차세대하둡과 주목해야할 오픈소스

윤진석
<edwardyoon@apache.org>

발표자는 누구?

- 윤진석 (Edward J. Yoon)
  - 아파치 재단 위원
  - Apache Hama 프로젝트 의장
  - Apache BigTop 프로젝트 관리 위원회
  - Apache Hadoop, Whirr 개발자/커미터
  - 오라클 직원

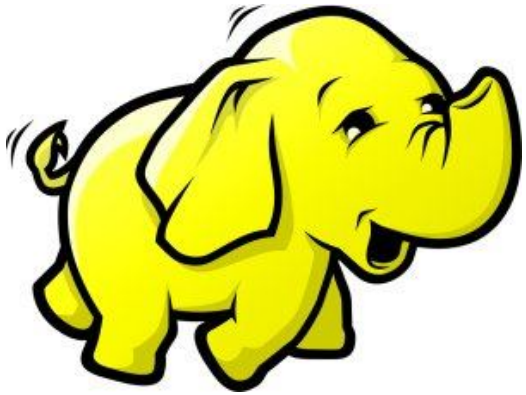
차례

- Hadoop 1.0
- 그리고, 차세대 Hadoop 2.0

- 이제 앞으로 주목해야할
  오픈 소스와 기술 트렌드

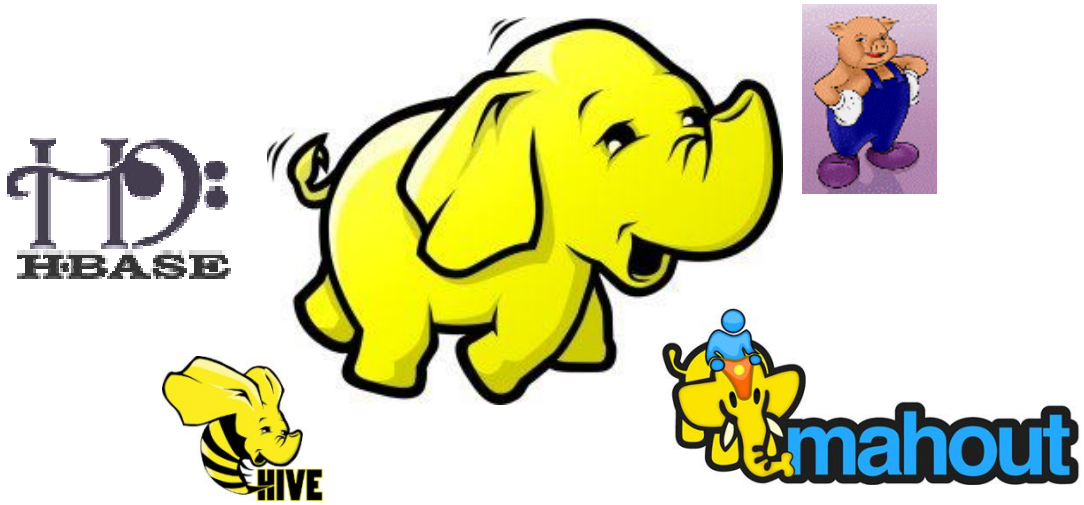
하둡 1.0

- HDFS – 분산 파일 시스템
- Map/Reduce - 분산 처리 엔진



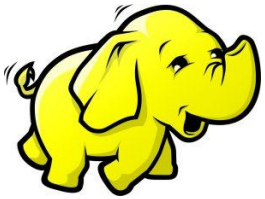
하둡 1.0

- HDFS – 분산 파일 시스템
  - HBase, Cassandra (NoSQLs)
- Map/Reduce - 분산 처리 엔진
  - Mahout, Pig, Hive



하둡 1.0 의 특성

Era of **Web Documents**



2004년 ~ 2009년

Era of **Web Applications**

차세대 하둡

- HDFS – 분산 파일 시스템
- Map/Reduce – 분산 처리 엔진

- **Map/Reduce v2 YARN**
  - MPI, BSP 등 분산 처리 엔진의 다양화

이유?

- 데이터 복잡도의 증가
- 고급 분석의 요구
  - Map/Reduce 모델의 한계

Era of **Web Applications**

MR과 차세대 컴퓨팅 엔진 비교

- Map/Reduce
  - 데이터 가공 (Relational algebraic computing)
    - 데이터 집계 or 통계
    - 간단한 확률 계산

- MPI 또는 BSP 컴퓨팅 엔진
  - 과학 연산 (Scientific computing)
    - 네트워크 분석 (e.g., social network)
    - 기계 학습
    - 수치선형대수

주목해야할 오픈소스

- Open MPI
  - MPI 라이브러리
- Apache Hama
  - Hadoop 기반 BSP 컴퓨팅 엔진
- GraphLab
  - BSP 모델 기반 그래프 처리 및 기계학습 라이브러리

Question!