LESSON 1

하둡의 역사

하둡의 역사

하둡의 아버지

더그 커팅
(Doug Cutting)

루씬 제작자

- 검색엔진용 오픈 소스 텍스트 인덱스 엔진

하둡의 역사

- 구글 검색엔진과 같은 대형 검색엔진 제작에 관심
- 데이터를 대량으로 저장할 수 있는 빅파일시스템과 분산처리구조에 관심을 가지고 있었음
- 구글의 두 가지 논문에 영감을 얻어 하둡을 제작
  - The Google File System (2003)
  - MapReduce : Simplified Data Processing on Large Cluster (2004)
- 2006년부터 제작(Apache Top Level Project)
- 야후에 취직 → 클라우데라로 이직
- GFS → HDFS, MapReduce → MapReduce
- 하둡의 이름은 더그커팅의 아들이 가지고 놀던 노란색 코끼리 인형!
  - 하둡의 기술의 이름은 코끼리와 관련 있는 이름을 지음

하둡의 역사

Hadoop

Mahout

Oozie

Horton

머하웃(Mahout) : 코끼리를 모는 사람

LESSON 2

하둡의 기초사항

하둡의 기초사항

- 하둡은 자바(Java)로 만듦
  - 다른 기술과 연결 가능
- 하둡은 유닉스기반에서 주로 사용
  - MS가 윈도우 기반으로 포팅하고 있으나 파일시스템 요구사항을 맞추기 쉽지 않음
  - 리눅스 중 우분투를 기반으로 수업을 진행
- 배포판
  - Apache Hadoop, CDH(클라우데라판), HDP(호튼웍스판)...
- 하둡의 빅3회사
  - Apache 재단
  - 클라우데라(Claudera) → 더그커팅이 있는 회사
  - 호튼웍스(HortonWorks) → 더그커팅이 예전에 다니던 회사(야후)
  - MapR(구글/아마존과 협업)

LESSON 3

하둡의 특징

하둡의 특징

- Open Source cf. github
- 데이터가 있는 곳으로 코드를 이동
- Scale Out vs. Scale Up
- 병렬처리를 가능하게 하는 단순한 데이터 모델
- 오프라인 배치 프로세싱에 최적화
  - 실시간 처리가 안 됨

LESSON 4

하둡 아키텍처

하둡 아키텍처

하둡(Hadoop) = HDFS + MapReduce

- 빅파일 시스템 : HDFS(Hadoop File System)
  - 네임노드(마스터)/데이터노드(슬레이브)
  - 세컨더리 네임노드
    - SPOF(Single Point Of Failure)문제
- 분산처리 프레임워크 : MapReduce
  - 잡트래커(마스터)/태스크트래커(슬레이브)

하둡 아키텍처

HDFS

- 하나의 HDFS 에 하나의 네임스페이스 제공
- 파일을 여러 개의 블럭으로 나누어 저장
- 블럭사이즈: 64MB(실제로는 128MB 많이 사용)
- 큰 파일을 다루는 데 적합

- 운영체제의 파일 시스템을 그대로 사용한다
- 복제수(Replication Factor)
  - 여러 군데에 같은 블럭을 복사(HA:High Availability)

하둡 아키텍처

HDFS

- Write Once Read Many
- File Overwrite not Append

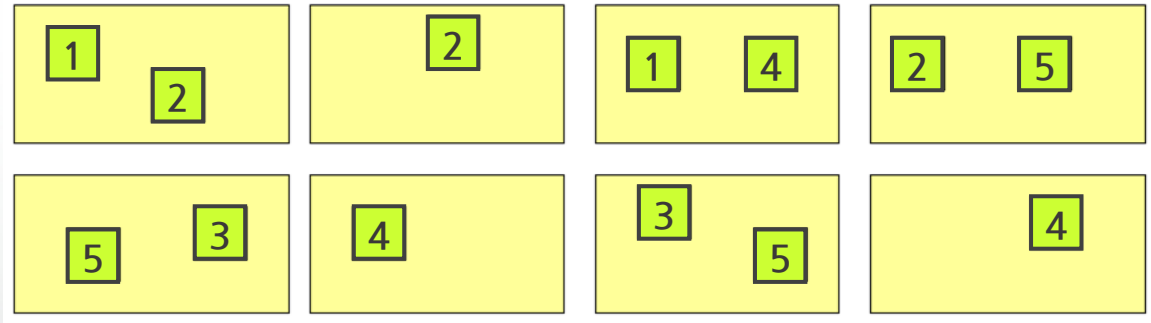
하둡 아키텍처

HDFS

Block Replication

Namenode (Filename, numReplicas, block-ids, …)
/users/sameerp/data/part-0, r:2, {1,3}, …
/users/sameerp/data/part-1, r:3, {2,4,5}, …

Datanodes



하둡 아키텍처

잡트래커(JobTracker)

- Job: 하나의 MapReduce프로그램
- 하나의 잡은 여러 개의 맵 태스크와 리듀스 태스크로 이루어짐
  - 태스크 트래커: 2개의 맵 Task와 2개의 리듀스 Task(기본)
- 역할
  - 사용자 하둡 잡 실행요청(jar파일, 입력데이터위치, 출력데이터 위치 등)을 받아
  - 잡의 태스크들을 태스크 트래커로 나눠서 실행하고 종료할 때까지 관리
- 입력데이터/출력데이터 위치는 반드시 HDFS상에 존재해야 함
- 태스크 트래커는 데이터노드와 같은 물리서버에 존재

하둡 아키텍처

MapReduce 프레임워크