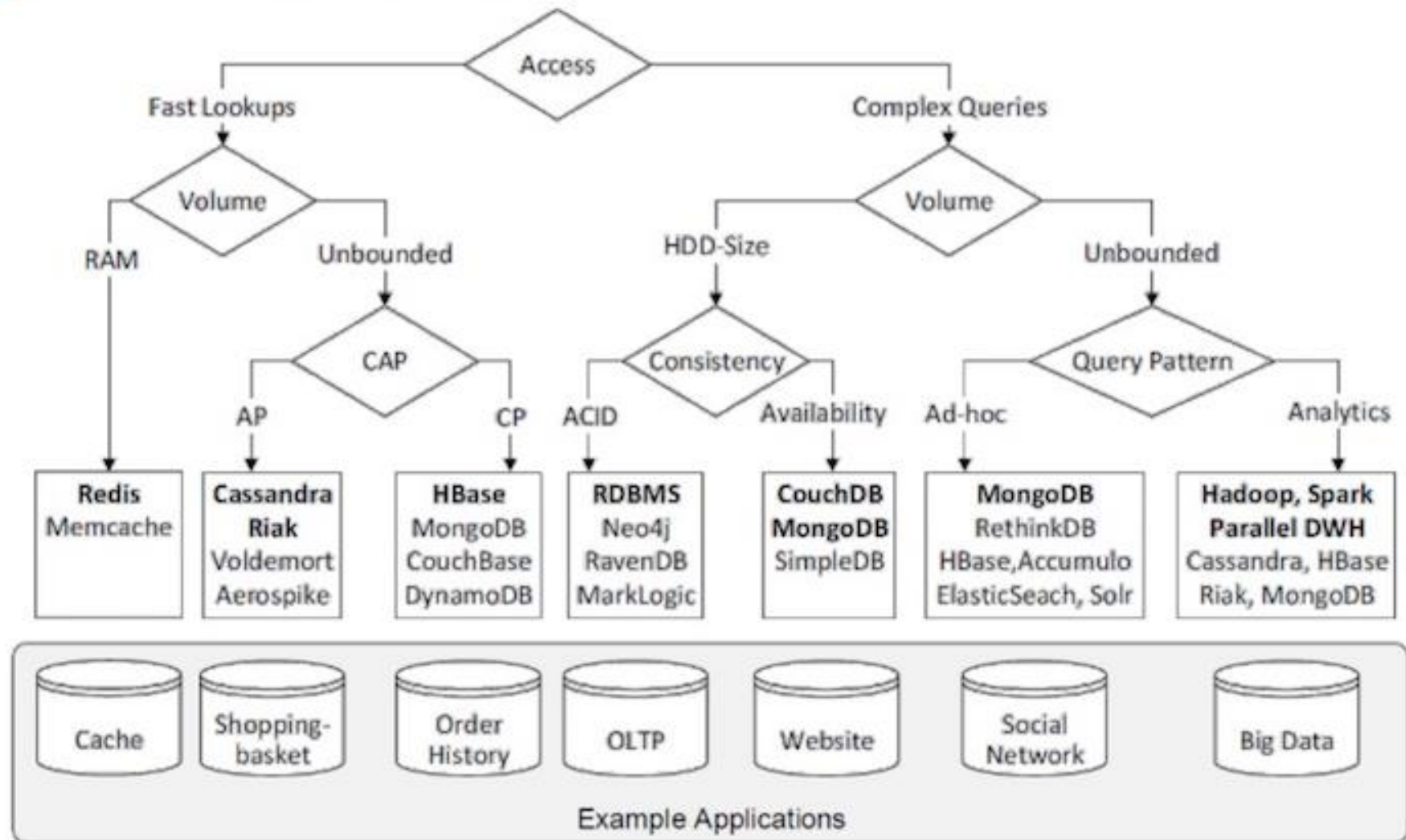
Hadoop Distributed File System







# NoSQLDecisionTree



# 빅데이터의 개념(정의), 목적 및 활용

"대용량 데이터를 활용/분석해서 가치 있는 정보를 추출하고, 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 정보화 기술"

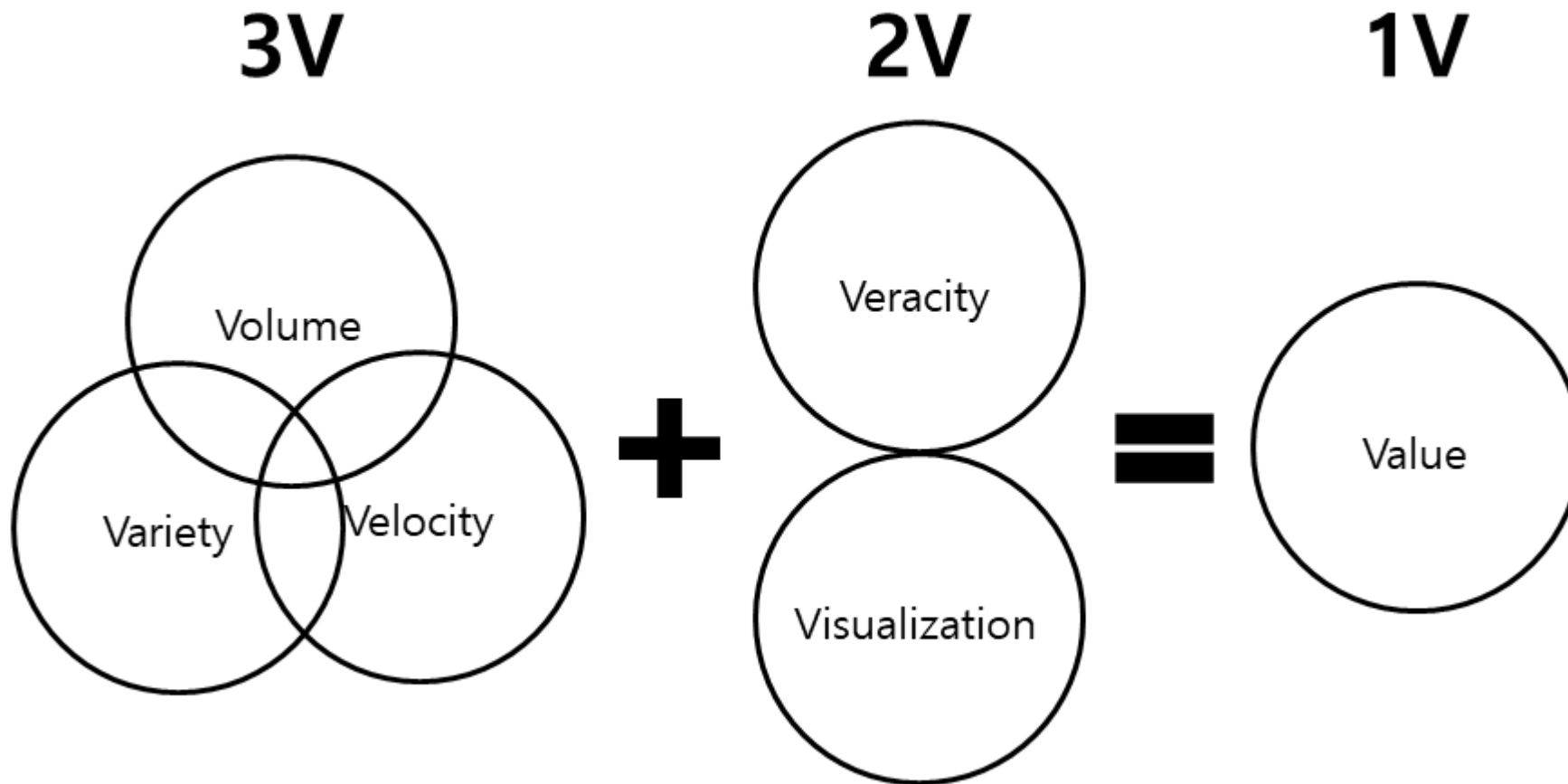
– 국가정보화전략위원회

"단순한 데이터의 크기가 아니라 데이터의 형식과 처리 속도 등을 함께 아우르는 개념으로 기존 방법으로는 데이터의 수집, 저장, 검색, 분석 등이 어려운 데이터를 총칭해서 일컫는 용어"

– ITWorld, 2012

2011년 메타그룹(현 가트너)의 애널리스트인 더그레이니(Doug Laney)는 다소 혼란스러운 빅데이터의 정의를 3V라는 표현으로 매우 명확하게 정리했는데, 이는 데이터의 크기(Volume), 데이터 입출력 속도(Velocity), 데이터 종류의 다양성(Variety)이라는 세 개의 차원으로 빅데이터를 정의

# 빅데이터의 정의 6V



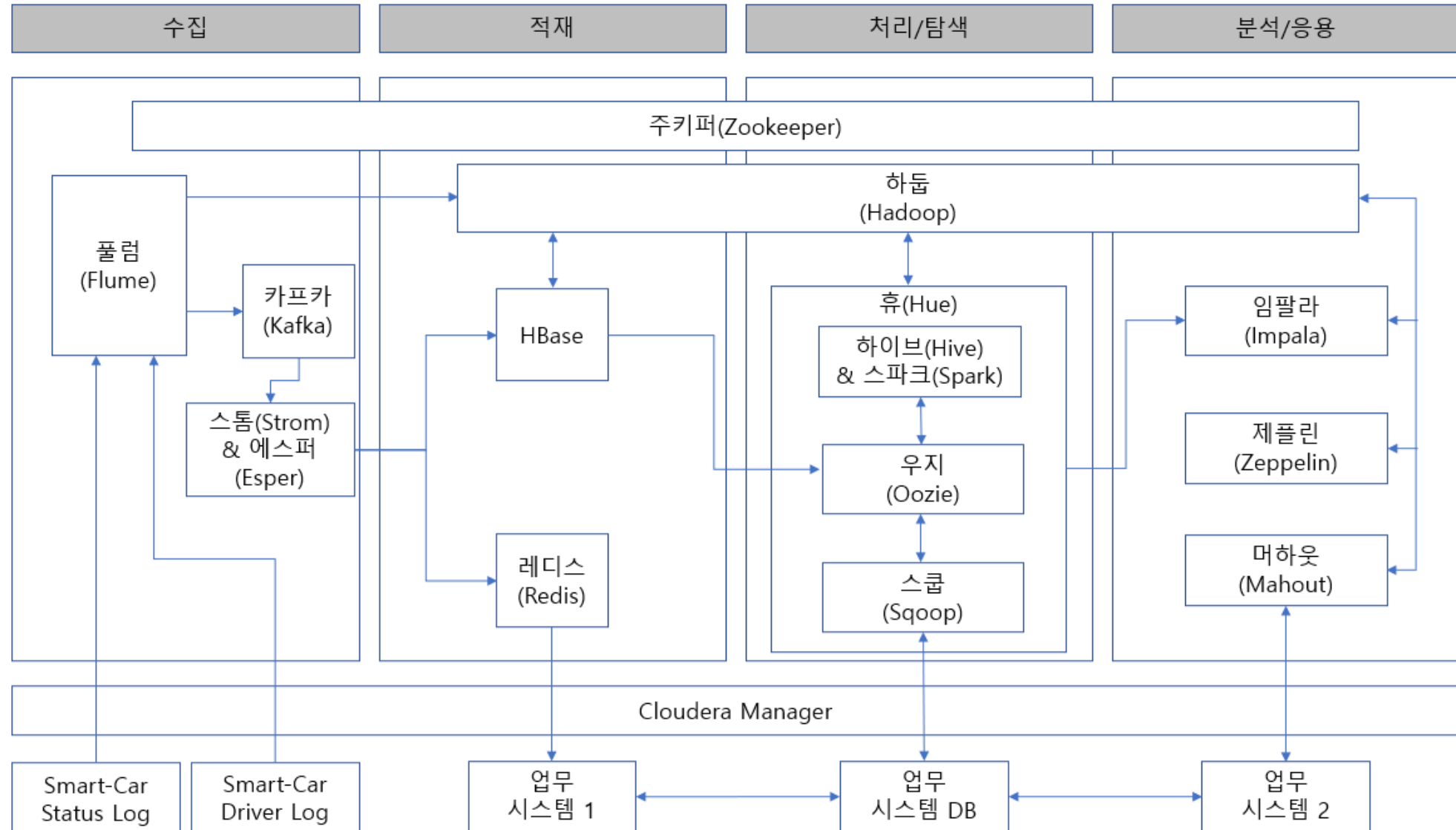


# 빅데이터의 정의 6V

IBM이 진실성(Veracity)이라는 요소를 더해 4V를 정의했고, 이후에 시각화(Visualization) 와 가치(Value)가 추가로 정의되면서 6V까지 확장됐다.

- 크기(Volume): 방대한 양의 데이터(테라, 페타바이트 이상의 크기)
- 다양성(Variety): 정형(DBMS, 전문 등) + 비정형(SNS, 동영상, 사진, 음성, 텍스트 등)
- 속도(Velocity): 실시간으로 생산되며, 빠른 속도로 데이터를 처리/분석
- 진실성(Veracity): 주요 의사결정을 위해 데이터의 품질과 신뢰성 확보
- 시각화(Visualization): 복잡한 대규모 데이터를 시각적으로 표현
- 가치(Value): 비즈니스 효익을 실현하기 위해 궁극적인 가치를 창출

# 빅데이터 시스템 파일럿 아키텍처



# 빅데이터 인사이트

- 첫 번째 현상 이해에서는 대규모 데이터로부터 통계량을 추출해 과거에 발생한 일에 대한 이해와 원인을 파악하고,
- 두 번째 현상 발견에서는 지금까지 알지 못했던 데이터 패턴들을 발견하고 해석해 무슨 일이 새롭게 일어났는지를 알아낸다.
- 세 번째 현상 예측에서는 이해와 발견을 기반으로 예측 모형(모델)을 만들고, 현재 발생하고 있는 데이터를 모형에 입력해 미래에 발생할 현상을 예측하게 된다.

보통 빅데이터 시스템을 도입한 후 현상 이해를 시작으로 발견과 예측의 인사이트 단계로 발전해 나간다. 특히 현상 예측은 머신러닝(딥러닝) 같은 고급 분석 기술을 이용해 예측 모델을 만들어 업무 시스템에 적용해 최적화까지 진행하는 단계로서 빅데이터에 대한 거버넌스와 함께 높은 기술 수준까지 요구된다.

# 빅데이터 파일럿 프로젝트 클러스터 환경 구축

[https://github.com/JSJeong-me/KOSA\\_BIGDATA\\_DEEPLARNONG](https://github.com/JSJeong-me/KOSA_BIGDATA_DEEPLARNONG)

정 준 수 / Ph.D ( jsjeong@hansung.ac.kr )

- 前) 삼성전자 연구원
- 前) 삼성의료원 (삼성생명과학연구소)
- 前) 삼성SDS (정보기술연구소)
- 現) (사)한국인공지능협회, AI, 머신러닝 강의
- 現) 한국소프트웨어산업협회, AI, 머신러닝 강의
- 現) 서울디지털재단, AI 자문위원
- 現) 한성대학교 교수(겸)
- 전문분야: Computer Vision, 머신러닝(ML), RPA
- <https://github.com/JSJeong-me/>

