**01 하둡 살펴보기**

**하둡(Hadoop)이란 ?**

하둡은 대용량 데이터를 분산 처리할 수 있는 자바 기반의 오픈소스 프레임워크.

하둡은 분산 파일 시스템인 HDFS에 데이터를 저장하고, 분산 처리 시스템인 Map-Reduce를 이용해 데이터를 처리.

**왜 하둡인가 ?**

1. 값이 저렴.
2. 데이터 저장 용량이 부족할 경우, 필요한 만큼 리눅스 서버만 추가하면 됨 🡪 편리함
3. 데이터의 복제본을 저장하기 때문에 데이터의 유실이나 장애가 발생했을 때도 데이터 복구가 가능.

🡪 저렴한 구축 비용과 비용 대비 빠른 데이터 처리, 그리고 장애를 대비한 특성.

**하둡 에코 시스템**

하둡은 비즈니스에 효율적으로 적용할 수 있게 다양한 서브 프로젝트를 제공. 이러한 서브 프로젝트가 상용화되면서 하둡 에코시스템이 구성됨;

*\*\* OLAP(On-line Analytical Processing)*

**하둡의 단점**

1. 데이터 수정 불가 – HDFS에 파일을 저장하면 파일의 이동이나 이름 변경은 가능하지만 파일의 내용은 수정이 불가능.

**02 하둡 분산파일시스템**

**HDFS 기초**

HDFS와 기존 대용량 파일 시스템의 가장 큰 차이점은 **저사양 서버**를 이용해 스토리지를 구성할 수 있다는 것 🡪 비용 절감

4가지 목표

1. **장애 복구**

장애를 빠르게 감지하고 대처할 수 있게 설계되었음.

1. **스트리밍 방식의 데이터 접근**

HDFS는 클라이언트의 요청을 빠른 시간 내에 처리하는 것보다는 동일한 시간 내에 더 많은 데이터를 처리하는 것을 목표로 함.

인터넷 뱅킹, 인터넷 쇼핑몰과 같은 서비스에서는 HDFS 부적합

1. **대용량 데이터 저장**하나의 파일이 기가바이트에서 테라바이트 이상의 크기로 저장될 수 있게 설계되었음.
2. **데이터 무결성 = 데이터의 일관성**

**HDFS 구조**

블록 구조 파일 시스템

* HDFS에 저장되는 파일은 특정 크기의 블록으로 나누어져 분산된 서버에 저장됨 – 128MB

Name Node & Data Node

* Master Server(Name Node)
  + 메타 데이터 관리
  + 데이터 노드 모니터링
  + 블록 관리
  + Client 요청 접수
* Slave Server(Data Node)
  + Client가 HDFS에 저장하는 파일을 로컬 디스크에 유지

(파일 – 실제 데이터가 저장돼 있는 Raw Data 파일, 메타데이터가 설정돼 있는 파일)