데이터 사이언스와 빅데이터

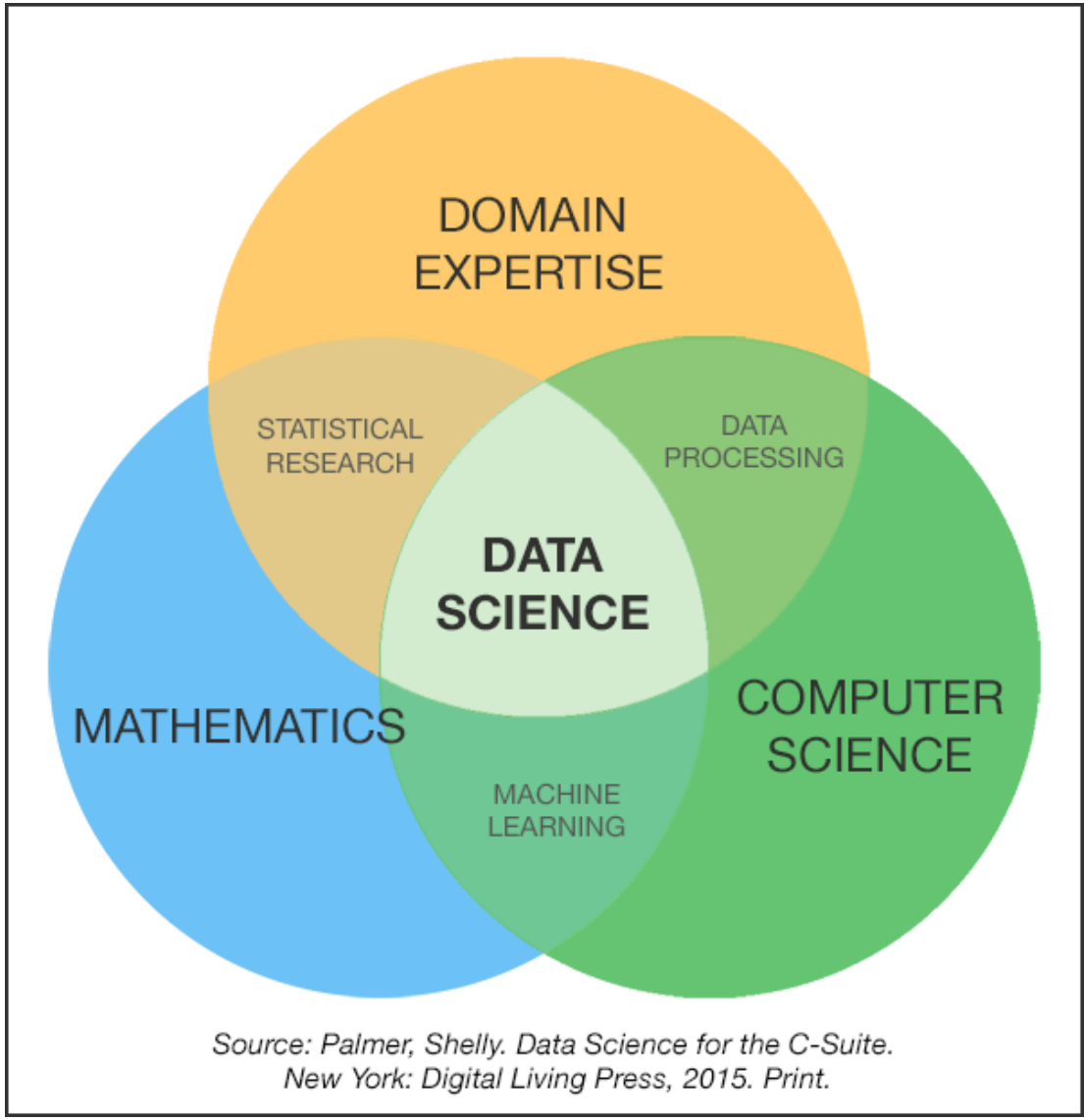
데이터 사이언스

- 데이터 사이언스는 단일 학문이 아님
  - 도메인 지식
  - 수학 및 통계
  - 컴퓨터 과학의 융합형 학문

- 데이터 사이언스 : 데이터를 과학적으로 다루는 일
  - 데이터 사이언티스트의 일

가치를 더할 수 있는 일을 찾고
데이터를 이용해서 문제를 해결하는 것

데이터로부터 머신러닝, 컴퓨터, 통계 등의
기술을 활용, 인사이트를 도출하는 일



데이터 사이언스 밴 다이어그램

데이터 사이언스
: 데이터 사이언스와 데이터

- 데이터
  - 기록 또는 자료
  - 관찰이나 측정을 통해 수집한 사실 또는 값
  - 예전 : 대부분의 데이터가 버려졌다
  - 현재 : 컴퓨팅 파워의 증가와 저장 기술의 발달
    -> 대부분의 데이터를 저장, 활용 가능

- 데이터 사이언스
  - 데이터 마이닝과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합 분야

    - 위키피디아

  - 데이터와 연관된 모든 것을 의미

    - Journal of Data Science

  - 데이터 사이언스에 필요한 역량은 프로그래밍, 수학과 통계 그리고 특정 분야에 대한 전문성이다.

    - 컨웨이(경제학자)

데이터 사이언스
빅 데이터와 스몰 데이터

- 스몰 데이터: 빅 데이터와 비교, 기존 기술을 이용해 취급할 수 있는 작은 데이터
  - 명확한 구분 기준은 없으나 한 대의 컴퓨터에서 큰 부담 없이 처리할 수 있는 데이터를 칭함
  - 대략 레코드 수로 수백만~수천만 건, 데이터 양으로 수 GB 정도의 데이터를 칭함

- 빅 데이터와 스몰 데이터 모두 그냥 데이터이며 본질적 차이는 없다
  - 대량의 데이터는 과거라면 버릴 수밖에 없었던 데이터였지만, **컴퓨팅 파워의 증대**와 **저장 장치의 비용과 효율의 증대**로 저장, 분석, 활용할 수 있게 된 것일 뿐이다

- 데이터 분석 방법은 스몰 데이터 시절부터 이미 존재

→ **결국 효율의 문제**

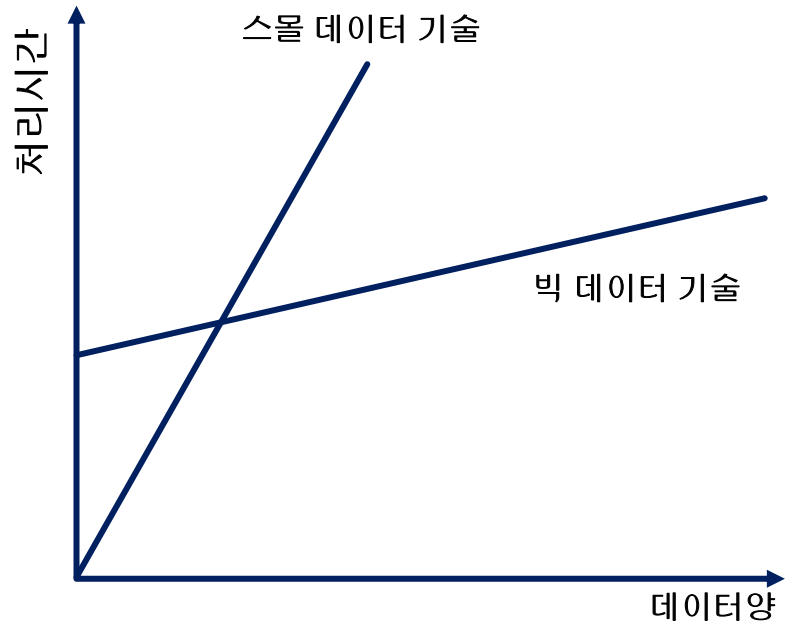
데이터 사이언스
빅 데이터와 스몰 데이터

- 이미 세상에 가득한 스몰 데이터
  - 사내에서 작성된 Excel 파일
  - 웹으로부터 수집한 CSV 파일
  - 관계형 데이터베이스에 정리된 고객 정보 파일 등

- 스몰 데이터 기술은 빅 데이터 기술만큼 중요

→ 효율적인 스몰 테이터 처리 방법을 알지 못한 채
빅 데이터 기술만 학습해서는 충분하지 않음

→ 빅 데이터 기술과 스몰 데이터 기술을
적재적소에 구사하는 것이 이상적



- 스몰 데이터 기술에서는 데이터양이 증가하면
  처리 시간이 급격히 증가
- 빅 데이터 기술의 경우 시간의 증가는 억제되지만
  데이터양이 적은 상황에서는 스몰 데이터 기술이 더 우수

데이터 사이언스
: 빅 데이터의 형태와 종류

▶ 빅데이터 구분

형태	특징	종류
정형 데이터 (Structured)	고정된 필드에 정의된 데이터	RDBMS
반정형 데이터 (Semi-Structured)	고정된 필드는 아니지만 스키마를 포함하는 데이터	XML, HTML, JSON, CSV
비정형 데이터 (Non-Structured)	고정된 필드에 저장되어 있지 않은 데이터	텍스트, 이미지, 동영상

정형 데이터

- 단순한 형태로 정리가 잘 되어있어 분석하기 쉬운 데이터 형태
- 기존 데이터 분석에 주로 사용되던 형태로 분석에 용이
- 예: 기업 또는 기관에서 주로 가지고 있는 고객 정보와 매출 등

반정형 데이터

- 데이터 속성인 메타데이터를 가지며 스토리지에 저장되는 데이터

비정형 데이터

- 복잡한 형태로 잘 정리가 안 되어 있어 분석하기 힘든 데이터 형태
- 최근에 많이 발생하고 있는 소셜 데이터와 영상, 이미지, 음성 등의 다양하고 복잡한 형태의 데이터들을 통칭.
- 정리가 안 되고 복잡하기 때문에 분석이 어렵다. 따라서 빅데이터 분석 기법들과 관련 하드웨어들이 주목

데이터 사이언스
: 기존 데이터 vs 빅 데이터

구분	기존 데이터	빅 데이터
데이터 양	테라바이트(TB) 수준	페타바이트(PB) 이상
데이터 유형	정형 데이터	소셜 미디어, 로그 파일, 스트리밍, 동영상 등 비정형 데이터 중심
프로세스 및 기술	1. 프로세스 및 기술이 단순 2. 처리/분석이 정형화 3. 원인/결과 중심	1. 처리 복잡도 높음 2. 처리에 새로운 기술이 필요 3. 잘 정의된 데이터 모델이 필요 없음 4. 상관관계 중심

- 기존 데이터와 빅 데이터의 처리 차이점
  - 빠른 의사결정이 상대적으로 덜 요구된다
  - 처리 Processing 복잡도가 높다
  - 처리할 데이터 양이 방대하다
  - 비정형 데이터의 비중이 높다
  - 처리/분석 유연성이 높다
  - 동시처리량(Throughput)이 낮다

데이터 사이언스
: 빅 데이터의 예

- 웹 검색 엔진 데이터
  - 빅데이터 시스템의 근간인 하둡은 2003년 Google File System, 2004년 MapReduce 두 논문에서 시작
  - 검색엔진을 만들면서 대용량 데이터 처리에 대한 여러 이슈를 해결하기 위한 시스템에 대한 설명
  - 검색엔진에서 사용하는 데이터들이나 검색엔진에서 만들어 내는 데이터들은 빅데이터의 전형적인 예

- 웹 페이지 데이터
  - 웹 검색엔진의 검색 대상이 되는 웹 페이지들
  - 검색엔진은 수 조개의 웹 페이지를 크롤링 해서 인덱스를 만들어 낸다.
    ex) 크롤링 된 1조 개의 웹 페이지(4KB) -> 전체 데이터는 대략 4PB
  - 웹 페이지들은 계속 업데이트가 일어 나기 때문에 검색엔진은 이를 반영하기 위해 계속 크롤링하고 저장하고 인덱스를 다시 만들어주는 작업을 한다.(분산 파일 시스템, 처리 시스템이 필요)

데이터 사이언스
: 빅 데이터의 예

- 검색어 로그와 클릭 로그 데이터
  - 검색엔진을 사용하면 만들어지는 검색어 로그 데이터
  - 검색 결과 클릭 로그 데이터
  - 구글 트렌드 서비스 (http://www.google.com.trend)
  - 독감 예상 서비스 (http://www.google.org/flutrends)

- 디바이스에서 생성되는 데이터
  - 스마트폰 : 전화 기록, 위치정보, 앱 다운로드 정보, 앱 사용 정보 등등
  - 시스코(CISCO)에 따르면 2015년 현재 모바일 디바이스 생성 데이터는 매 달 6.3XB(엑사바이트)
  - 서버, 스위치와 같은 네트워크 장비에서 발생되는 데이터
  - 각종 IoT 디바이스들의 센서들에서 측정되는 데이터
    ex) 보잉사의 제트 엔진은 30분마다 10TB 의 데이터를 만들어 낸다.
    스마트미터, 강우량이나 풍량 등의 날씨 측정 센서

데이터 사이언스
: 빅 데이터의 예

- 소셜 미디어 데이터
  - 페이스북, 트위터, 링크드인, 포스퀘어 와 같은 소셜 미디어에서 만들어 내는 데이터
  - 비지니스 분야에서 유용하게 활용할 수 있는 엄청난 크기의 데이터
  - 기업들이 마케팅에 활용하고자 데이터 마이닝에 많은 노력을 하고 있는 데이터

빅 데이터 기초 지식
: 빅 데이터의 정착

- 2000년대를 거치며 데이터 분석을 위한 환경은 크게 변화
  - 대량의 데이터를 활용하여 새로운 가치를 창출하거나 의사 결정을 위해 이용하는 일이 보편화
  - 클라우드 서비스의 보급으로 기술적 제약이 크게 낮아짐

- 2010년대 초반, 많은 기업들이 데이터 처리에 분산 시스템을 도입하기 시작
  - '빅 데이터'라는 용어가 널리 일상적으로 사용되면서 데이터를 비즈니스에 활용하려는 움직임이 활발해짐 -> 하나의 기술 분야로 정착
  - 하지만, 현재로서는 **'빅 데이터 기술이 큰 어려움 없이 안심하고 사용할 수 있다'고 말하기 어려운 상황**이며, 실제로 **'데이터를 모아서 무엇을 할 것인가?'**에 대해서도 명쾌하게 해답을 내리기 어려운 상황

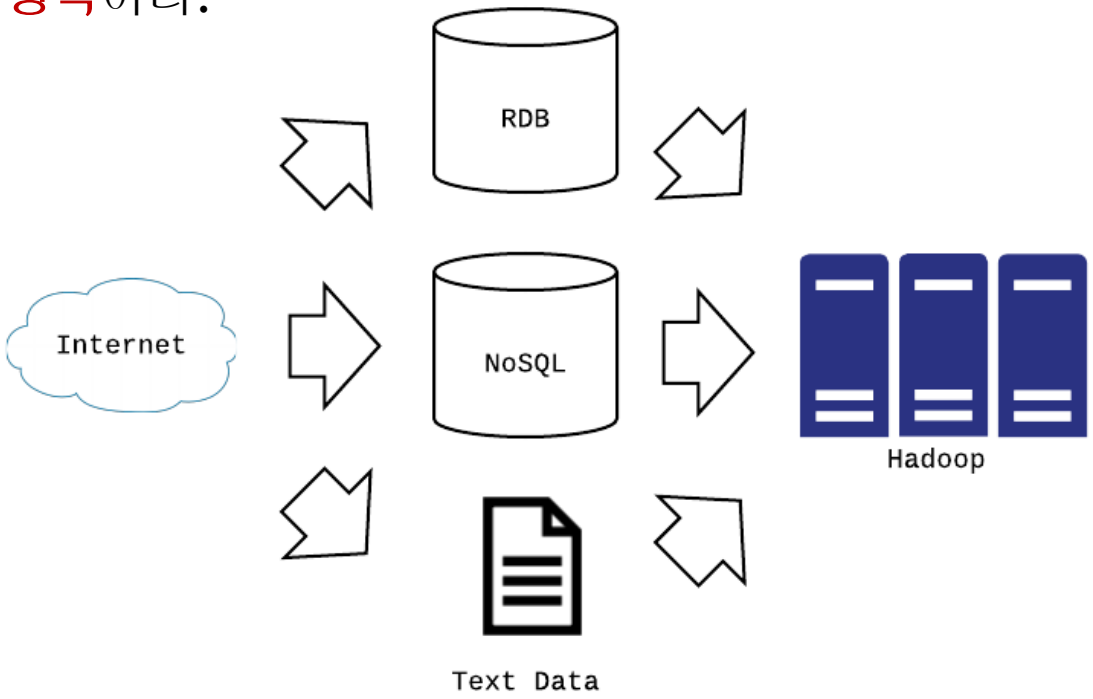
빅 데이터 기초 지식
: 빅 데이터의 정착

- 빅 데이터 취급이 어려운 이유
  - 데이터의 분석 방법을 모른다
  - 데이터 처리에 수고와 시간이 걸린다

데이터가 있어도 **가치를 창조하지 못하면 의미가 없고**,
지식이 있어도 **시간을 많이 소비한다면 할 수 있는 일은 한정적**이다.

- 빅 데이터 기술의 요구 : Hadoop과 NoSQL의 대두
  - 인터넷과 스마트폰의 보급으로 **대량**의 **비정형** 데이터들이 **빠른 속도**로 증가
  - 기존 RDB로는 취급할 수 없는 데이터들이 증가

데이터 처리를 위해 **기존과 다른 구조**가 필요



빅 데이터 기초 지식
: 빅 데이터의 정착

- Hadoop : **다수의 컴퓨터**에서 **대량의 데이터**를 처리하기 위한 시스템
  - 대량으로 발생하는 **데이터를 저장**해둘 스토리지와 **데이터를 순차적으로 처리**하기 위한 구조가 필요
  - 수백 대, 수천 대 단위이 컴퓨터를 활용해야 함 -> 이것을 관리하는 것이 Hadoop 프레임워크

  - Google에서 개발된 분산 처리 프레임워크인 MapReduce를 참고하여 제작
  - 초기 Hadoop에서 MapReduce를 동작 시키려면 데이터 처리 내용 기술을 위해 자바 언어로 프로그래밍을 해야 했음
    -> 기술적 장벽 높음, 기술 확산에 걸림돌

  - Hive 등, SQL과 같은 쿼리 언어를 Hadoop에서 실행하기 위한 소프트웨어들이 등장하면서 많은 사람들이 Hadoop을 이용한 분산 시스템을 활용할 수 있게 됨

Hadoop 중요 역사

시기	주요 이벤트
2004년 12월	구글에서 MapReduce 논문 발표
2007년 9월	Hadoop 최초 버전(0.14.1) 배포
2009년 5월	Hive 최초 버전(0.3.0) 배포
2011년 12월	Hadoop 1.0.0 배포

빅 데이터 기초 지식
: 빅 데이터의 정착

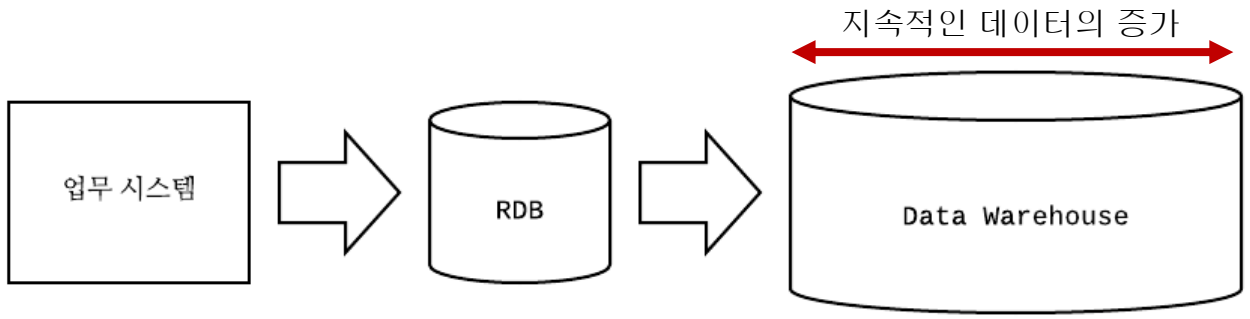
- NoSQL : 전통적인 **RDB의 제약을 제거**하는 것을 목표로 한 데이터베이스의 총칭
  - 키 밸류 스토어 : Key-Value Store / KVS - 다수의 키와 값을 관련 지어 저장
  - 도큐먼트 스토어 : Document Store - 복잡한 데이터 구조를 저장
  - 와이드 컬럼 스토어 : Wide-Column Store - 여러 키를 사용하여 높은 확장성을 제공
- NoSQL 데이터베이스의 특징
  - RDB에 비해 **고속의 읽기/쓰기**가 가능
  - **분산 처리**에 강점
- Hadoop과 NoSQL의 결합
  - NoSQL에 기록하고 Hadoop으로 분산처리하기 흐름이 확산
  - 기존의 기술로는 불가능하거나 고가의 하드웨어가 필요한 경우에도 **현실적인 비용으로 데이터를 처리**할 수 있게 됨

NoSQL 데이터베이스 중요 역사

시기	주요 이벤트	종류
2009년 8월	MongoDB 1.0 배포	도큐먼트 스토어
2010년 7월	CouchDB 1.0 배포	도큐먼트 스토어
2011년 9월	Riak 1.0 배포	키-밸류 스토어
2011년 10월	Cassandra 1.0 배포	와이드 컬럼 스토어
2011년 12월	Redis 1.0 배포	키-밸류 스토어

빅 데이터 기초 지식
: 빅 데이터의 정착

- 빅 데이터가 확산되기 이전에도 엔터프라이즈 데이터 웨어 하우스(Enterprise Data Warehouse/EDW) 제품들이 데이터 분석 기반으로 활용
  - 전국 각지에서 보내진 점포의 매출, 고객 정보 등이 오랜 기간에 축적
  - 축적된 데이터를 분석함으로써 업무 개선과 경영 판단의 자료로 활용
- 전통적 데이터 웨어하우스도 대량의 데이터를 처리할 수 있으며 여러 방면에서 오히려 Hadoop 보다 우수
  - 하지만, 많은 데이터 웨어하우스 제품들이 안정적인 성능을 실현하기 위해 하드웨어와 소프트웨어가 통합된 통합 장비(Appliance)로 제공, 데이터 용량을 늘리려면 하드웨어를 교체하는 등 추후 확장이 쉽지 않음

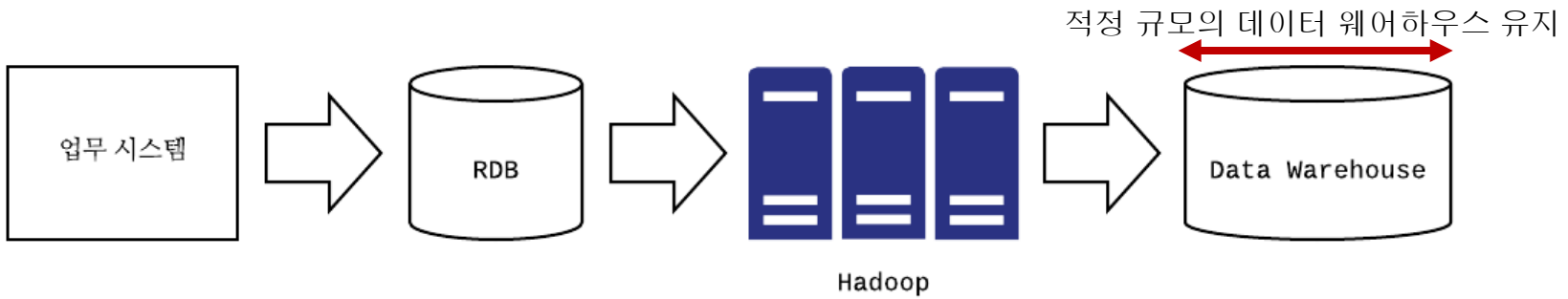


기존의 데이터 웨어하우스 활용

빅 데이터 기초 지식
: 빅 데이터의 정착

- 분산 시스템의 장점은 Scale Out이 가능하다는 점

- Data Warehouse의 부하를 줄이기 위해 Hadoop을 이용, 협업
  - 기하급수적으로 늘어나는 데이터의 처리는 Hadoop에 일임
  - 비교적 적은 데이터, 정제된 데이터, 중요한 데이터만 데이터 웨어하우스에서 관리



데이터 웨어하우스와 Hadoop의 협력 모델

빅 데이터 기초 지식
: 빅 데이터의 정착 - 클라우드 서비스의 보급과 빅데이터

- 빅 데이터 시스템의 특징 : 여러 컴퓨터에 분산 처리한다
  - 현실적 한계: 분산 시스템을 위한 하드웨어를 준비하고 소프트웨어를 설치, 설정, 관리한다는 것은 쉬운 일이 아님
- 최근에는 클라우드 서비스 벤더(아마존, 구글, 마이크로소프트)들이 데이터처리를 위한 클라우드 서비스를 제공
  - 시간 단위로 필요한 자원을 확보할 수 있음
  - 작은 프로젝트 단위에서도 손쉽게 데이터 웨어하우스를 구축, 데이터 분석 기반을 마련하는 것이 가능
    -> 빅 데이터 환경의 대중화,

amazon web services

Microsoft Azure

Google Cloud

데이터 처리를 위한 클라우드 서비스

시기	주요 이벤트	종류
2009년 4월	Amazon Elastic MapReduce	클라우드를 위한 Hadoop
2010년 5월	구글 BigQuery	데이터 웨어하우스
2012년 10월	Azure HDInsight	클라우드를 위한 Hadoop
2012년 10월	Amazon RedShift	데이터 웨어하우스

빅 데이터 기초 지식

: 빅 데이터의 정착

- 오픈 소스 생태계와 빅 데이터 글로벌 빅 3 업체의 선전

- 오픈 소스 소프트웨어 생태계
  - 빅 데이터 기술은 거대한 오픈 소스 소프트웨어 생태계
  - 대용량 데이터 분산 저장/처리를 도와주는 오픈 소스 프로젝트 하둡이 빅 데이터 시장을 받쳐주고 있음
  - 하둡을 기반으로 비즈니스 사용 측면에서 요구되는 여러 오픈소스들이 추가되어 하나의 생태계를 이룸
    -> 하둡 에코 시스템(Hadoop Ecosystem)
  - 하둡 기술을 주도하는 기업, 조직이 전 세계 빅 데이터 시장을 주도하는 추세

- 빅 데이터 글로벌 빅 3
  - 클라우데라(Cloudera), 호튼웍스(HortonWorks), 맵알(MapR)의 과감한 투자와 마케팅 -> 빠르게 상업화
  - 이들 3대 업체들이 하둡 기술을 중심으로 빅데이터 시스템 소프트웨어 스택을 개발, 공개
    -> 빅 데이터 생태계에 절대적인 영향력을 행사
  - 국내 빅 데이터 기술 수준은 아직 글로벌 빅3가 주도하는 하둡 기반 오픈 소스 프로젝트들을 단순 활용하는 수준

빅 데이터 기술

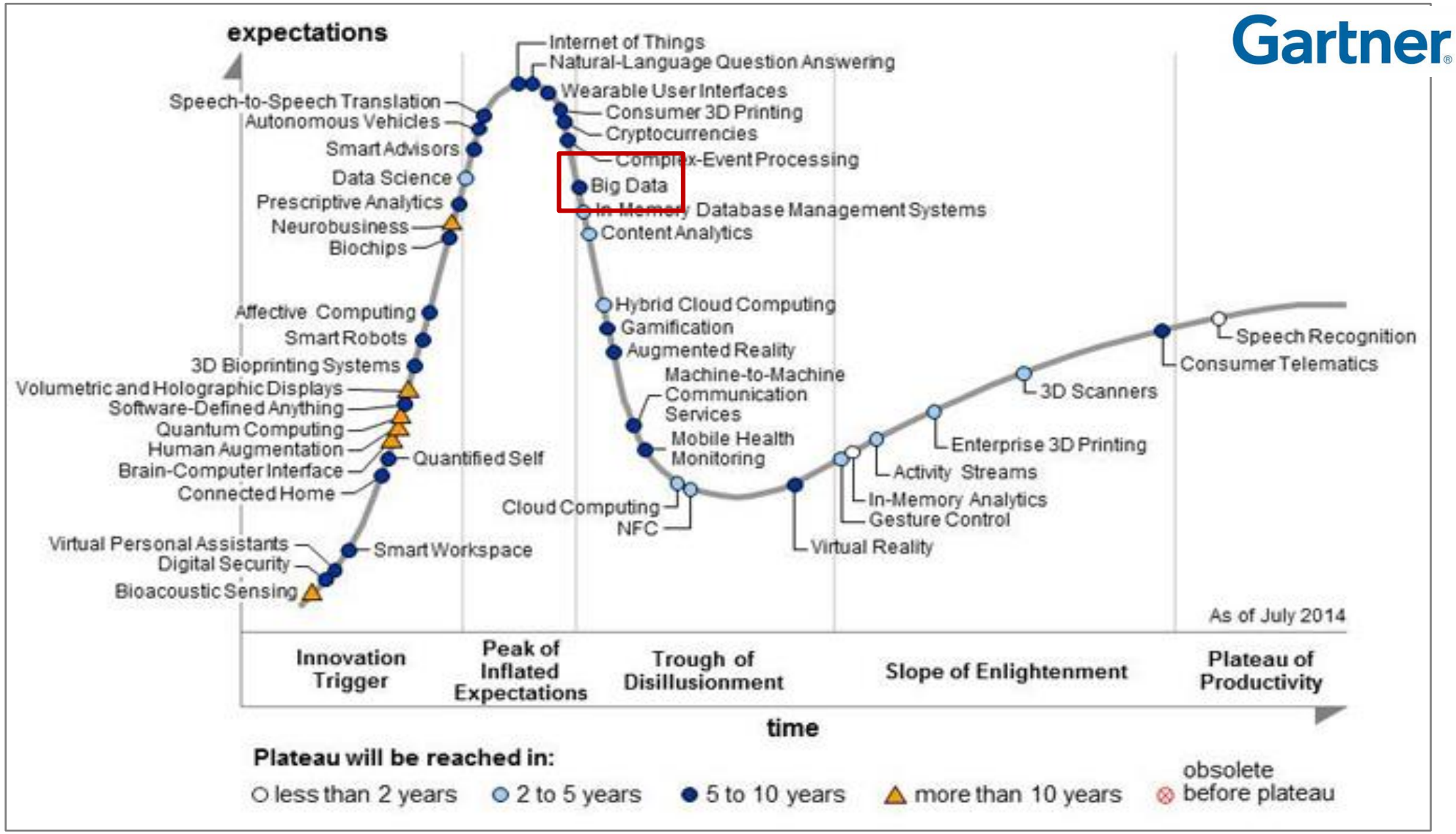
- 빅 데이터 기술의 변화 (아키텍처)
  - 인프라스트럭처(Infrastructure)
    - 서버 (x86, 리눅스 서버)
    - 네트워크 (대규모 빅 데이터 서버 및 스토리지 지원을 위한 대용량 네트워크)
    - 스토리지 (대규모 데이터 저장을 위한 내외부 스토리지 장비)
  - 소프트웨어 플랫폼
    - 빅 데이터 전방위 기술을 포괄하는 스택(순수 오픈 소스 스택, 기업 배포판 스택)
    - 하둡을 기반으로 한 오픈 소스 생태계
    - 빅 데이터 수집/적재/처리/분석 등의 지원 솔루션)
    - 빅 데이터 시스템 관리 및 모니터링 툴 제공
  - 서비스
    - 빅 데이터 컨설팅 및 구축
    - 빅 데이터 전문 운영 및 유지 보수
    - 데이터 분석 서비스
    - 교육 및 운영 인력 양성

빅 데이터 기술

- 빅 데이터 기술의 변화 (기간별)
  - ~ 2009년
    - 수집/적재를 위한 스토리지 안프라
    - 낮은 비용의 스토리지 구축을 위한 솔루션으로 인식
    - 저비용 x86 서버 하드웨어를 활용한 대용량 데이터 저장소
  - ~2012년
    - 분산 컴퓨팅 기반 분석 기술이 시작
    - RDB로 처리하기 어려운 대용량 데이터 연산, 추출 및 집계 기술
  - ~2016년
    - 빅 데이터의 가치와 활용 측면에서 분석 기술이 적극 활용되어 고급 분석 기술로 확대 발전
    - 고급 분석(머신러닝, 텍스트 마이닝)을 통한 통찰력과 예측들을 데이터 서비스로 활용
    - 조직의 의사결정 도구 등 IT, 통신, 공공, 금융 이외에도 의료, 방성, 제조, 통신 등 사업 전방위에 활용되기 시작

빅 데이터 기술

: 전망 (by Gartner, 2014)



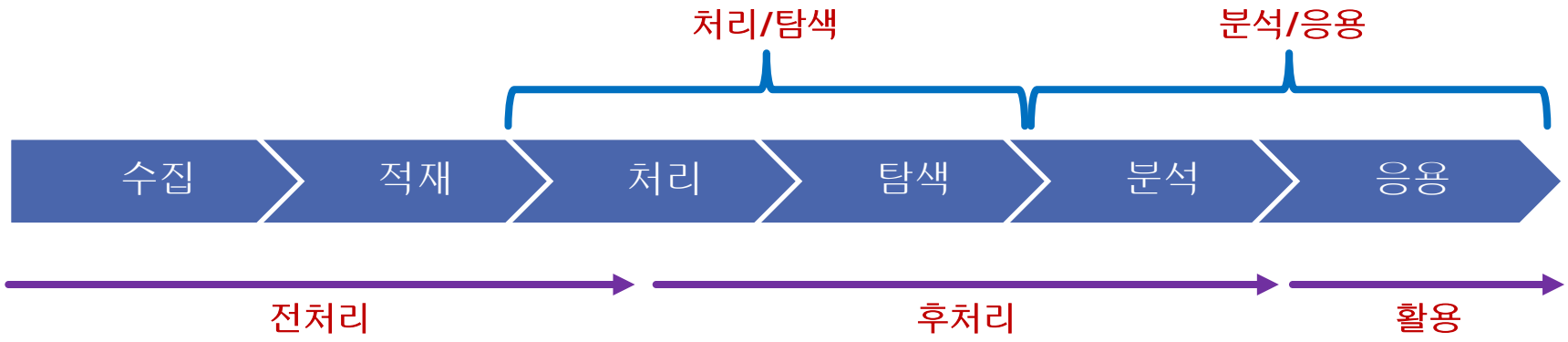
2014 Hype Cycle for Emerging Technologies

빅 데이터 기술

- 빅 데이터 기술이 기존의 데이터 기술과 다른 점은 "다수의 **분산 시스템을 조합**하여 **확장성**이 뛰어난 **데이터 처리 구조**를 만든다"는 점

- 데이터 파이프라인(Data Pipeline): 차례대로 전달해나가는 데이터로 구성된 시스템을 칭함
  - **어디에서 데이터를 수집**하여 **무엇을 실현하고 싶은지**에 따라 다양하게 변화한다
  - 처음에는 간단한 구성으로 끝나지만, 하고 싶은 일이 증가함에 따라 시스템은 점차 복잡해지고 시스템들을 **어떻게 조합시킬지**가 문제가 된다

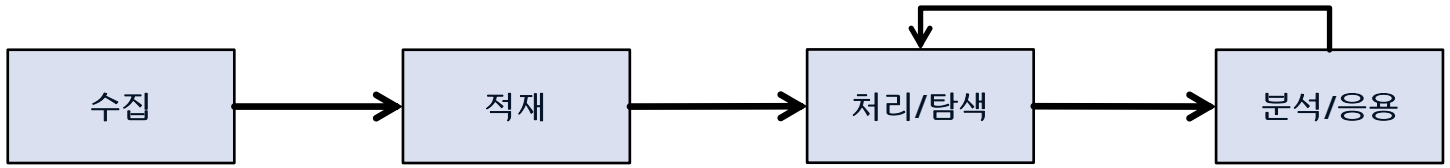
- 빅 데이터 처리 단계



빅 데이터 기술

- 빅 데이터 아키텍처 6개 계층(Layer)

빅 데이터 기술



- 6단계는 보통 4개의 단계로 축약하여 진행하며 처리/탐색과 분석/응용은 반복 진행하여 품질과 분석 수준을 향상시켜 나가게 된다
  - 기술들은 보통 하둡 생태계의 오픈 소스들로 구성하는 것이 일반적

- 수집 기술
  - 서비스/조직 내외부의 다양한 시스템으로부터 원천 데이터를 효과적으로 수집하는 기술
  - 빠르고 다양한 데이터를 효과적으로 수집해야 하기 때문에 선형 처리 그리고 분산 처리가 가능해야 한다
  - 외부 데이터(SNS, 블로그, 포털 등) 수집시에는 크롤링, NLP 등 비정형 처리를 위한 기술이 선택적으로 적용될 수 있다
  - 실시간 수집의 경우에는 CEP, ESP 기술 등 이벤트를 감지해 콜백 처리를 한다
  - 수집된 데이터는 정제, 변환, 필터링 작업을 거쳐 분산 스토리지에 적재하게 된다
  - 6V 중, Volume, Variety, Velocity를 효과적으로 처리하는 기능에 집중한다
  - 관련 소프트웨어 : Flume, Fluentd, Scribe, Logstash, Chuckwa 등이 있으며 실시간 처리를 위해 Storm, Esper 등도 고려된다

빅 데이터 기술

- 적재 기술
  - 수집된 데이터를 분산 저장소(Distributed Storage)에 영구 또는 임시로 적재하는 기술
  - 빅 데이터의 분산 저장소는 4가지 정도로 구분할 수 있다
    1. HDFS(Hadoop Distributed File System) : 대용량 파일 영구 저장을 목적으로 한다
    2. NoSQL(Hbase, MongoDB, Cassandra 등) : 대규모 메시징 데이터를 영구 저장을 목적으로 한다
    3. Inmemory Caching(Redis, Memcached, Infinispan 등) : 대규모 메시지 처리 결과를 고속으로 저장하기 위해 사용
    4. MoM(Kafka, RabbitMQ, ActiveMQ 등) : 대규모 메시징 데이터를 임시 저장하기 위한 목적으로 사용

- 수집된 데이터의 성격에 따라 적재 저장소를 선택해서 사용해야 한다
- 적재될 때는 추가적인 전처리 작업이 선행될 수도 있다 (비정형 데이터 -> 정형 데이터, 비식별화 처리
- 비즈니스 요구에 따라 HDFS에 적재 후, 후처리로 이루어질 수도 있다 (MapReduce)
- 6V 관점에서는 Volume, Velocity, Veracity를 효과적으로 처리하는 데 집중한다

빅 데이터 기술

- 처리/탐색 기술
  - 대용량 저장소에 적재된 데이터를 분석에 활용하기 위해 정형화/정규화 하는 기술
  - 데이터를 통해 가치를 찾아내기 위해서는 데이터를 이해해야 하기 때문에 적재된 데이터를 관찰하고 탐색적 분석을 수행한다.
  - 탐색적 분석에는 SQL on Hadoop이 주로 사용된다
  - 대화용 Ad-Hoc 쿼리로 데이터를 선택, 변환, 통합, 축소 등의 작업을 수행한다.
  - 정기적으로 수행해야 하는 처리/탐색 작업은 Workflow로 프로세스화하고 자동화한다.
  - 처리/탐색 작업이 끝난 데이터셋들은 데이터 웨어하우스(Data Warehouse)로 측정 가능한 구조로 만들어져 분석을 편리하게 할 수 있게 한다
  - 처리/탐색 기술로는 Hue, Hive, Spark SQL 등이 있고, 후처리 작업을 자동화하기 위한 Workflow 작업에는 Oozie를 사용한다

빅 데이터 기술

- 분석/응용 기술
  - 대규모 데이터들로부터 새로운 패턴을 찾고 패턴에 대한 해석을 통해 통찰력(Insight)을 확보하기 위한 기술
  - 활동 영역에 따라 통계, 데이터 마이닝, 텍스트 마이닝 등 다양하게 분류될 수 있다
  - 그 이전에도 데이터 분석 기술과 도구는 존재했지만 데이터 크기, 생성 속도, 다양성 등에 대한 한계점을 낮은 비용의 대규모 분산 환경 구축으로 극복할 수 있게 되었다
  - 머신러닝 기술을 활용한 Clustering, Classification, Regression, Recommendation 등의 영역까지 확장 가능
  - 최근에는 대규모 배치 분석이 인메모리 기반의 실시간에 가까운 분석이 가능해지고 있다
  - 6V의 거의 모든 요소가 해당된다
  - 분석/응용 기술로 Impala, Zepplin, Mahout 등이 있으며 분석된 데이터는 Sqoop을 활용해 RDBMS로 Export도 가능하다

빅 데이터 기술

- 분산 데이터의 처리 : 분산 스토리지에 저장된 데이터를 처리하기 위해서는 '분산 데이터 처리 (Distributed Data Processing)' 프레임워크가 필요
  - 이 과정에서 데이터의 양과 처리 내용에 따라 많은 컴퓨터 자원이 필요
  - 분산 데이터 처리의 주 역할은 **나중에 분석하기 쉽도록** 데이터를 **가공**해서 그 결과를 **외부 데이터베이스에 저장**하는 것

- SQL로 분산 데이터 집계 : 대다수 사람들은 데이터 집계에 있어 SQL을 사용하는 것에 익숙

  방법 1. 쿼리 엔진(Query Engine) 도입 : 예) Hive

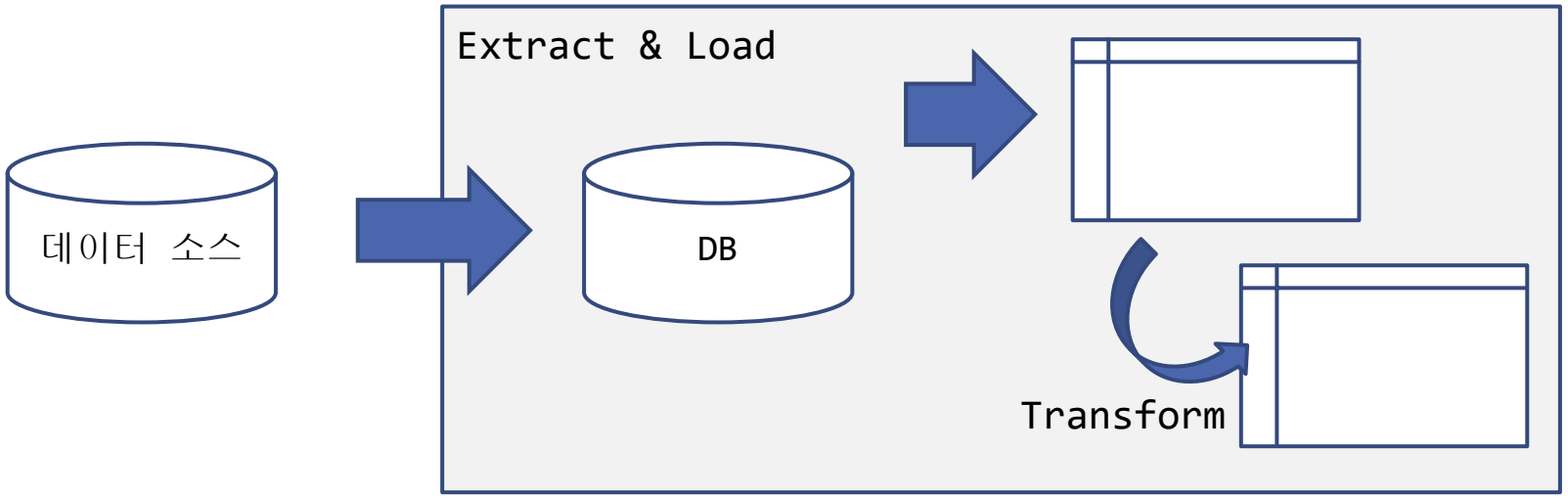
  방법 2. 외부의 데이터 웨어하우스 제품을 이용
  -> 이를 위해 분산 스토리지에서 추출한 데이터를 데이터 웨어하우스에 적합한 형식으로 변환
  : ETL(Extract-Transform-Load) 프로세스

  -> 경우에 따라서는 데이터를 읽어들인 후 가공하기도 하는데, 이런 프로세스를
  ELT(Extract-Load-Transform) 프로세스로 구분하기도 한다

빅 데이터 기술

- ETL(Extract-Transform-Load)

- ELT(Extract-Load-Transform)



빅 데이터 기술
: 데이터 웨어하우스와 데이터 마트

- 데이터 웨어하우스 : 일반적인 RDB와는 달리 '대량의 데이터를 장기 보존하는 것'에 최적화
  - 정리된 데이터를 한번에 전송하는 것은 뛰어나지만, 소량의 데이터를 자주 쓰고 읽는 데는 적합하지 않음
  - 예) 업무 시스템에서 꺼낸 데이터를 하루가 끝날 때 정리하여 쓰고, 이것을 야간 시간대에 집계하여 보고서를 작성

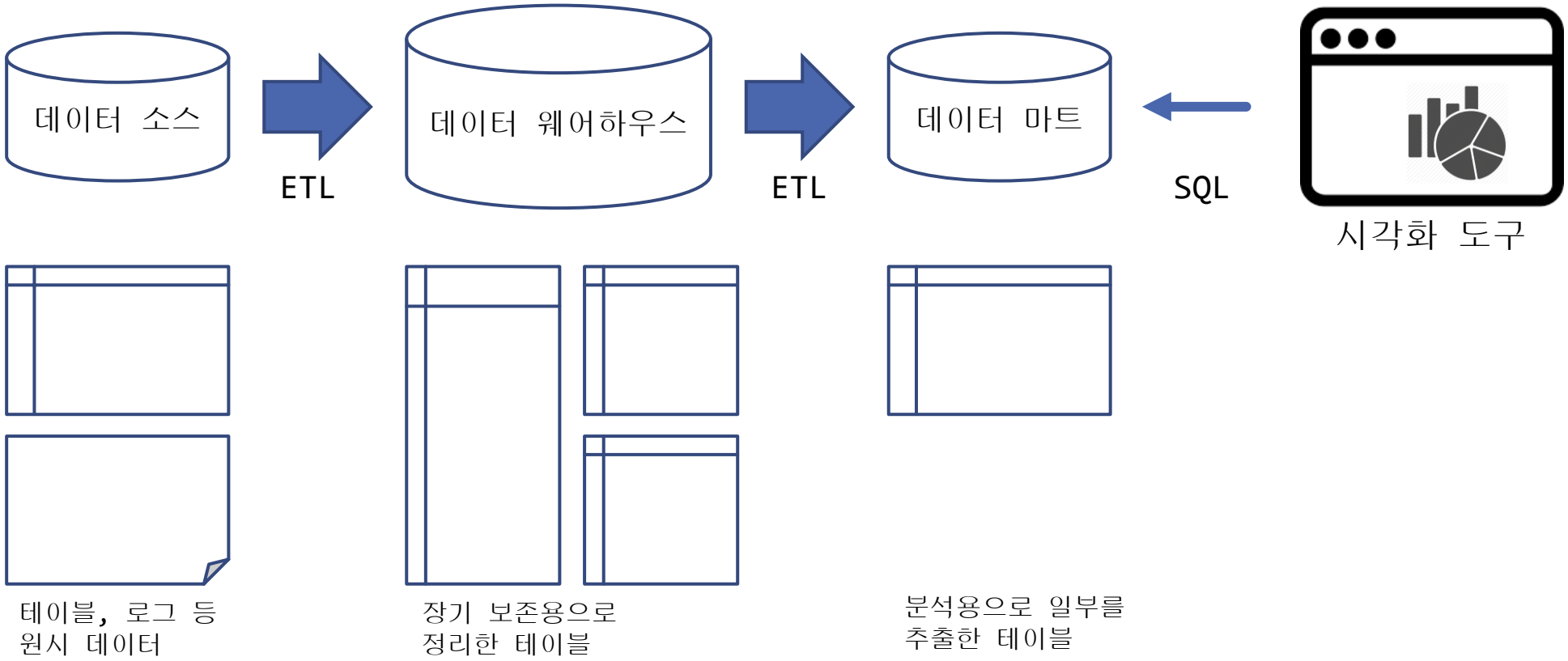
- 데이터 웨어하우스의 측면에서 봤을 때
  - 데이터 소스(Data Source) : 업무 시스템을 위한 RDB, 로그 등
  - ETL 프로세스 : 데이터 소스에 보존된 원시 데이터(Raw Data)를 추출하고 필요에 따라 가공한 후 데이터 웨어하우스에 저장하기까지의 흐름

  - 데이터 웨어하우스는 업무에 있어 중요한 데이터 처리에 사용되기 때문에 아무때나 함부로 사용해서는 곤란 (시스템 과부하 초래)

    -> 데이터 분석 등의 목적에 사용되는 경우 데이터 웨어하우스에서 필요한 데이터만을 추출하여 데이터 마트(Data Mart)를 구축

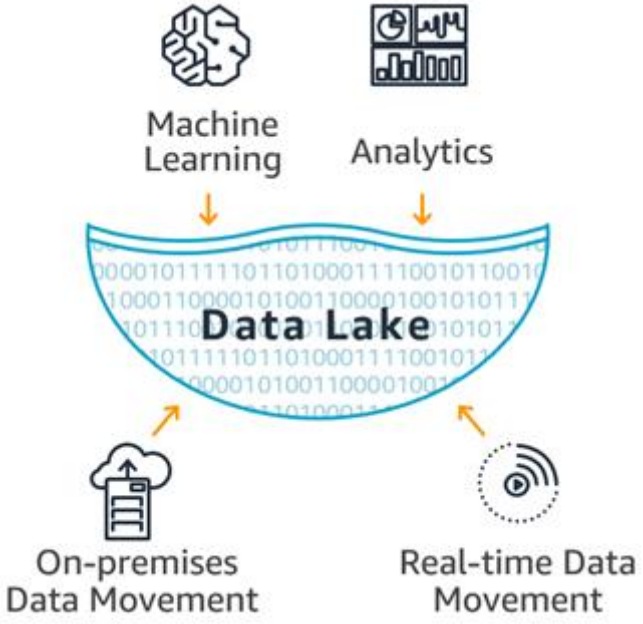
빅 데이터 기술
: 데이터 웨어하우스와 데이터 마트

- 데이터 웨어하우스를 중심으로 하는 데이터 파이프라인
  - 데이터 웨어하우스와 데이터 마트 모두 SQL로 데이터를 집계한다
    - 먼저 테이블 설계를 제대로 정한 후에 데이터를 투입한다



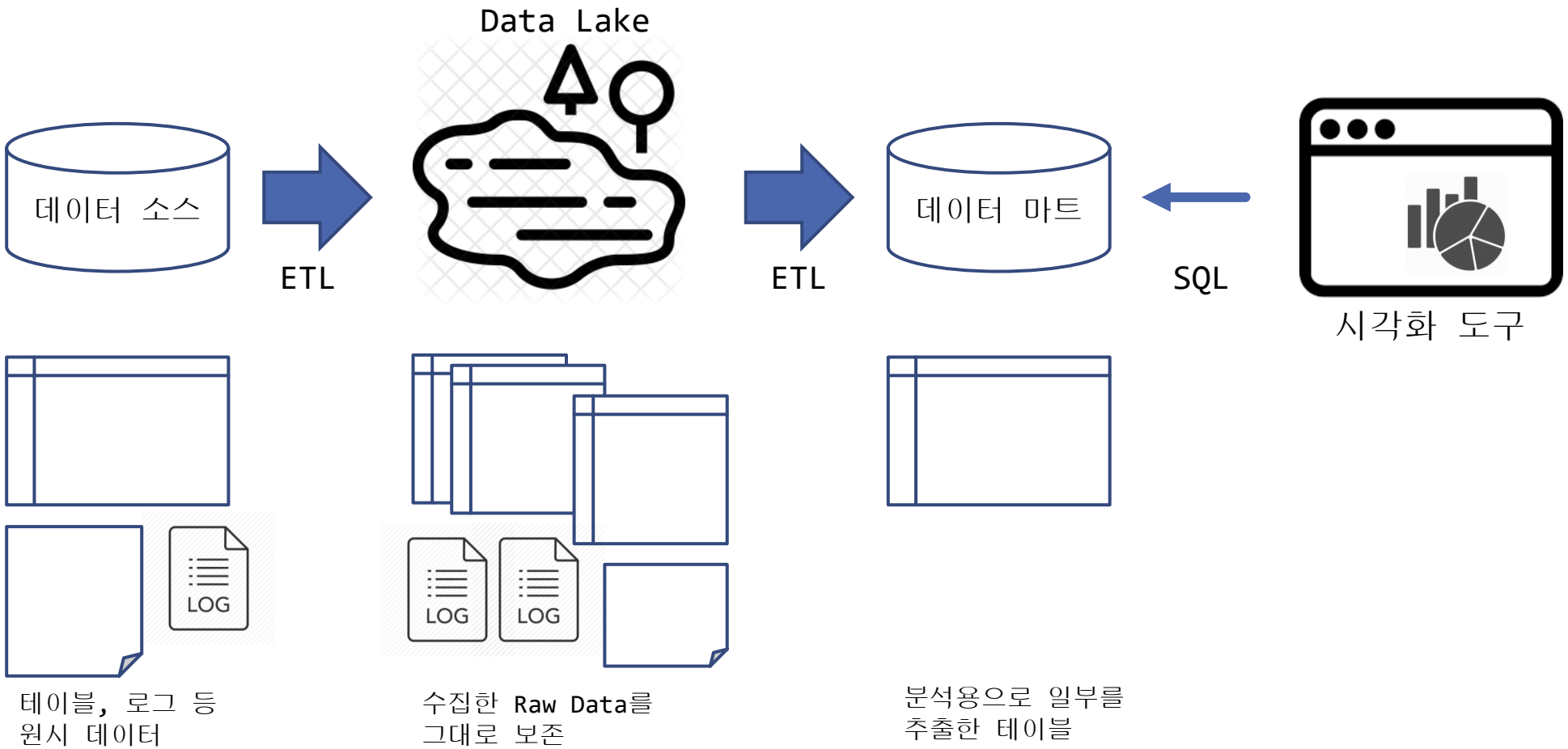
빅 데이터 기술
: 데이터 레이크

- 빅 데이터 시대가 되면서 **ETL** 프로세스 자체가 복잡해짐
  - 모든 데이터가 데이터 웨어하우스를 가정해서 만들어지지 않으며 외부에서 수집된 텍스트와 바이너리 데이터들은 그대로 데이터 웨어하우스에 집어넣을 수 없는 경우도 많다
  - 우선 데이터가 있고, **나중에 테이블을 설계**하는 것이 빅 데이터
  - 모든 **데이터를 원래의 형태대로 축적**해두고 나중에 그것을 필요에 따라 가공하는 구조가 필요
    - 데이터 레이크(Data Lake) : 여러 곳에서 데이터가 흘러 들어오는 '데이터를 축적하는 호수'에 비유
    - 임의의 데이터를 저장할 수 있는 분산 스토리지가 데이터 레이크로 이용
    - 데이터 형식은 자유지만, 대부분 범용적인 텍스트 형식인 CSV, JSON 등의 포맷이 활용된다



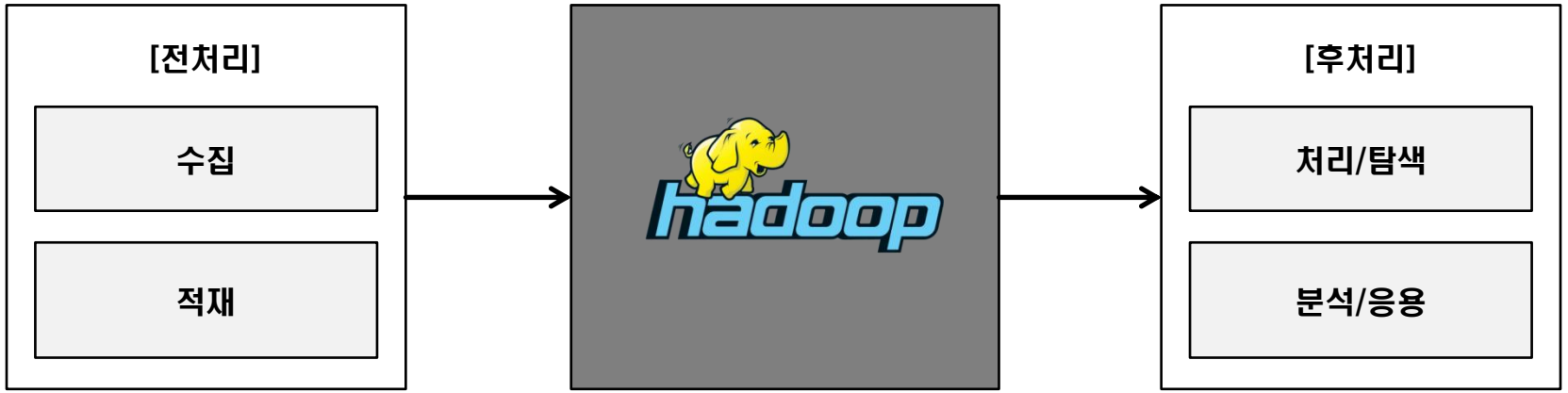
빅 데이터 기술
: 데이터 레이크

- 데이터 레이크를 중심으로 하는 데이터 파이프라인

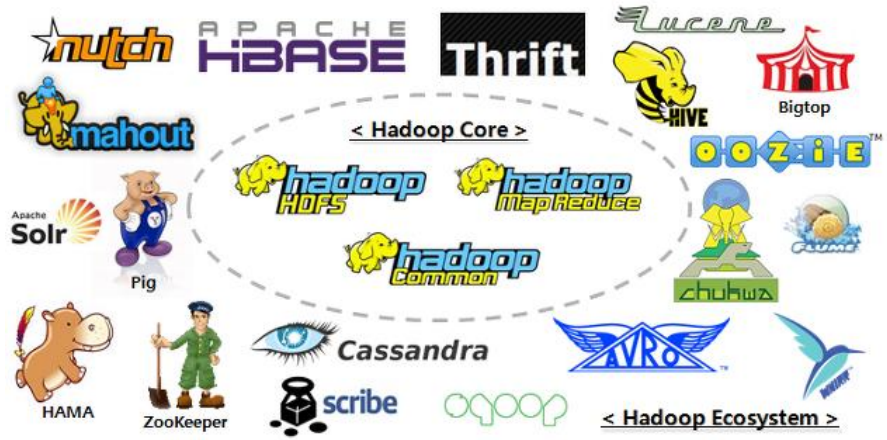


빅 데이터 기술

- 하둡을 중심으로 한 빅 데이터 시스템 아키텍처 레이어



- 하둡(Hadoop)을 중심으로 **전처리**, **후처리** 2개 영역으로 구분하는 경우도 있음
- 많인 빅 데이터 시스템의 소프트웨어 아키텍처는
  수집, **적재**, **처리/탐색**, **분석/응용** 등 4개 레이어로 구축
- 각 레이어 담당 소프트웨어들은 전문 영역이 존재하지만,
  하둡 에코 시스템에 대한 기본적인 이해와 기술이 공통적으로
  뒷받침되어야 한다

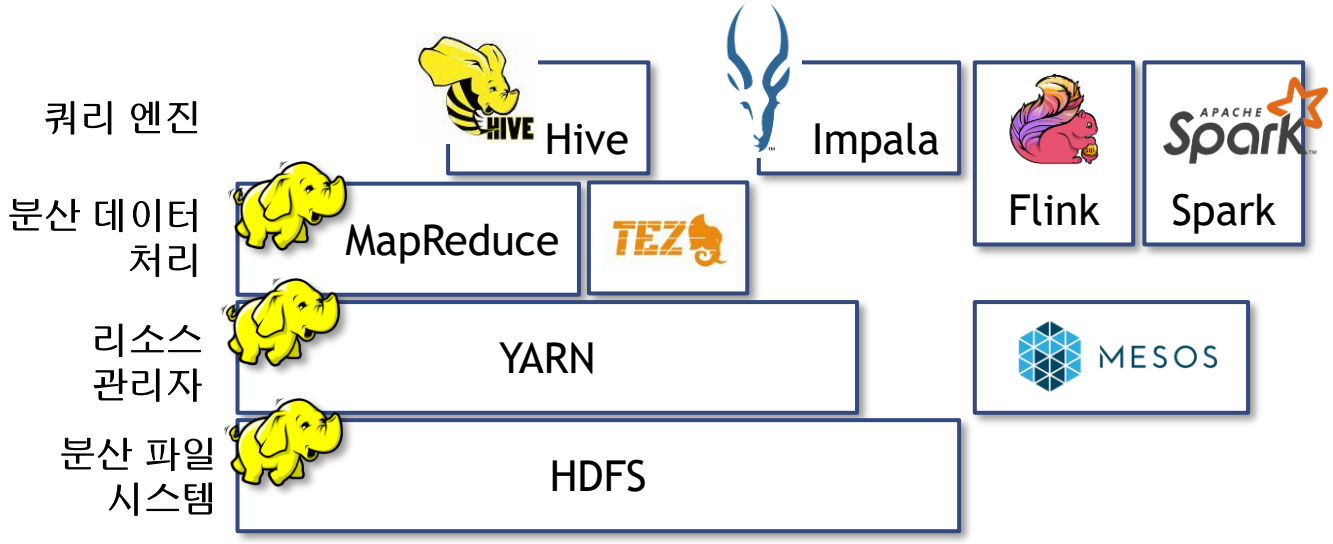


빅 데이터 기술
: Hadoop Ecosystem

- Hadoop은 단일 소프트웨어라기보다 분산 시스템을 구성하는 **다수의 소프트웨어**로 이루어진 집합체
  - 그 외에 하둡을 기반으로 하거나 하둡을 지원하는 **분산 환경 소프트웨어들이 상호 협력**하며 생태계(Ecosystem)을 이룬다

시기	이벤트
2003년	Nutch 프로젝트 발족
2004년	Google MapReduce 논문
2006년	Apache Hadoop 프로젝트 발족
2011년	Apache Hadoop 1.0.0 배포
2013년	Apache Hadoop 2.2.0 배포
2014년	Apache Spark 1.0.0 배포
2016년	Apache Flink 1.0.0 배포
2016년	Apache Mesos 1.0.0 배포

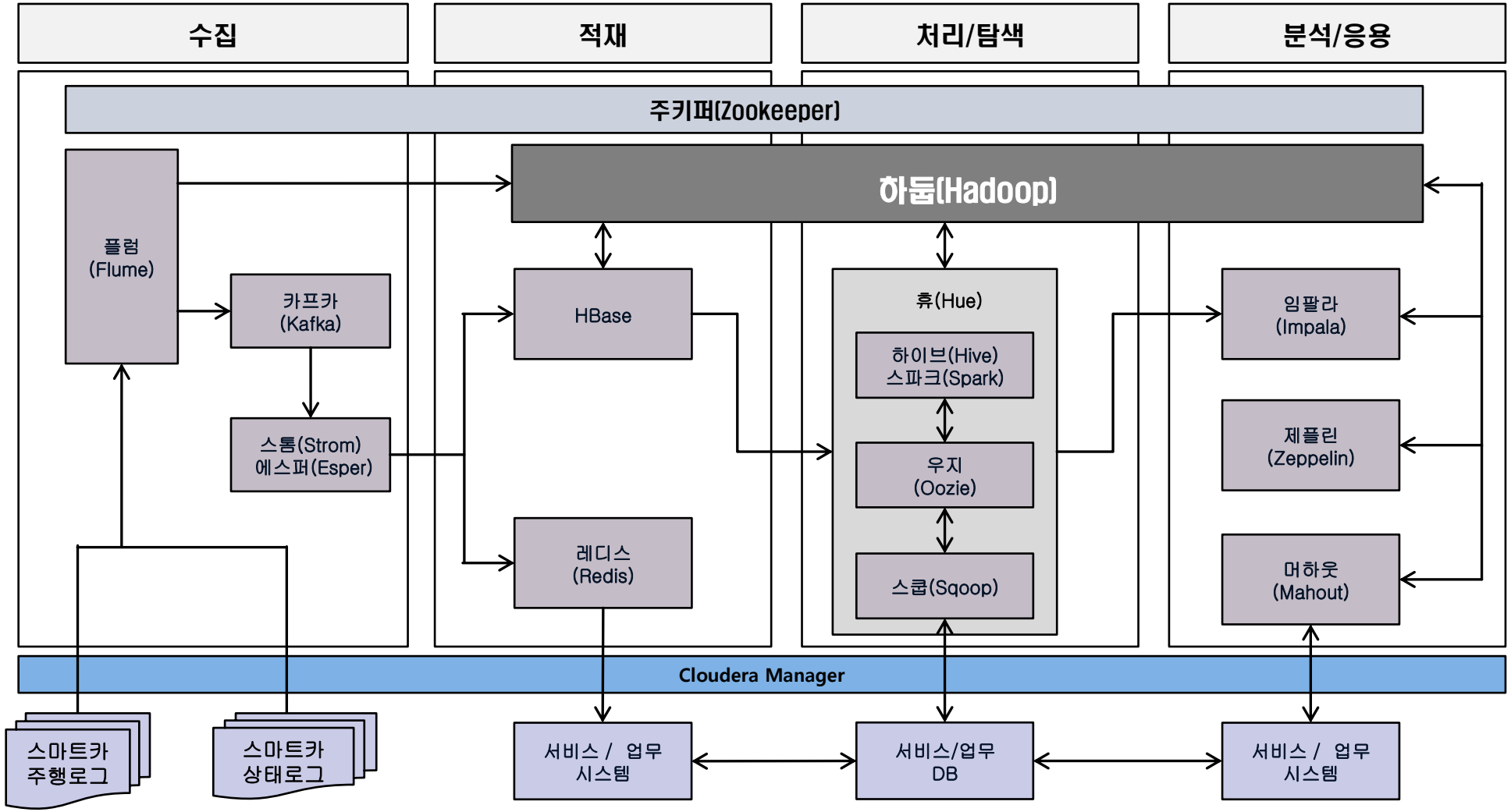
하둡과 주변 프로젝트의 역사



빅데이터 관련 Apache 프로젝트

빅 데이터 기술
: Hadoop Ecosystem

- Hadoop Ecosystem을 활용한 아키텍처 구성 사례



빅 데이터와 데이터 파이프라인

- 빅 데이터가 어려운 이유 중 한 가지는 **정해진 답이 없으며** 소프트웨어 스택 구성시에도 **여러 가지 선택지**가 있다는 것
  - 달성하고자 하는 **목표**가 같으면 그 다음은 절차상의 문제일 뿐이다

- 기본적으로 파악해야 할 두 가지 사항
  - 저장할 수 있는 **용량**에 제한이 없을 것
  - 데이터를 **효율적으로 추출할 수단**이 있을 것

- 새로운 도구와 서비스가 계속 개발되고 있지만, **데이터 파이프라인 전체의 기본적인 흐름은 변하지 않는다**는 점을 항상 염두에 두자

중요한 것은 **데이터의 흐름을 만드는 것**이며
그 과정에서 **기술은 교환**될 수 있다