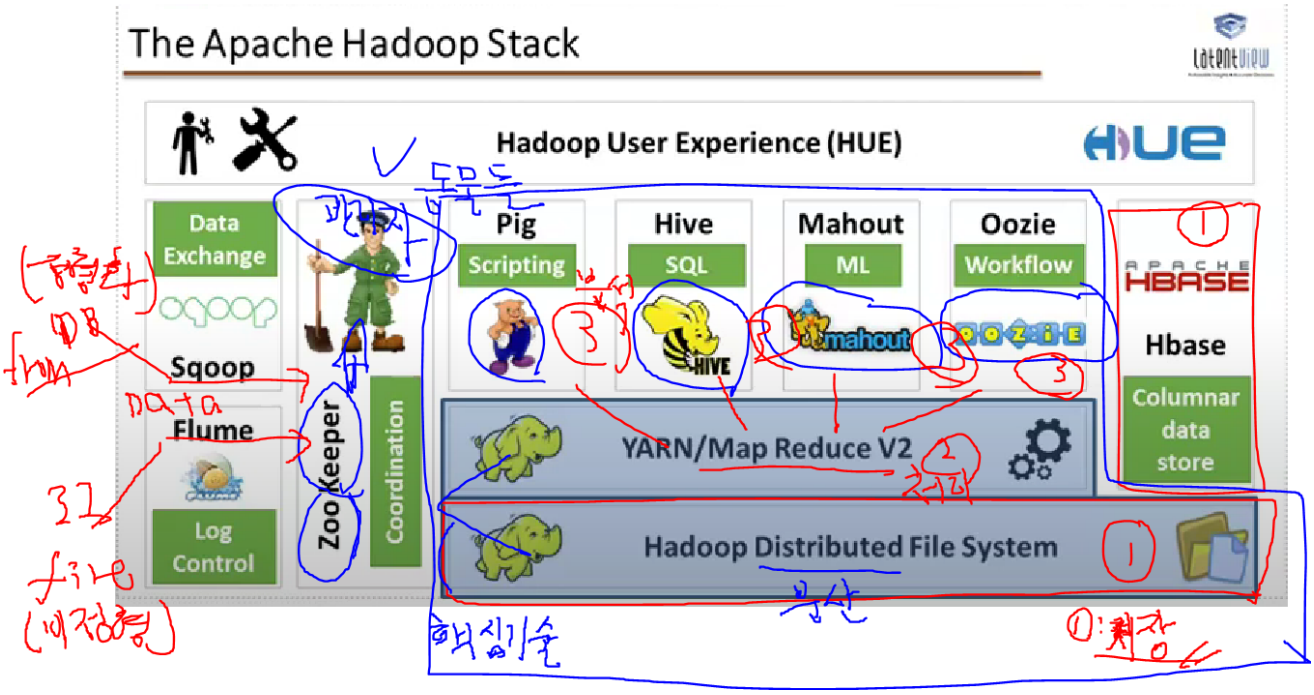
3. 하둡 에코 시스템 (Hadoop Eco System)

1. 개념도(1)



- HDFS - 원천 데이터 분산 저장
- MapReduce/YARN - 데이터를 (Key,Value)으로 분산 처리
- HBase - DB형태로 저장
- Pig / Hive / Mahout / Oozie - HBase의 데이터들을 분석해주는 툴
- HUE - 하둡에 접속하는 툴, 화면(인터페이스)

2. 하둡 기능 보완 S.W(2)

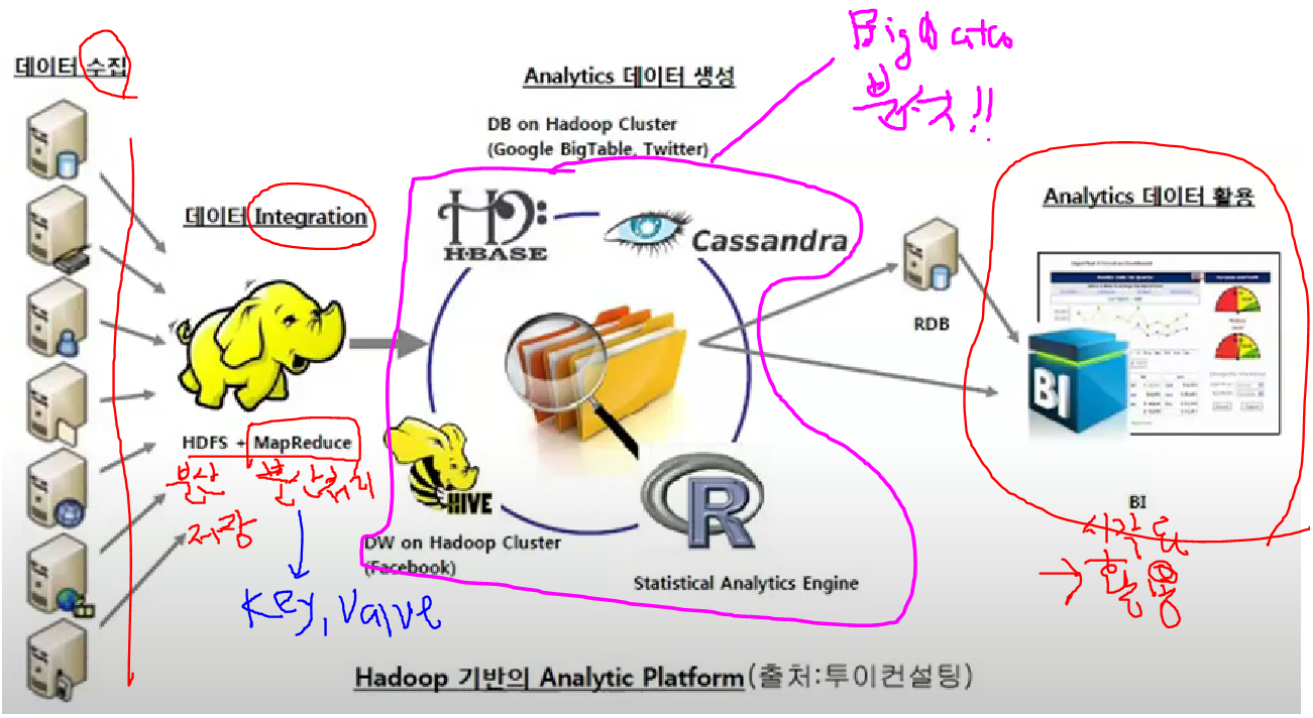
하둡의 기능을 보완하는 서브 오픈소스 소프트웨어들

구분	주요 기술	기술 별 주요 기능
빅데이터 수집	▪ 플럼(Flume) ▪ 스쿱(Sqoop)	▪ 비정형 데이터 수집 ▪ 관계형 DB로부터 데이터 가져오기
빅데이터 저장, 활용	▪ Hbase	▪ 컬럼 기반 NoSQL 데이터베이스
빅데이터 처리	▪ 하이브(Hive) ▪ 피그(Pig) ▪ 마후트(Mahout)	▪ 유사 SQL 기반 빅데이터 처리 ▪ 스크리브 언어 기반 빅데이터 처리 ▪ 기계학습 알고리즘 기반 빅데이터 처리
빅데이터 관리	▪ 우지(Oozie) ▪ H카탈로그(HCatalog) ▪ 주키퍼(Zookeeper)	▪ 빅데이터 처리 과정(Process) 관리 ▪ 빅데이터 메타 정보 관리 ▪ 빅데이터 서버 시스템 관리

하둡 에코시스템

- Hbase (일종의 DBMS) - 저장
- MapReduce (Key, Value)으로 올려보내
- HDFS (데이터가 분산 저장 됨) - 데이터를 근본적/원초적으로 저장

3. 하둡 프로세스



4. 배포본의 필요성

- 에코시스템을 구성하는 것들은 개별적인 프로그램
- 개별 프로그램들의 버전의 업그레이드는 독자적으로 이뤄짐
  -> 이에 따라 호환이 안되는 경우도 발생
- 이런 경우를 예방하기 위해, 특정 주기를 두고 업그레이드 시켜 호환성에 문제가 없도록 배포판을 만드는 것이다.

5. Pig와 Hive

- 맵리듀스를 통해서 데이터가 처리되어야 분석을 할 수 있음

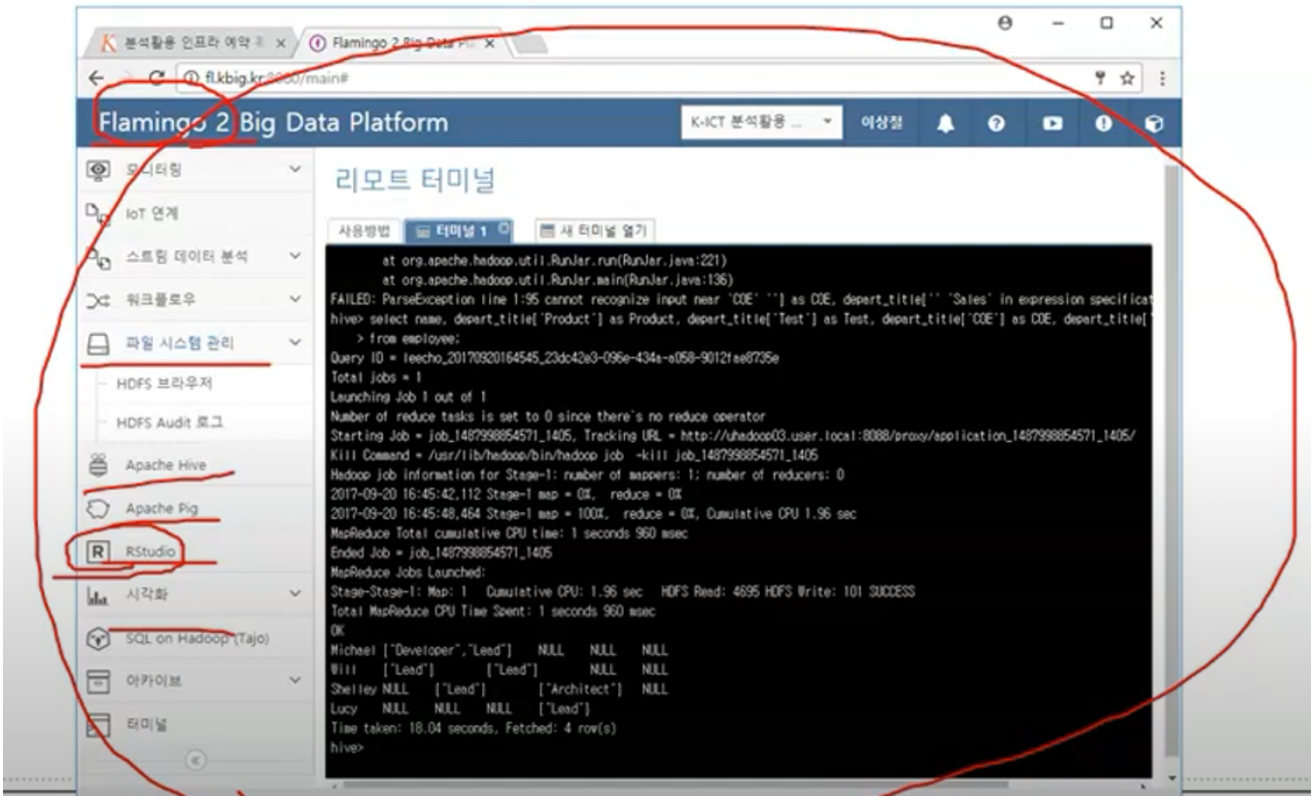
- 맵리듀스의 분석 및 처리는 JAVA 코딩을 해야함 -> diff -> 자동으로 코딩을 해주는 상위버전의 툴을 만든 것이 -> Pig와 Hive (자동으로 JAVA 코딩 처리)

- Pig : MapReduce 프로그램을 만들어주는 고수준 언어 (Yahoo)

- Hive : SQL(유사) 구문에서 MapReduce를 자동생성 (Facebook)

****초창기 때에는 HDFS와 MapReduce를 전문가들이 코딩했다면,
배포한을 통해 쓸 수 있는 상위 프로그램들을 통해 어려운 JAVA코딩이 아닌,
쉽게 할 수 있는 방법이 생긴 것

cf) HUE 인터페이스에 HIVE를 실행중인 장면



- 플라밍고(빅데이터 플랫폼)를 깔면 배포판이 깔리는 것
- 좌측 메뉴를 보면, 생태계를 이루는 프로그램들을 쓸 수 있게끔 얹어져 있음
  (Hive, Pig, R 등)

하둡 - 프레임워크, 분산 저장, 분산 처리