KOSA
Korea Software Industry Association

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

이번 과정에서 다룰 내용들

빅데이터

한대의 컴퓨터로는 저장하거나 연산하기 어려운 규모의 거대 데이터

분산

여러대의 컴퓨터로 나눠서 일을 처리함

저장

여러대의 컴퓨터에 나눠서 저장한다.

분석

데이터가 저장된 컴퓨터에서 데이터를 분석하고 그 결과를 합친다.

이외에 비정형데이터의 분석(NoSQL), 네트워크 등

빅데이터의 개념(정의), 목적 및 활용

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

- 국가정보화전략위원회

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

- ITWorld, 2012

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

빅데이터 시스템 구축 및 딥러닝 분석 과정 소개

스마트카의 빅데이터 파일럿 프로젝트를 단계별로 진행하면서 빅데이터의 수집/적재, 처리/탐색, 분석/응용 영역의 아키텍처와 활용 기술에 대해 설명

•빅데이터의 수집/적재

- 빅데이터의 개요와 파일럿 프로젝트의 도메인을 이해하고, 파일럿 실습 환경을 구성한다
- 플럼, 카프카를 이용해 스마트카에서 발생하는 상태 정보와 운전자의 운행정보를 수집한다.
- 스톰, 에스퍼, 하둡, HBase, 레디스로 스마트카의 대용량 파일과 실시간 데이터를 적재한다.

•빅데이터의 처리/탐색

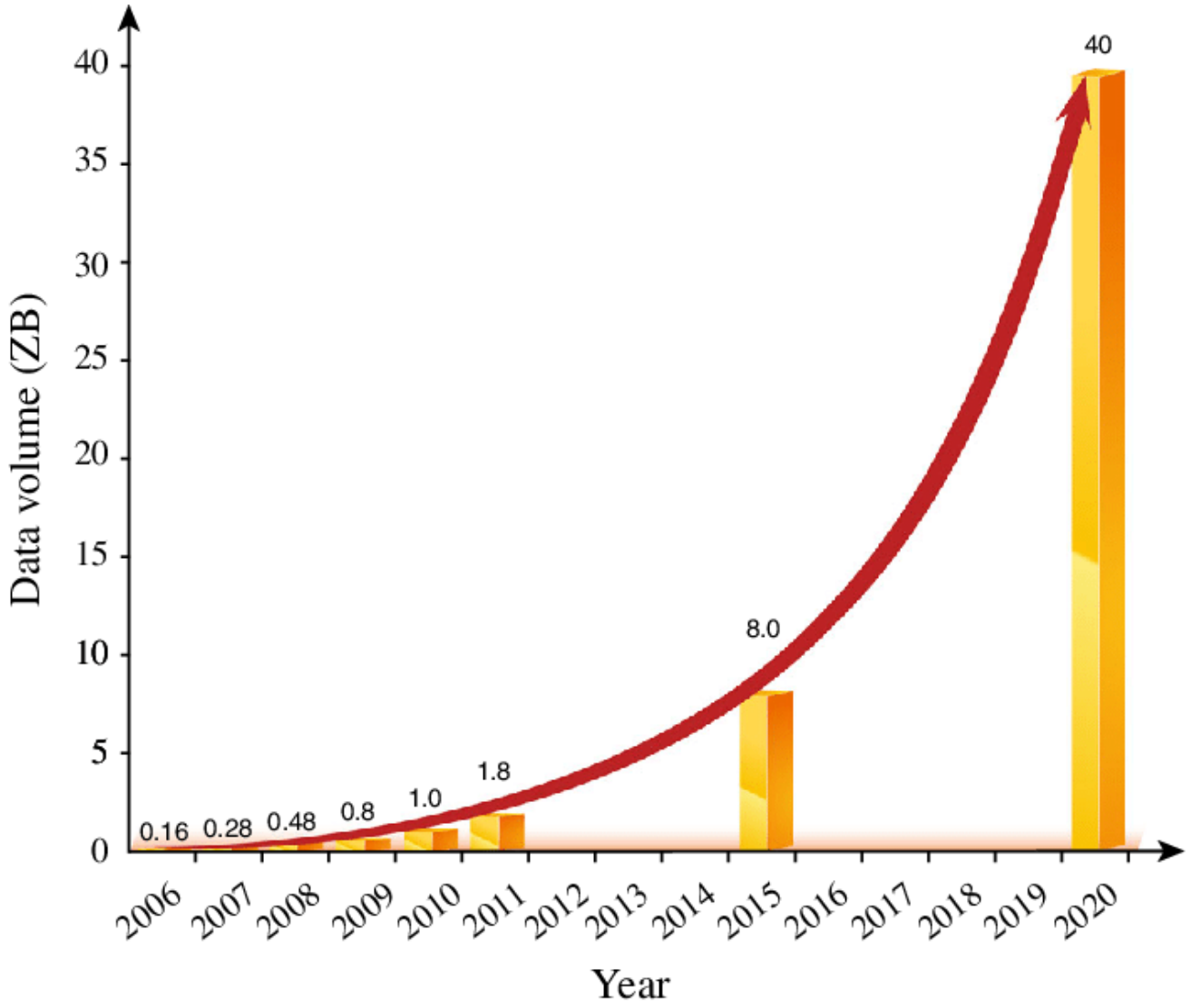
- 하이브, 스파크SQL의 애드혹 쿼리로 데이터 선택, 변환, 통합, 축소 등의 전처리 작업을 한다.
- 휴, 우지를 통해 데이터 가공/탐색 과정을 프로세스화해서 빅데이터 마트를 구성한다.
- 빅데이터 분석 결과를 하둡에 저장하고 스쿱을 이용해 외부 RDMS에 Export한다.

빅데이터 시스템 구축 및 딥러닝 분석 과정 소개

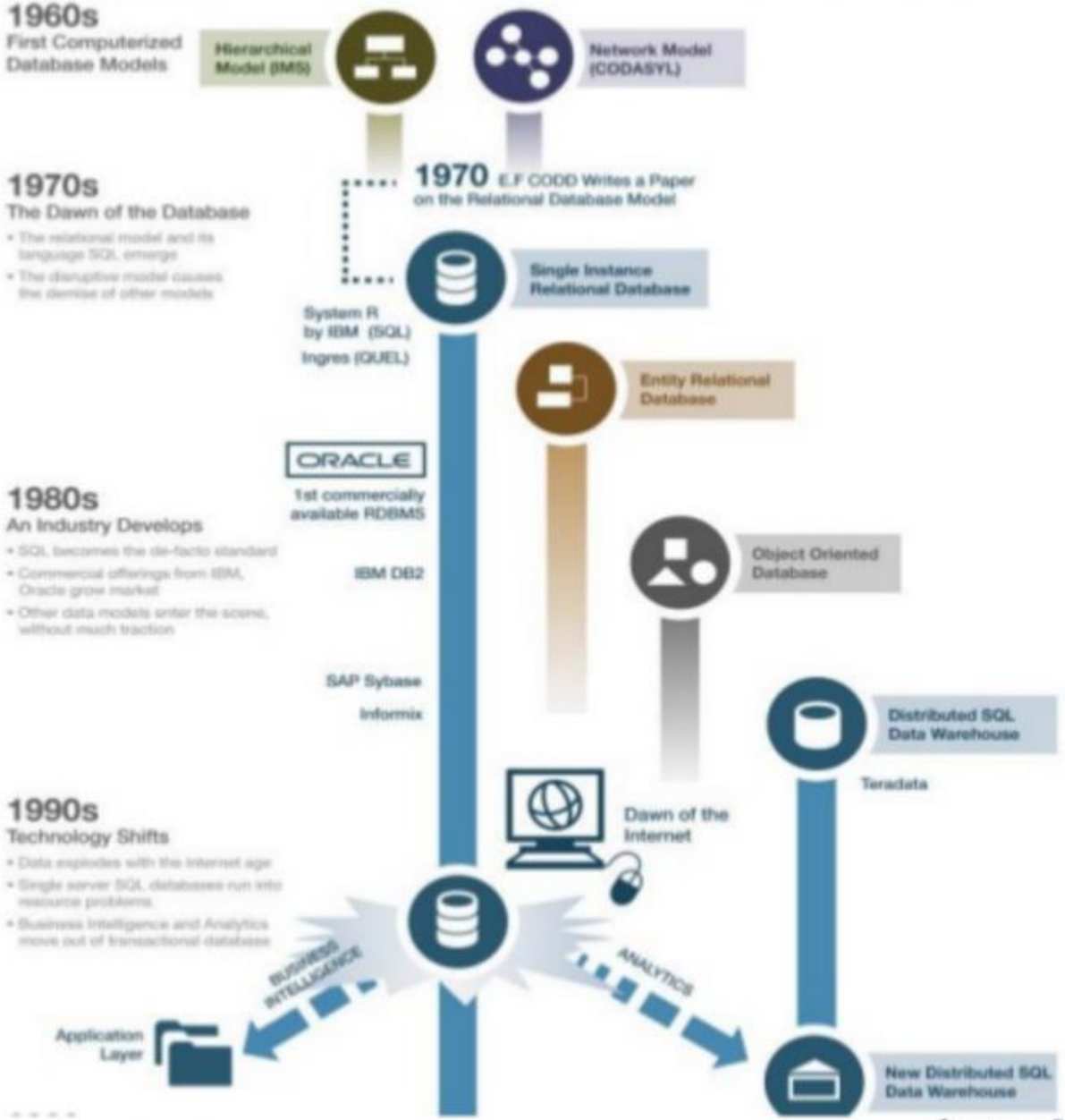
•빅데이터의 분석/응용

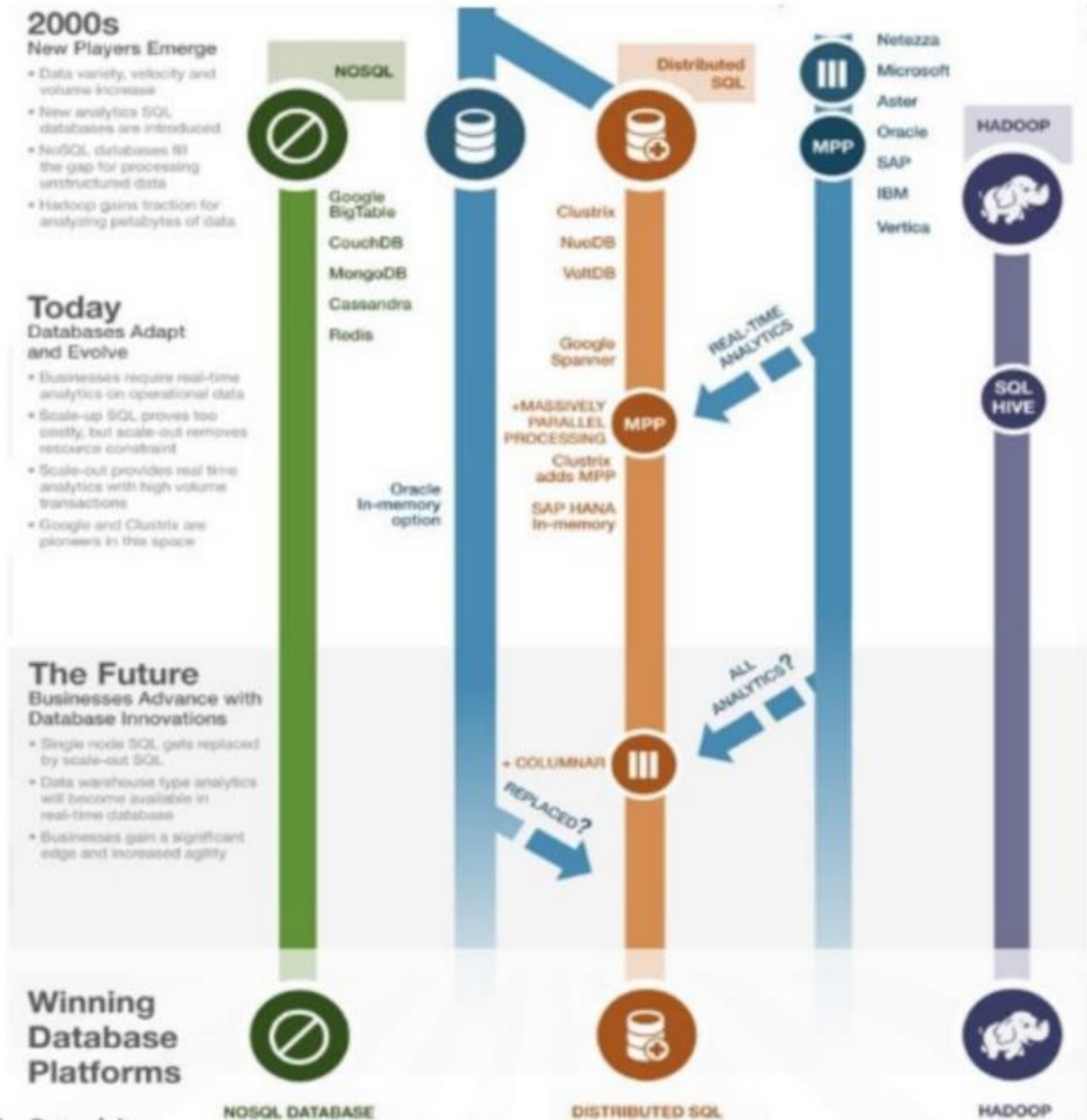
- 임팔라, 제플린으로 스마트카 데이터를 대상으로 고성능 인메모리 분석을 수행하여 인사이트를 발견하고 결과를 시각화한다.
- 스파크ML과 머하웃으로 스마트카의 마트 데이터를 활용해 추천, 분류, 군집 등의 머신러닝 분석을 진행한다.
- Python을 이용해 스마트카 운전자의 연소득 회귀모델과 텐서플로/케라스로 주행중 차량의 이상탐지 딥러닝 모델을 만들어 REST API 서비스를 구성한다.

데이터의 증가 추세

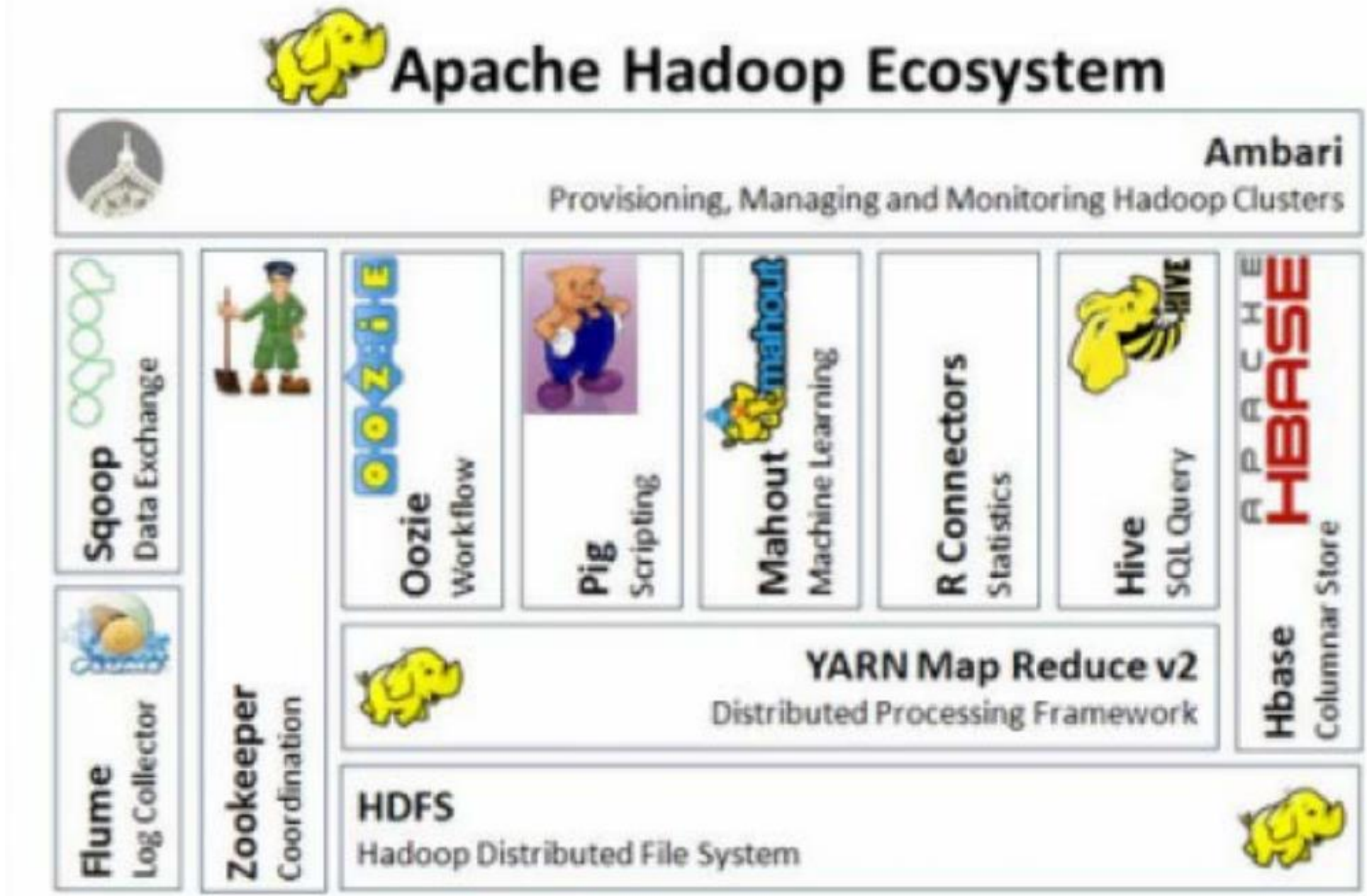

Data volume (ZB)
0
5
10
15
20
25
30
35
40
0.16
0.28
0.48
0.8
1.0
1.8
8.0
40
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year

Database Evolution History

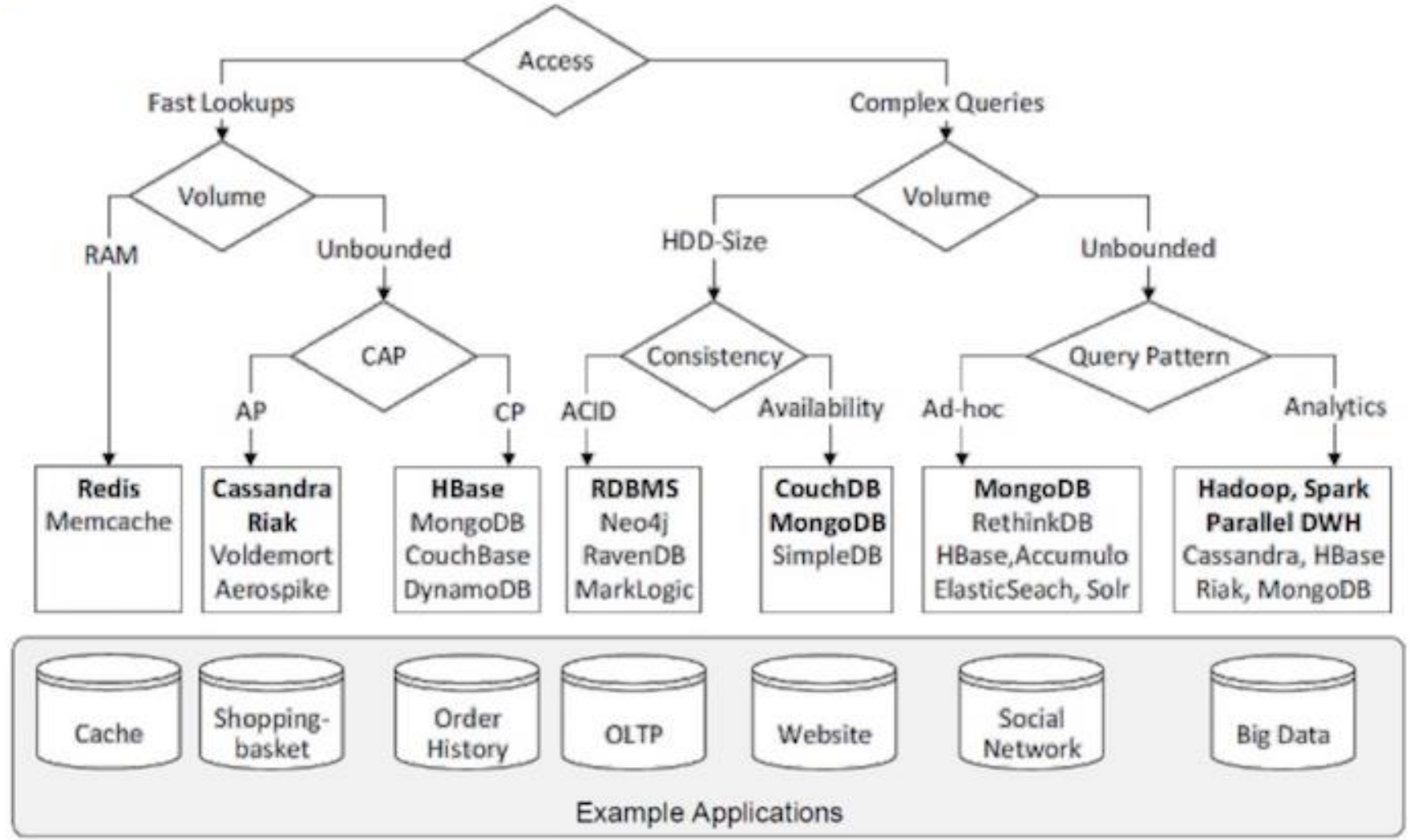




Source: Robin Purohit


Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
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

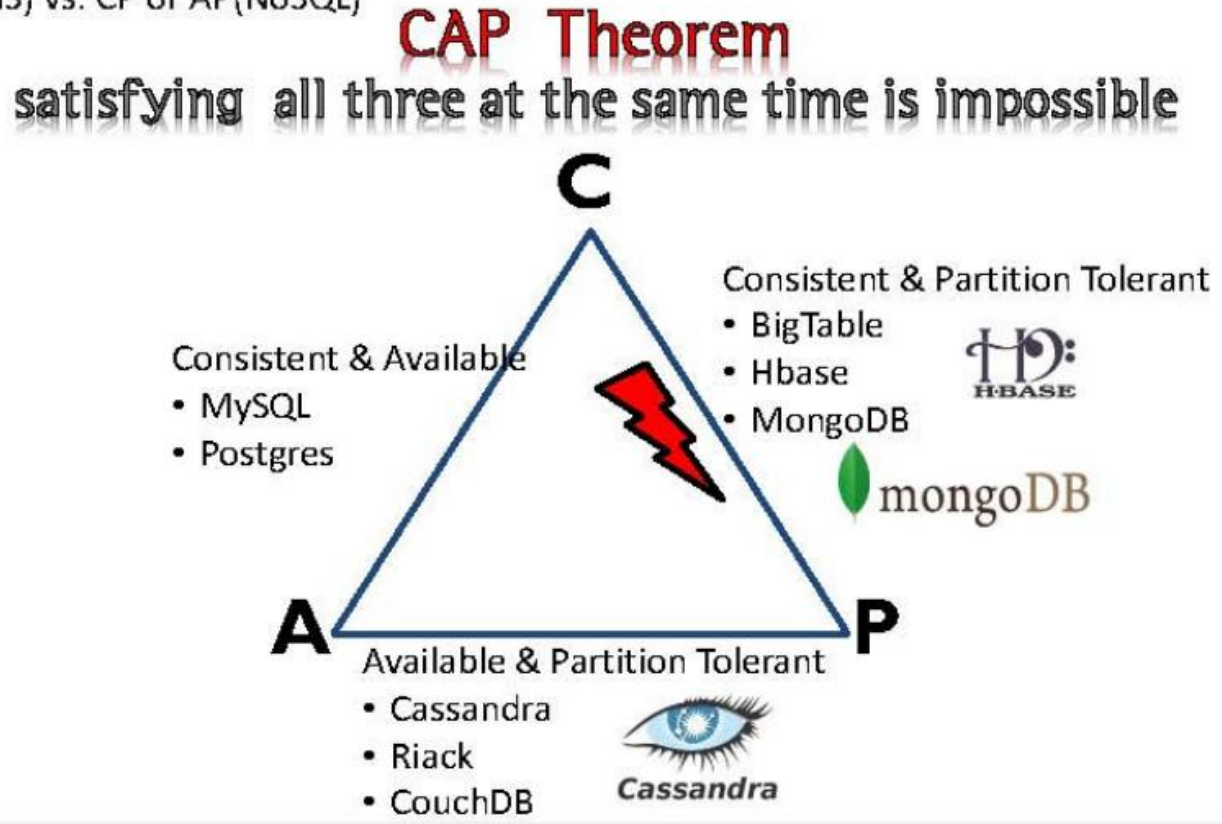
NoSQLDecisionTree



Source: Felix Gessert

NoSQL

- 샌프란시스코 Bay Area, 대용량 오픈소스 데이터베이스 토론 그룹에서 처음 사용
- NoSQL은 데이터를 빠르고(Rapid) 효율적인(Efficient) 처리를 가능하게 하는 개념의 집합으로, 성능, 신뢰성, 민첩성에 초점을 두고 있다.
  - CA(RDBMS) vs. CP or AP(NoSQL)



MapReduce

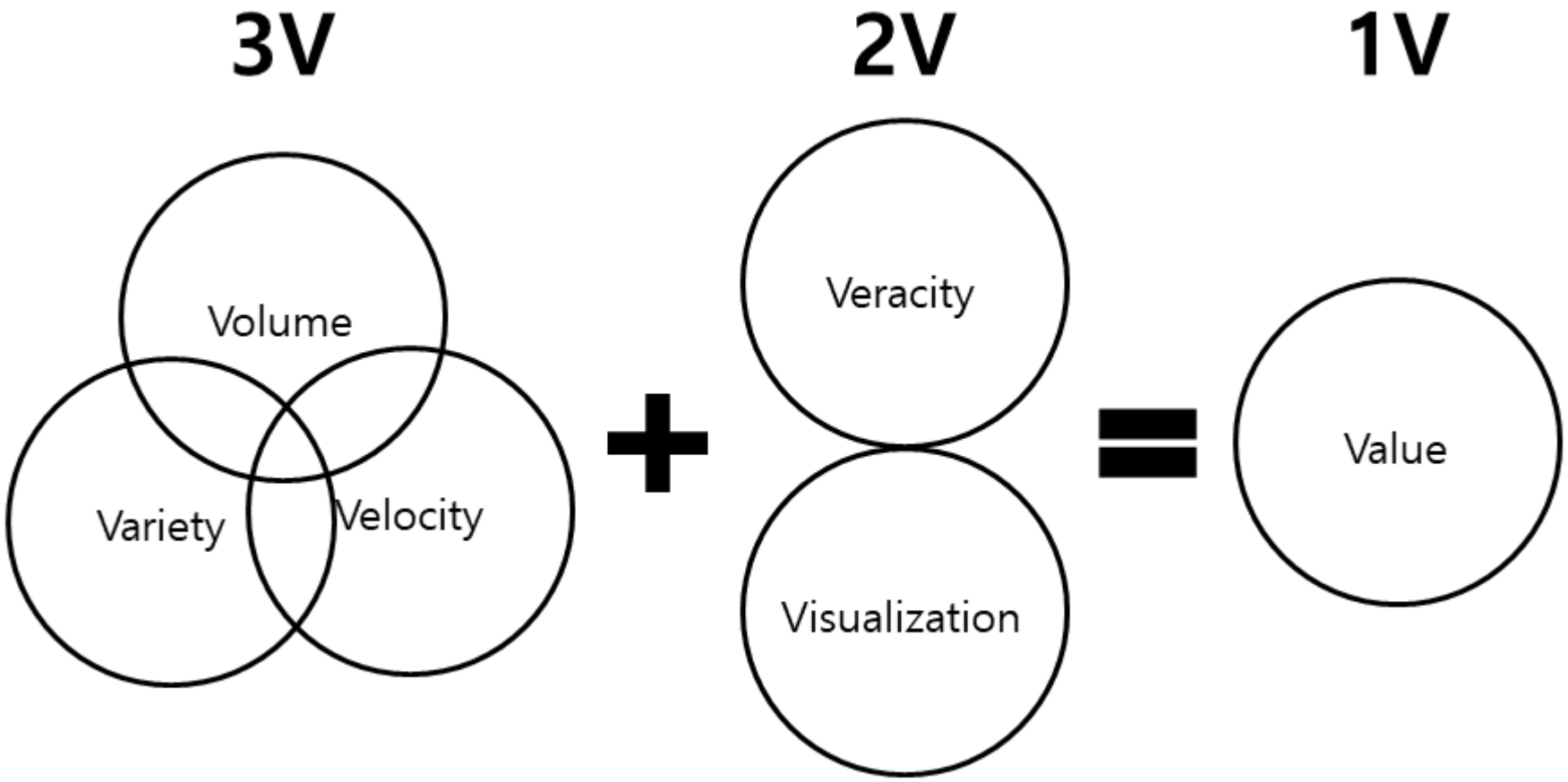
- 정의
  - MapReduce는 구글에서 분산 컴퓨팅을 지원하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임워크로 HDFS에 저장된 파일이용
- 구성
  - 프레임워크는 함수형 프로그래밍에서 일반적으로 사용되는 map()과 reduce() 함수 기반으로 주로 구성.
  - map() : (key, value) 쌍을 처리하여 또 다른 (key ,value) 쌍을 생성하는 함수
  - reduce() : 맵으로부터 생성된 (key, list(value))들을 병합(merge)하여 최종적으로 list(value) 들을 생성하는 함수
- 목적
  - 페타바이트(Petabyte) 이상의 대용량 데이터를 신뢰할 수 없는 컴퓨터로 구성된 클러스터 환경에서 병렬 처리를 지원하기 위해서 개발
- 단점
  - Java 언어를 습득할 필요.

빅데이터의 정의 6V

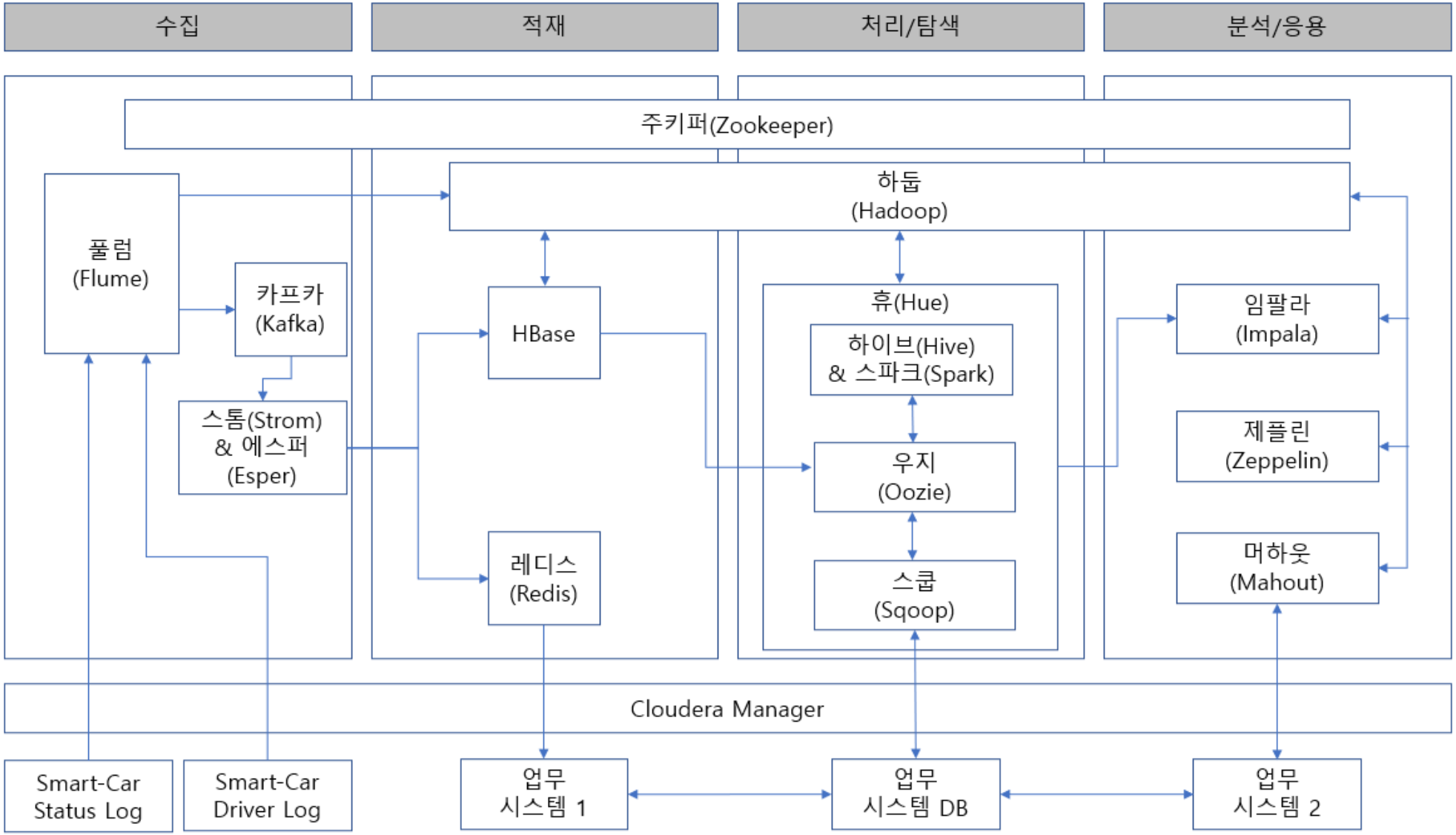
IBM이 진실성(Veracity)이라는 요소를 더해 4V를 정의했고, 이후에 시각화(Visualization) 와 가치(Value)가 추가로 정의되면서 6V까지 확장됐다.

- **크기(Volume): 방대한 양의 데이터(테라, 페타바이트 이상의 크기)**
- **다양성(Varity): 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**

빅데이터의 정의 6V


3V
2V
1V
Volume
Variety
Velocity
Veracity
Visualization
Value

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 데이터 를 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_DEEPLEARNONG

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/