빅데이터 이해하기

빅데이터의 정의(6V)

3V + 2V = 1V

3V : Volume(크기) Variety(다양성) Velocity(속도)

2V : Veracity(진실성) Visualization(시각화)

1V: Value(가치)

빅데이터 목적

빅데이터 인사이트 - 현상 이해, 현상 발견, 현상 예측

이해 인사이트

- 시계열별 회원 가입 추이
- 고객별 서비스 평균 이용 시간
- 서비스 유입 또는 이용 경로
- 신규 상품 및 서비스 관심도
- 상품 및 서비스 휴면/해지 율

발견 인사이트

- 고객 증가 감소 원인
- 매출 증가 감소 원인

예측 인사이트

- 상품 가입 이탈 고객은? 등

빅데이터 활용

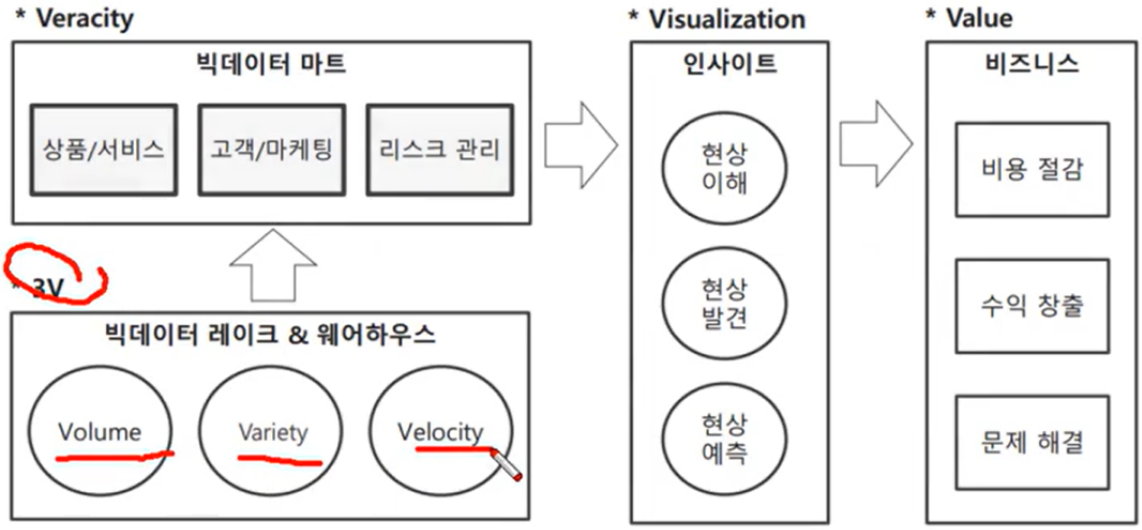


그림 1.7 빅데이터 활용 방안

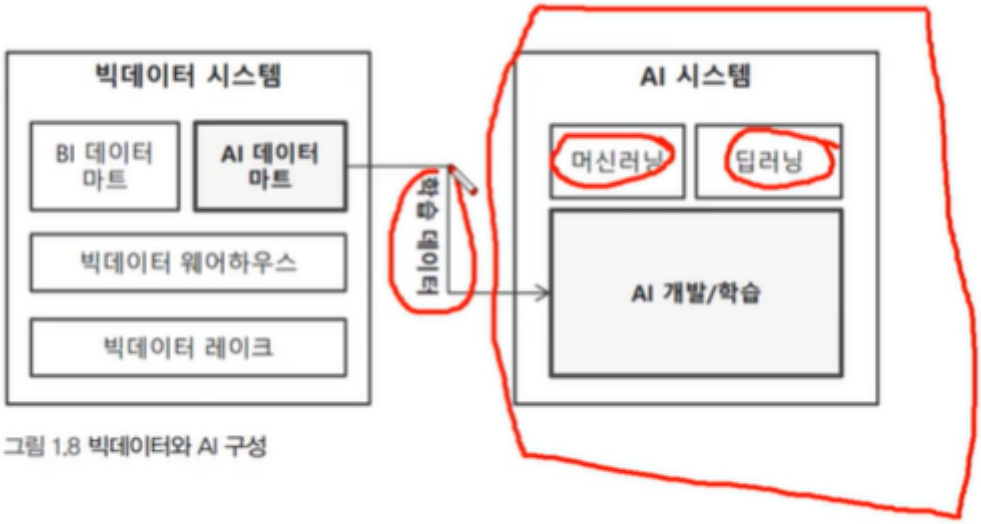


그림 1.8 빅데이터와 AI 구성

빅데이터의 오해

RDBMS와 BigData는 상호보완적

빅데이터 프로젝트

플랫폼 구축형 프로젝트

그림 1.9 빅데이터 프로젝트 조직 1 – 플랫폼 구축형

- 전형적인 빅데이터 SI 구축형 사업
- 빅데이터 하드웨어와 소프트웨어 설치 및 구성
- 수집 → 적재 → 처리 → 탐색 → 분석 기능 구현

빅데이터 분석 프로젝트

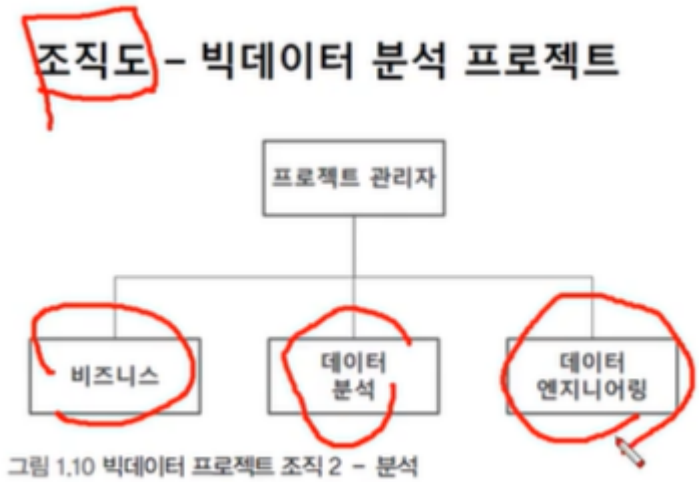


그림 1.10 빅데이터 프로젝트 조직 2 – 분석

- 플랫폼 구축 완료 후 수행
- 빅데이터 탐색으로 데이터 이해 높아질 때 시작
- 조직의 가치사슬 중, 대규모 분석이 필요한 시점에 추진
- 분석 주제 영역 → 마케팅 고객, 상품 서비스, 리스크 관리

빅데이터 운영 프로젝트

그림 1.11 빅데이터 프로젝트 조직 3 – 운영 및 관리

- 구축 완료된 플랫폼을 중장기적으로 유지 관리
- 대규모 하드웨어/소프트웨어로 운영 비용 높음
- 분야별 전문가 그룹이 확보되야 함
- 빅데이터 거버넌스 체계를 수립해야 함

빅데이터 기술 변화

표 1.2 빅데이터 전문 기술 영역

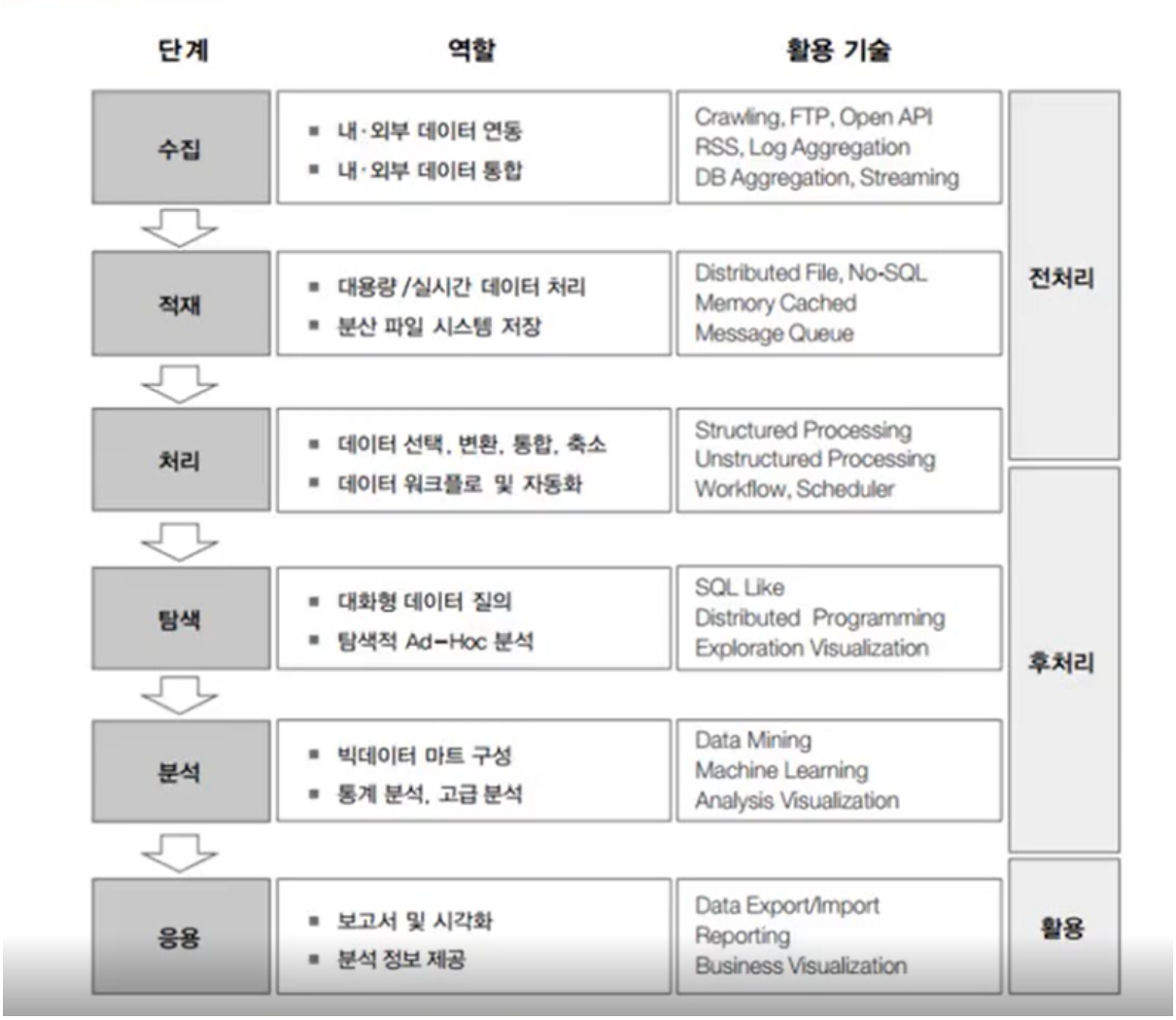
빅데이터 전문 영역		설 명	국내외 사업자
인프라스트럭처	서버	• x86급의 CPU, 메모리, 디스크 등을 장착한 서버 • 리눅스 운영체제가 설치된 서버(RedHat, CentOS 등)	HP IBM
	네트워크	• 대규모 빅데이터 서버 및 스토리지 지원을 위한 대용량(10G) 네트워크	Cisco
	스토리지	• 대규모 데이터를 저장하기 위한 내외부 스토리지 장치	Dell RedHat 등
소프트웨어 플랫폼		• 빅데이터의 전방위 기술을 포괄하는 스택 구성 (순수 오픈소스 스택 또는 기업 배포판 스택) • 빅데이터 수집/적재/처리/분석 등의 지원 솔루션 • 빅데이터 시스템 관리 및 모니터링 툴 제공 • 빅데이터 + AI 플랫폼 확장	Cloudera MapR HortonWorks KT넥스알 그루터 클라우다인 등
IT 서비스		• 빅데이터 컨설팅 및 구축 이행 • 빅데이터 전문 운영 및 유지보수 • 빅데이터 데이터/분석 서비스 • 빅데이터 교육센터 운영 및 인력 양성	KT DS LG CNS 삼성 SDS SK C&C 다음소프트 등

인프라스트럭쳐

소프트웨어 플랫폼 : 하둡을 기반으로 생태계를 만듬

IT서비스

빅데이터 구현 기술



구축 순서

- 수집

표 1.3 6V 관점의 빅데이터 수집 기술

6V	수집 기술	중요성
Volume	대용량 데이터(테라바이트 이상) 수집 대규모 메시지(1,000TPS 이상) 수집	상
Variety	정형/반정형/비정형 데이터 수집 예) Log, RSS, XML, 파일, DB, HTML, 음성, 사진, 동영상 등	상
Velocity	실시간 스트림 데이터 수집	상
Veracity	N/A	하
Visualization	N/A	하
Value	N/A	하

    - 분산기능의 선형적 확장이 필요

- DB, File, API, message 등 정형 및 비정형 데이터를 대용량으로 수집
- 외부데이터(소셜미디어, 블로그, 포털, 뉴스 등)를 수집 할 때 크롤링이 선택적으로 적용됨
- 수집처리
  - 대용량 처리
  - 실시간 수집
    - CEP, ESP 등 수집 중인 데이터에서 이벤트를 감지해 빠른 후속처리 실행
- 관련 소프트웨어
  - 플럼, 플런티드, 스크라이브, 츄카, 나이파이 등
- 실시간 처리
  - 스톰, 에스퍼

- 적재

표 1.4 6V 관점의 빅데이터 적재 기술

6V	적재 기술	중요성
Volume	대용량 데이터(테라바이트 이상) 적재 대규모 메시지(1,000TPS 이상) 적재	상
Variety	정형/반정형/비정형 데이터 수집	중
Velocity	실시간 스트림 데이터 적재	상
Veracity	데이터의 품질과 신뢰성을 확보해 적재	상
Visualization	N/A	하
Value	N/A	하

- 분산저장소
  - HDFS(대용량 파일 영구 저장)
    - 주로 HDFS를 사용하지만 실시간 및 대량으로 발생하는 작은 메시지 데이터를 HDFS에 저장할 경우 파일 수가 기하급수적으로 늘어나 관리 node와 병렬처리의 효율성이 크게 떨어짐
  - 대규모 메시징 데이타 전체 저장 - No-SQL(Mongo DB 등)
  - 대규모 데이터의 일부를 임시저장 하기 위한 In-memory cache - redis 등
  - 대규모 데이터 전체를 버퍼링하기 위한 MOM(Message Oriented Middleware) - kafka 등

    - 빅데이터가 적재될 때, 추가적인 전처리 작업이 필요할 수 있는데, 파일 형태에 따라 후처리 작업으로 할 수도 있음
    - 데이터 전처리가 도움이 될 순 있지만, 데이터의 일관성과 성능이 이와 trade off 되기 때문에 주의해야 함
- 처리/탐색

표 1.5 6V 관점의 빅데이터 처리/탐색 기술

6V	처리/탐색 기술	중요성
Volume	대용량 데이터(테라바이트 이상)에 대한 후처리 및 탐색	상
Variety	N/A	하
Velocity	N/A	하
Veracity	데이터의 품질과 신뢰성을 확보하기 위한 후처리 및 탐색	상
Visualization	후처리된 데이터셋을 시각화해서 탐색	상
Value	N/A	중

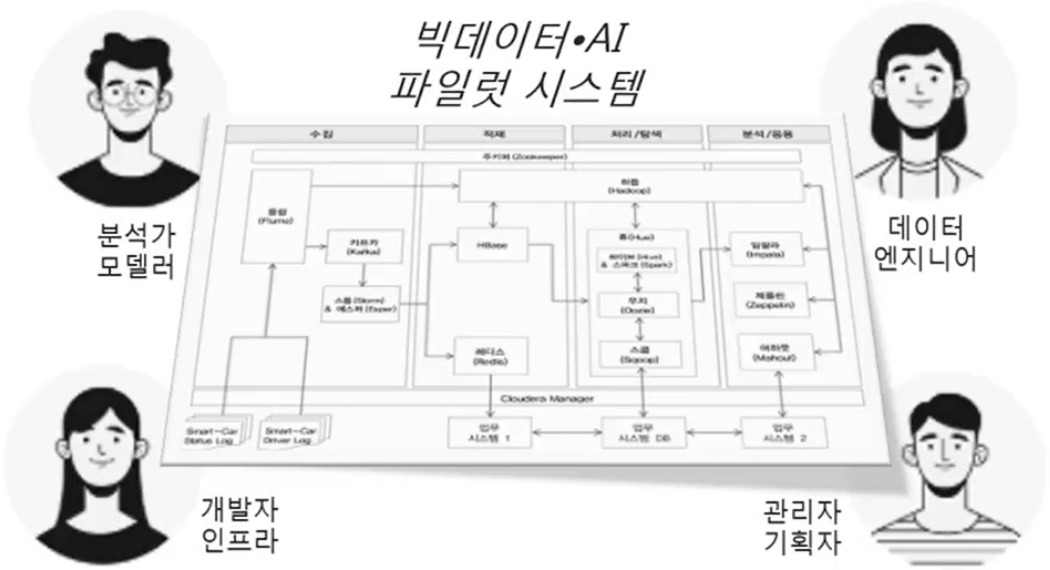
    - 데이터를 이해하는 것이 선행
    - 탐색적 분석, 탐색결과를 정기적으로 구조화
    - 탐색적 분석
        - SQL-on-Hadoop
    - 처리/탐색 기술
        - hue, hive, spark, sql 등
    - 후처리
        - uzi
- 분석/응용

표 1.6 6V 관점의 빅데이터 분석/응용 기술

6V	분석/응용 기술	중요성
Volume	대용량 데이터(테라바이트 이상) 분석	상
Variety	정형/반정형/비정형 등의 다양한 데이터 분석	상
Velocity	인메모리 기반으로 실시간 데이터 분석	상
Veracity	신뢰도 높은 분석 결과를 비즈니스에 적용	상
Visualization	분석 결과 및 창출된 가치를 시각화	상
Value	분석된 결과를 비즈니스에 적용해 가치 창출	상

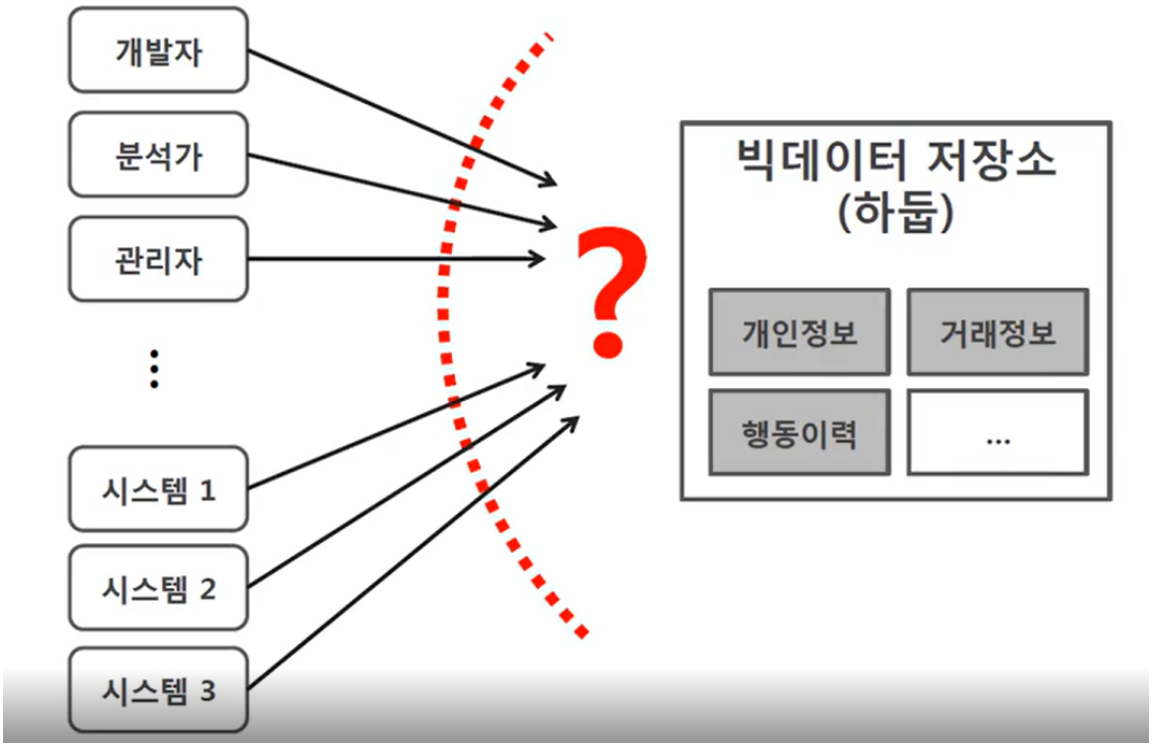
- 파일기반 보단 In-memory 기반 분석기술이 늘고 있음
- 처리 기술
  - R
  - tensorflow
  - Imfla 등

빅데이터에서 R&R



빅데이터 보안

- 데이터 보안
  - 개인정보 비식별화
  - 비식별화 + 대체키 활용
- 접근제어 보안



  - 하둡은 인증관리와 접근관리가 취약함
  - 보통 3rd-party 기술 이용
    - apache knox

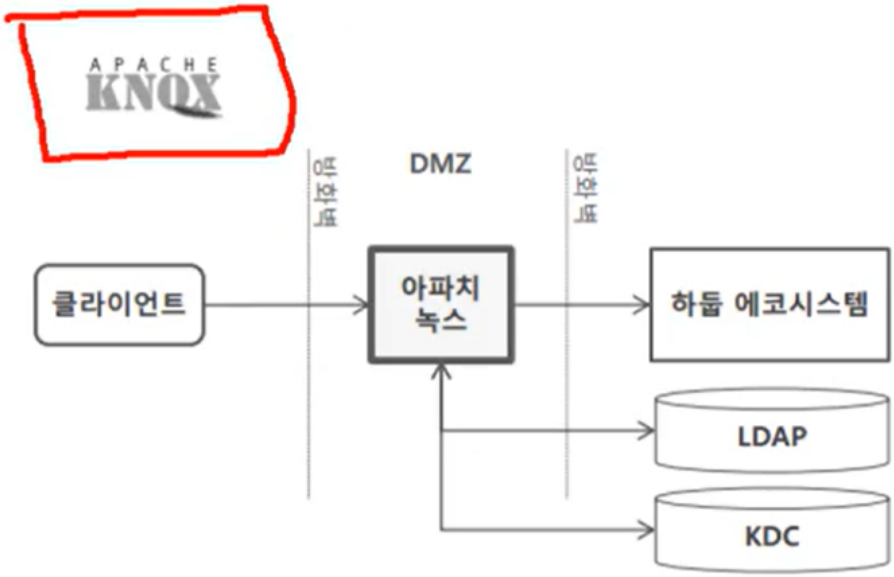
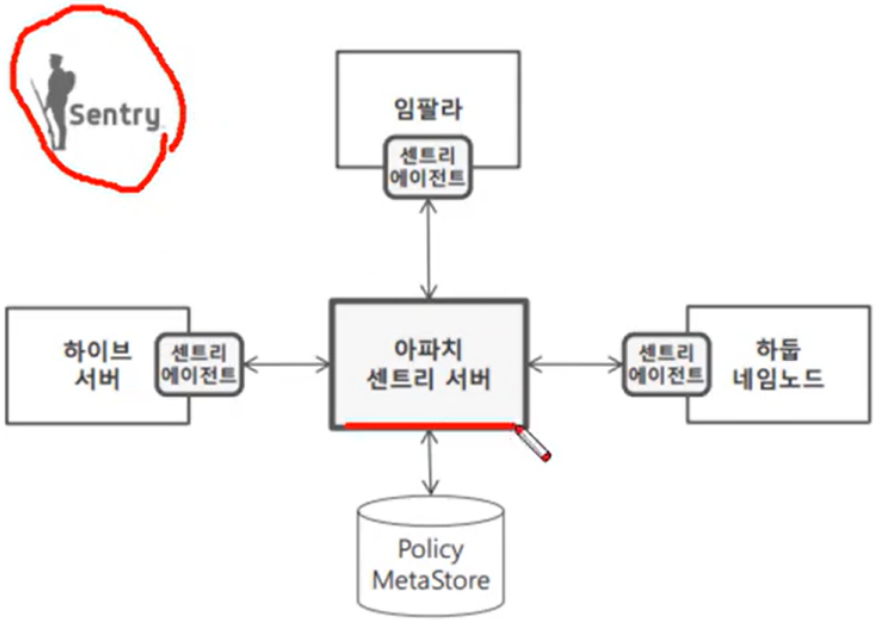


그림 1.24 빅데이터 접근제어 보안 – 아파치 녹스

- 항상 apache knox를 거치도록 하여 접근하게 함
  - LDAP, KDC 에서 개인정보를 받아 처리

- Sentry



- Policy MetaStore에서 계정을 관리

- 각각의 노드들에 센트리 에이전트를 설치
  - 중앙 서버의 접근제어 서버를 통해 이용하도록

- Apache Ranger

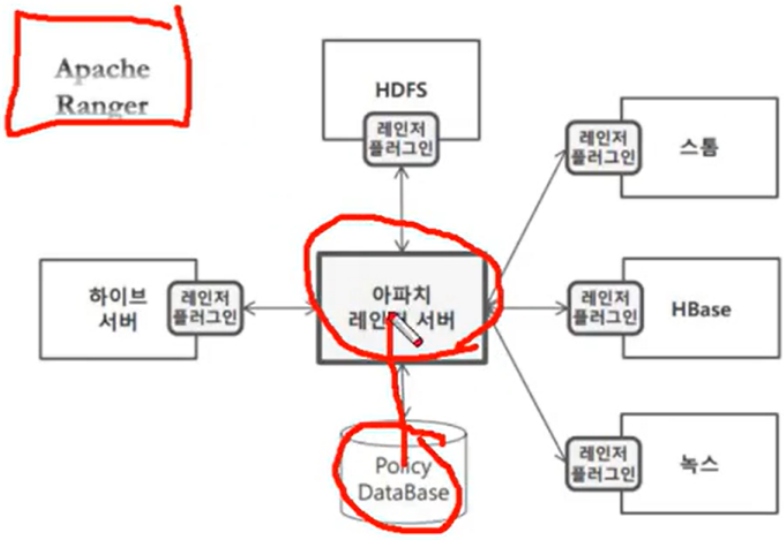


그림 1.26 빅데이터 접근제어 보안 – 아파치 레인저

  - Sentry와 유사

- 커베로스

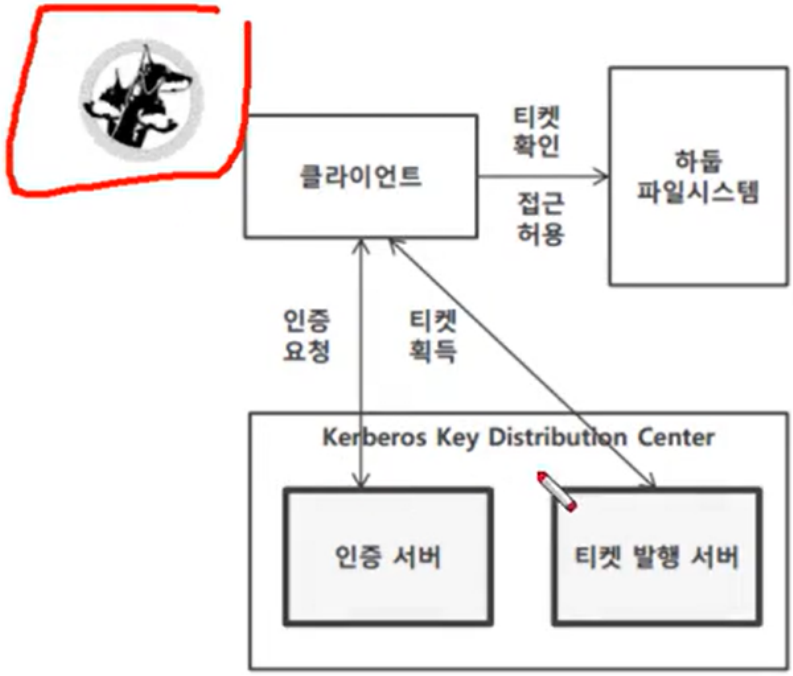


그림 1.27 빅데이터 접근제어 보안 – 커베로스

- KDC(Key Distribution Center)
  - 하둡 클라이언트는 KDC에서 티켓을 받고 하둡에 접근
    - 접근제어 통제도 KDC