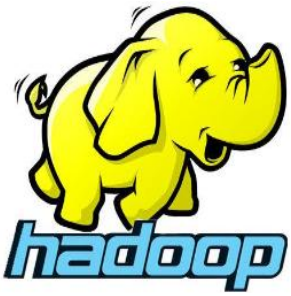
실무에 적용 가능한 Big Data 분석 개론

하둡 이해

한국기술교육대학교
온라인평생교육원

▣ 하둡의 개념과 발전과정

1. 하둡의 개념



대용량 분산 저장과 처리를 위한 프레임워크

HDFS (Hadoop Distributed File System)	맵리듀스(MapReduce)
빅데이터 파일을 여러 대의 서버에 분산 저장하기 위한 파일시스템	각 서버에서 데이터를 분산 처리하는 분산병렬처리를 위한 프레임워크

하둡(Hadoop) 특징

1. 오픈소스 자바 소프트웨어 프레임워크임
   - 대량의 자료를 처리할 수 있는 큰 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원함
2. 오픈소스 하둡
   - 빅데이터 활용을 가능하게 만든 빅데이터 플랫폼의 핵심 기술임
3. 저비용으로 방대한 양의 데이터를 저장 및 처리할 수 있음
   - 저가 장비 및 스토리지(저장장치)를 활용함

▣ 하둡의 개념과 발전과정

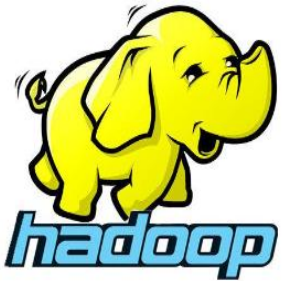
2. 하둡의 발전과정

너치 프로젝트

하둡 창시자 더그 커팅

2002년에 시작된 웹 검색엔진

전 세계에 오픈소스로 공개할 검색엔진 개발



너치를 개발할 당시, 웹이 규모가 폭발적으로 늘어나고 있었음

늘어나는 웹 페이지를 쉽게 색인 할 수
있는 기술의 개발이 필요해짐

크롤링(Crawling)과 빨리 찾아줄 수 있는 기술을 개발 함

너치 프로젝트 개발하고 오픈소스화 함

▣ 하둡의 개념과 발전과정

2. 하둡의 발전과정

너치 프로젝트(Apache Nutch)

10억 페이지 규모의 색인을 유지할 수 있었지만,
그 이상의 확장을 관리하기에는 구조적인 한계가 있음

인터넷 상에 존재하는 모든 페이지를 가져와서 저장할 수 없음

텍스트 기반의 탐색을 한다는 것은 기술적으로 불가능함

▣ 하둡의 개념과 발전과정

2. 하둡의 발전과정

너치 분산파일시스템(Nutch Distributed File System)

- 2003년 구글 파일 시스템(The Google File System)이라는 논문을 발표함
  - ➢ 급속히 늘어나는 구글의 데이터 처리를 위하여 설계된 대용량 분산 파일 시스템

아이디어 가져옴

더그 커팅

- 너치 분산파일시스템(Nutch Distributed File System)을 개발함
  - ➢ 웹 크롤링과 색인 과정에서 생성되는 굉장히 큰 파일들을 처리하기에 알맞은 구조를 가지고 있는 파일 시스템

▣ 하둡의 개념과 발전과정

2. 하둡의 발전과정

너치 분산파일시스템(Nutch Distributed File System)

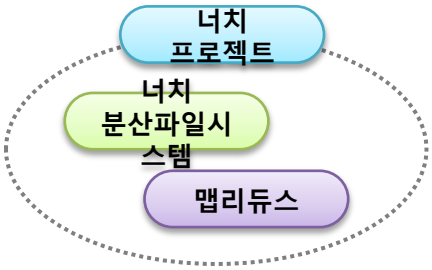
Google

· 2004년 맵리듀스를 발표함

➢ 구글 분산파일시스템 위에서 동작시켜 대용량 데이터를 간단하게 처리할 수 있음

더그 커팅

· 맵리듀스까지 프로젝트에 포함시킴



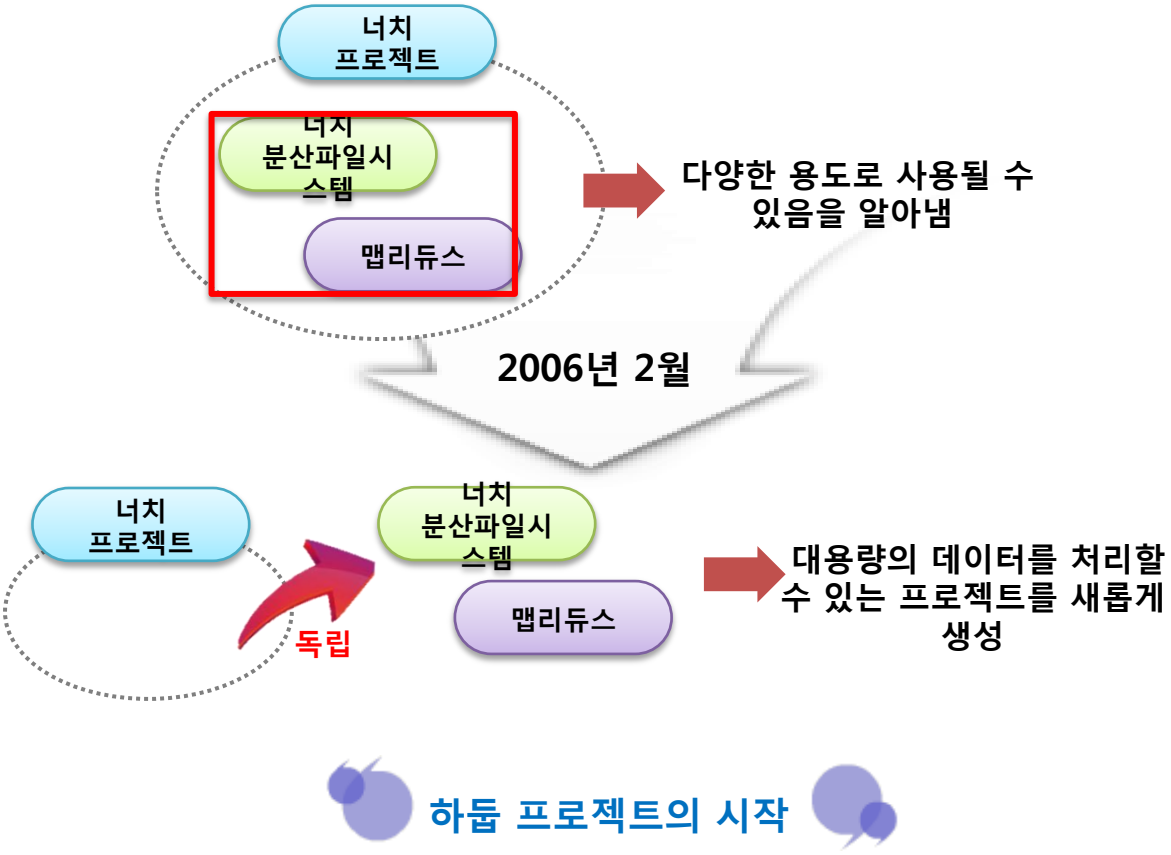
➢ 굉장히 많은 데이터를 저장할 수 있음

➢ 많은 데이터를 분산 처리 환경에서 크롤링하여 가져오고, 처리할 수 있는 기반을 구축함

▣ 하둡의 개념과 발전과정

2. 하둡의 발전과정

하둡 프로젝트



▣ 하둡의 개념과 발전과정

2. 하둡의 발전과정

하둡 프로젝트

더그 커팅

2006년, 하둡은 야후 안에서 엄청난 속도로 발전하고 성장하기 시작함

YAHOO	hadoop
• 10,000개의 하둡 코어를 이용하여 야후 서비스의 색인 제품들이 생성되고 있다고 발표함(2008.02)	• 오픈소스의 절대강자인 아파치 소프트웨어 재단에서 최고의 프로젝트에 등극하여 확실히 이름을 알리기 시작함(2008.02)

▣ 하둡의 개념과 발전과정

2. 하둡의 발전과정

하둡 프로젝트

▶ 분산 처리 속도를 높여 다시 한 번 발전하기 시작함

당시 데이터 처리 상황		하둡의 데이터 처리 상황
• 대부분의 하드디스크 용량 1TB • 초당 읽을 수 있는 용량 약 100MB ➡ 1TB 읽는데 2시간 30분 이상 소요	VS.	• 분산 환경에서 처리할 수 있도록 만들어졌기 때문에 비교할 수 없을 정도로 빠르게 처리할 수 있었음 ➡ 2007년, 1TB 읽고 정렬하는데 297초 소요

▶ 테라바이트 데이터 정렬을 위한 가장 빠른 시스템

2008.04	2008.11	2009.05
209초	**68초**	**62초**
• 910개 컴퓨터 클러스터로 1테라 바이트를 209초 만에 정렬함 • 세계 기록 경신 • 전년도 우승자 297초		• 야후가 하둡을 사용함

하둡은 다양한 목적에 맞게 다양한 데이터 처리 방식을 수용하는 플랫폼으로 계속 진화

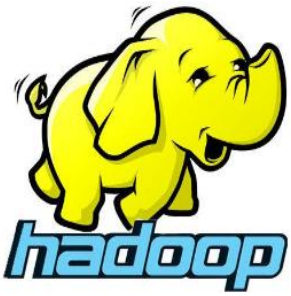
실무에 적용 가능한 Big Data 분석 개론

하둡 이해

한국기술교육대학교
온라인평생교육원

▣ 하둡의 구성요소

1. 분산파일시스템



대용량 분산 저장과 처리를 위한 프레임워크

HDFS (Hadoop Distributed File System)	맵리듀스(MapReduce)
빅데이터 파일을 여러 대의 서버에 분산 저장하기 위한 파일시스템	각 서버에서 데이터를 분산 처리하는 분산병렬처리를 위한 프레임워크

하둡 분산파일시스템 특징

하둡이 사용하는 분산 저장소
분산된 환경에서 다양한 형태, 초대용량의 데이터를 안전하게 저장할 수 있을 뿐만 아니라 저장되어 있는 데이터를 빠르게 처리할 수 있도록 설계됨

- 분산된 환경에서
  - 우수한 저장력
  - 우수한 데이터 처리

전체 성능이나 용량을 늘리기 위해 많은 서버를 이용하여 구축함

- 값싼 서버들 이용
- 높은 수준의 고장방지기능 이용

▣ 하둡의 구성요소

1. 분산파일시스템

하둡 분산파일시스템 특징

노드(Node)

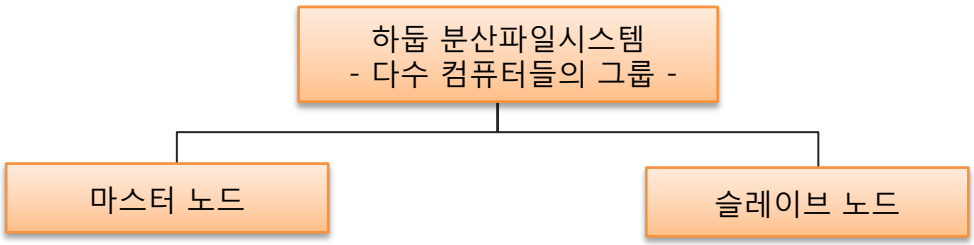
- **분산파일시스템에서는 마스터 노드가 동작하는 서버와 슬레이브 노드가 동작하는 서버로 구성되어 있음**

- 네트워크에서 사용되는 용어
- 일반적으로 네트워크에 연결되어 있고 전송 채널을 통해 자료를 주고 받을 수 있는 장비를 의미
- 하둡에서는 모든 서버들이 네트워크에 의해 연결되어 있고 통신 가능하도록 구성되어 있어 노드는 각 구성요소들이 동작하는 서버라고 이해하면 됨

▣ 하둡의 구성요소

1. 분산파일시스템

하둡 분산파일시스템 특징



▶ 마스터 노드

1 현재 분산 파일 시스템에서 사용하고 있는 모든 슬레이브 노드들을 관리

- 일반적으로 수십, 수백, 수천 개의 슬레이브 노드들이 동작함
- 어떤 슬레이브 노드가 분산 파일 시스템을 위해서 동작 중인지 혹은 고장으로 인해 더 이상 동작 하지 않는지의 정보를 실시간으로 파악

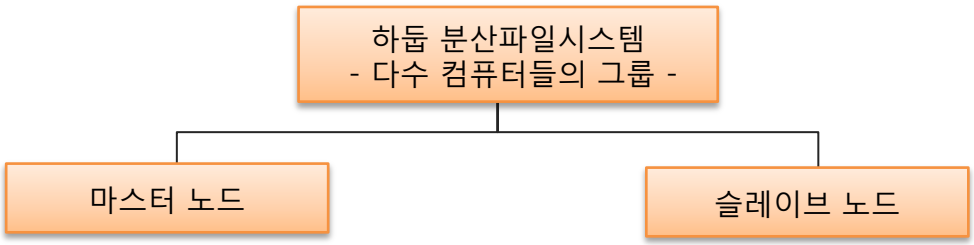
2 디렉터리와 파일에 대한 정보를 포함하는 메타데이터를 관리

- 사용자가 생성한 디렉터리 구조와 파일 목록을 보관하고 있어야 함
- 데이터가 어느 슬레이브 노드에 저장되어 있는지 정확히 알고 있어야 함
- 슬레이브 노드 정보를 실시간으로 확보함으로써 사용자가 데이터 업로드 요청 시 이 정보를 바탕으로 어느 슬레이브에 데이터를 저장할지 결정

▣ 하둡의 구성요소

1. 분산파일시스템

하둡 분산파일시스템 특징



▶ 슬레이브 노드

1. 사용자의 데이터 저장 및 전달
2. 하나의 파일을 여러 개의 슬레이브 노드에 동일하게 복제하여 관리

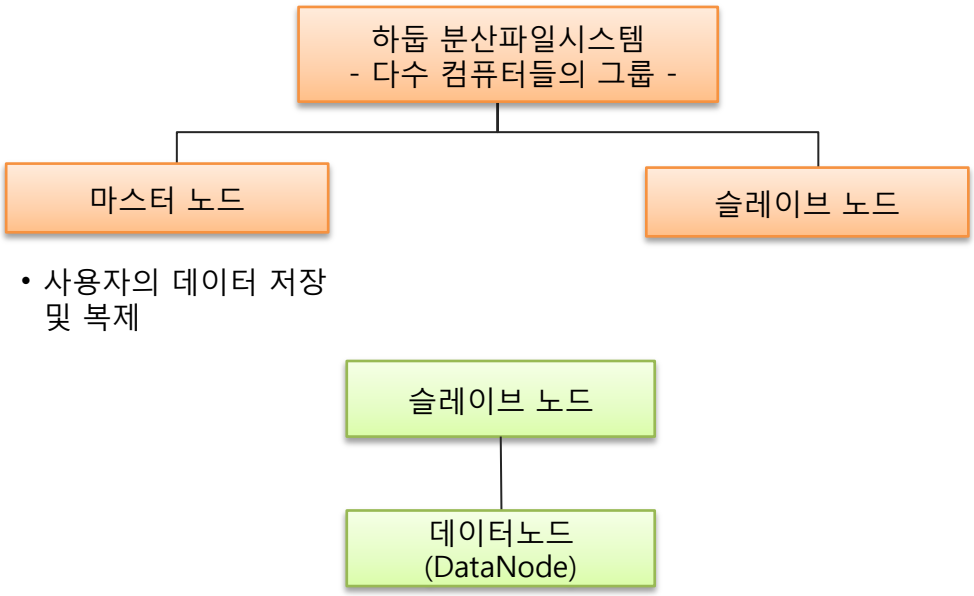
- 디스크 고장과 서버의 고장을 쉽게 처리하기 위해서 분산 파일 시스템에서는 여러 서버에 사용자의 데이터를 복제하도록 설계
- 사용자의 데이터를 안전하게 보관할 수 있고, 서버가 고장 나더라도 언제든지 데이터를 사용

▣ 하둡의 구성요소

1. 분산파일시스템

하둡 분산파일시스템 특징

마스터 노드와 슬레이브 노드의 역할

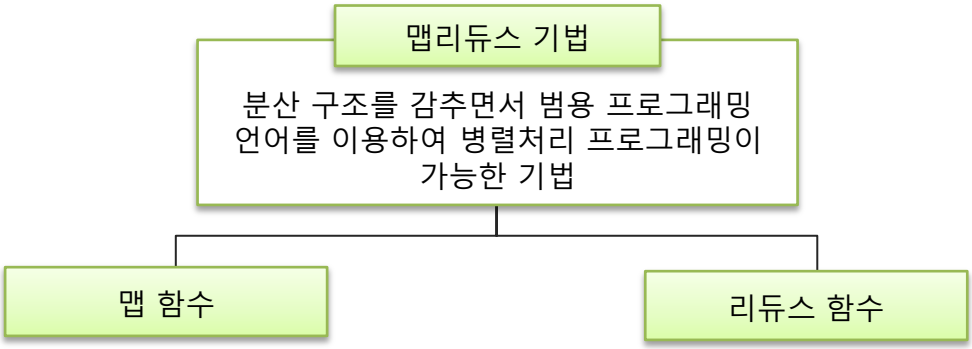


▣ 하둡의 구성요소

2. 맵리듀스

맵리듀스 개요

저렴한 머신들을 이용하여 빅 데이터를 병렬로 분산 처리하기 위한 프로그래밍 모델



맵리듀스 처리 성능에 영향을 주는 사항

- 여러 서버 중에서 하나의 서버가 제대로 동작하지 않거나 멈추는 문제
- 동시 처리를 위해 각 프로세스 간의 스케줄링 고려
- 장치 간의 네트워크 구성 고려

▣ 하둡의 구성요소

3. 하둡 에코시스템

하둡 에코시스템

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

<출처 : '빅데이터' 플랫폼의 미래', Technology Inside LG CNS R&D Journal, 이주열, 2013>

▣ 하둡의 구성요소

3. 하둡 에코시스템

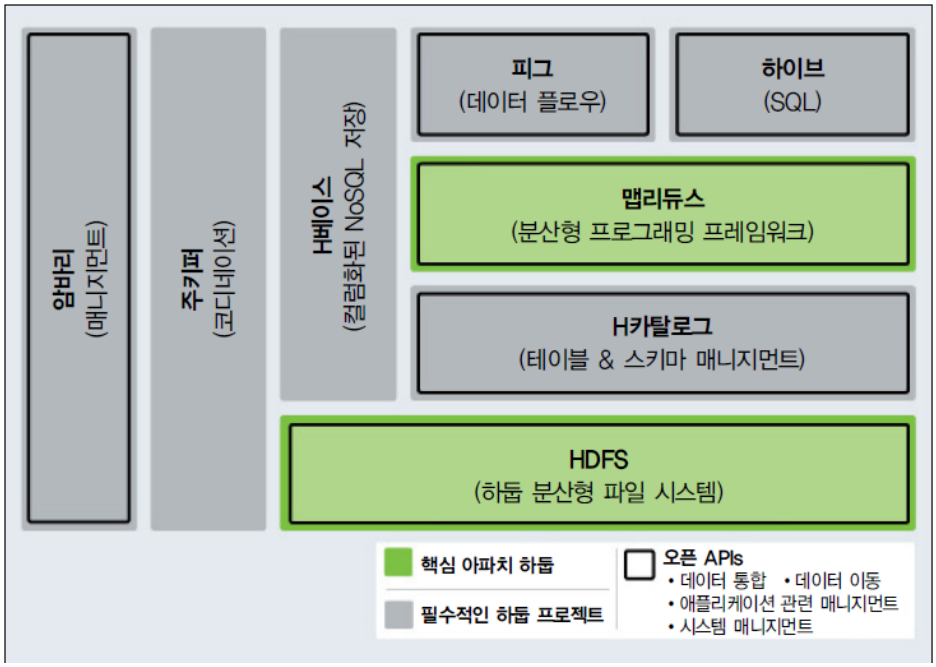
하둡 에코시스템 지도

<출처 : '빅데이터' 플랫폼의 미래', Technology Inside LG CNS R&D Journal, 이주열, 2013>

- 하둡 에코시스템의 주요 기술은 대부분은 동물 이름과 관련 있는데. Hadoop은 창시자 더그 컷팅의 아들이 노란 코끼리 장난감을 '하둡'이라 부르는 것을 보고 이름을 지었고, Hive는 벌떼, Pig는 돼지 그리고 이들을 관리하는 주키퍼는 동물 사육사라는 뜻임
- 하둡 생태계를 구성하고 있는 주요 솔루션은 플럼, 스쿱, HBase, 마후트 등이 있음
- 하둡 관련 오픈소스 프로젝트는 해마다 이어지고 있는데 이들 솔루션은 대부분 오픈소스화 되어 있음

▣ 하둡의 구성요소

3. 하둡 에코시스템



하둡 에코시스템 구성

< 출처 : IDG TechReport,빅 데이터를 위한 개방형 DB 프레임워크 "하둡"의 이해>

- 하둡 에코시스템에는 기본 요소인 하둡 분산파일시스템과 맵리듀스, 분산 데이터베이스인 H베이스, 관계형 대수 쿼리 언어 인터페이스인 피그(Pig), 데이터웨어하우징 솔루션인 하이브, 테이블 및 스토리지 관리 서비스인 H카탈로그, 그리고 매니지먼트인 암바리와 코디네이션인 주키퍼 등이 포함됨

▣ 하둡의 구성요소

3. 하둡 에코시스템

더그 커팅

하둡을 중심으로 성장한 프로젝트들의 전체 생태계가 있으며, 이는 계속해서 성장하고 있습니다.

하둡은 분산된 운영체제의 커널이며, 커널 주변의 다른 모든 구성요소들은 현재 개발 중에 있는데 피그와 하이브는 그런 것들의 좋은 예라 할 수 있습니다.

단순히 하둡만을 사용하는 사람은 없습니다.

▣ 하둡의 구성요소

3. 하둡 에코시스템

피그(Pig)

고급 수준의 데이터 처리 언어

- 하둡 프로그래밍을 간단하게 해주고 하둡의 확장성과 안정성을 유지시켜줌
- 고급 수준의 언어들에서 클래스 연산들로 여겨져야 하는 어떤 기능들은 맵리듀스 기능을 수행하기가 어려운데 이것은 이러한 점을 개선해 줌
- 대용량 데이터 집합을 좀 더 고차원적으로 처리할 수 있도록 해줌
- 다중 값과 중첩된 형태를 보이는 좀 더 다양한 데이터 구조를 지원하고 데이터에 적용할 수 있는 변환 종류도 훨씬 더 강력함

하이브(Hive)

하둡의 최상위층에 있는 데이터 웨어하우징 패키지

- 다수의 사용자 및 대용량 로그 데이터 처리를 위해 페이스북에서 개발한 정보 플랫폼 중 가장 중요한 구성 요소
  - ➢ 페이스북의 급증하는 소셜네트워킹에서 매일같이 생산되는 대량의 데이터를 관리하고 학습하기 위해 개발됨
- 페이스북은 HDFS에 대량의 데이터를 저장해 두고 하이브가 제공하는 믿을만한 SQL 기법을 이용하여 데이터를 분석함
- 많은 조직들이 채택하는 성공적인 아파치 프로젝트로 성장함