

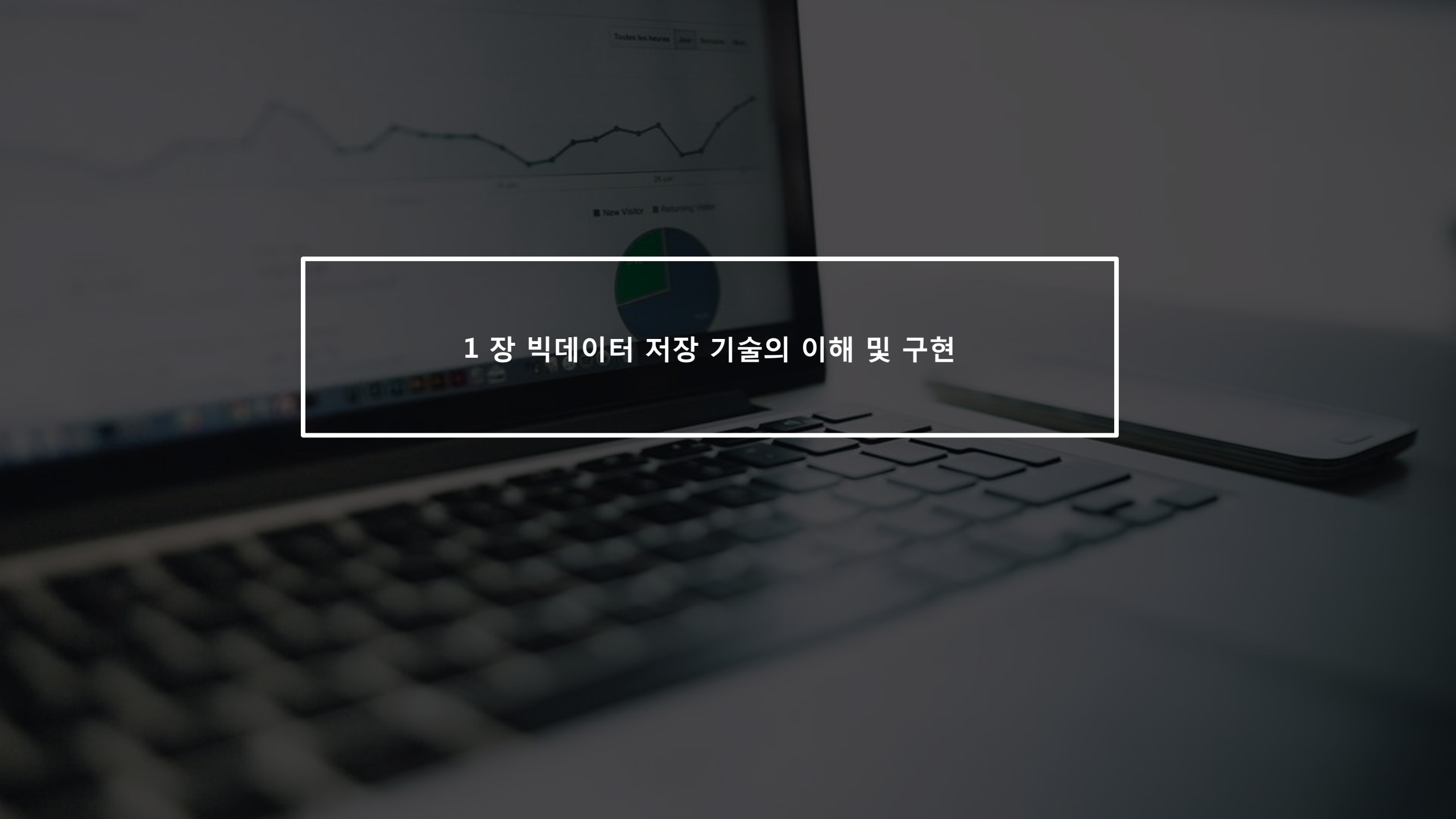


Toutes les heures

New Visitor

CMI KOREA

오픈 소스 하둡을 이용한  
빅데이터 저장 플랫폼



## 1 장 빅데이터 저장 기술의 이해 및 구현

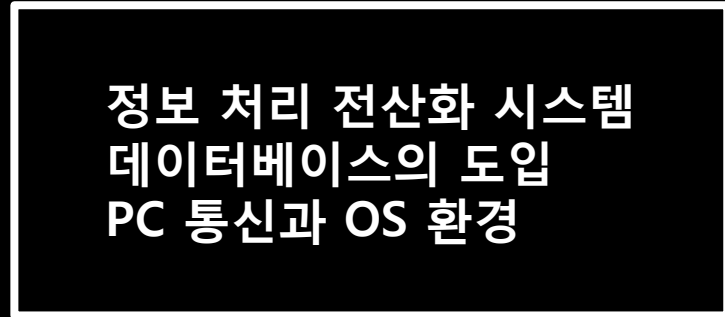
## 학습 목표

---

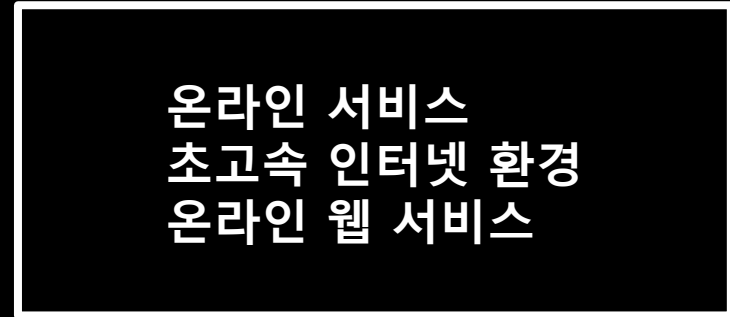
- 왜 빅데이터 처리 기술이 필요해 졌는지에 대한 이해
- 기존 기술과 빅데이터 처리 기술은 무엇이 다른지에 대한 이해
- 빅데이터 시대에 데이터베이스의 이슈는 무엇인지에 대한 이해
- 정보 통신 기술 발달로 인한 사회의 변화에 대한 이해
- 분산 파일 시스템의 필요성 및 특징

# 빅데이터 활용 인프라가 왜 필요해졌는가?

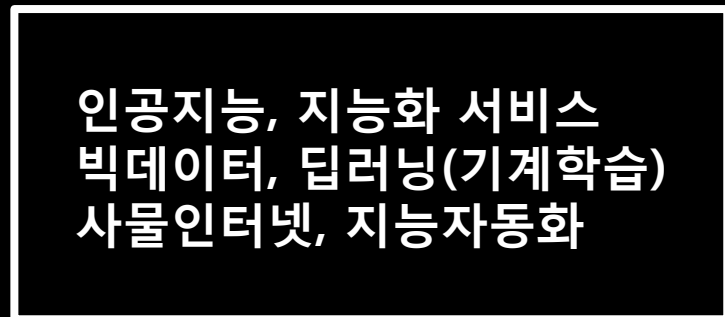
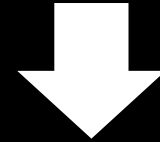
## 정보 기술의 변화에 따른 PC 환경



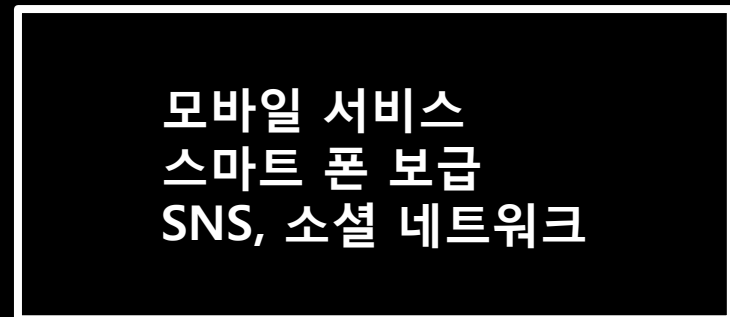
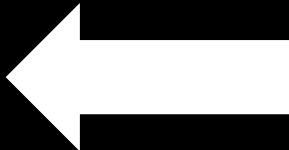
1인 1 PC 보급



인터넷, 포털



사물 정보화 구현



모바일 보급

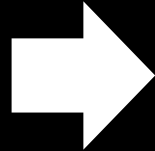
## 데이터 변화 및 흐름 변화

---

### 정보 기술의 변화에 따른 요구사항

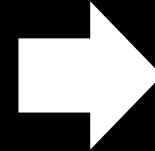
1. 데이터 생성 방법
2. 데이터 생성 크기
3. 데이터 생성 주기
4. 데이터 생성 유형

데이터의 다양한 변화



1. 정보 검색 서비스
2. 실시간 서비스
3. 자동화 서비스
4. 지능화 서비스

사회 문화적 변화



1. 지속적인 트렌드 발굴
2. 데이터, 정보 기술과 다양한 비즈니스의 긴밀한 연결
3. 심층적 분석 및 활용

비즈니스 요구 변화

## 데이터의 종류

정형	고정된 필드에 지정된 데이터, 일정한 규칙에 따라 체계적으로 정리한 데이터. 이런 데이터는 정형화된 그 자체로도 의미 해석이 가능하며, 바로 활용이 가능한 데이터를 포함한다. 예] 관계형 데이터베이스, 스프레드시트
반정형	고정된 필드에 저장되어 있지는 않지만, 메타데이터나 스키마 등을 포함하는 데이터. 반정형 데이터는 한글이나 MS워드 등으로 작성한 데이터이다. 페이스북, 트위터, 카카오톡 등 소셜 네트워크 서비스 사용자가 생성하는 데이터들이 이에 해당한다. 예]XML, HTML 텍스트 등
유사정형	노력과 시간을 들여 형식화할 수 있는 불규칙 형식의 문서 예] 웹 로그 데이터
비정형	고정된 필드에 저장되어 있지 않은 데이터 예]텍스트 분석이 가능한 텍스트 문서, 이미지.동영상.음성 데이터

## 빅데이터 활용 인프라가 왜 필요해졌는가?

빅데이터 : 형식이 다양하고 순환속도가 매우 빨라서 기존 방식으로 관리, 분석이 어려운 데이터를 다루는 기술과 분석 이를 활용한 비즈니스 등을 포함



### V3 (Volume, Velocity, Variety)

**Volume** : 데이터의 크기를 말합니다. 기업데이터나 센서데이터 등 제타 바이트 규모로 확장되는 데이터들은 통합 시스템에 존재하며, 분석 처리하기 위한 네트워크 데이터의 증가 등이 이에 속합니다.

**Velocity** : 데이터의 처리 속도를 이야기 합니다. 데이터를 수집하고 가공, 분석하는 일련의 과정을 일정 시간 내에 처리할 수 있어야 합니다.

**Variety** : 각 데이터 형태별 처리의 복잡성입니다. 기존의 정형된 데이터만 분석하는 것이 아니라, 다양한 형태의 데이터에 다른 기술들을 적용하여 필요한 데이터의 가치를 만드는 것을 이야기합니다.



### 왜 새로운 시스템 환경을 요구하는가?

#### 1. 정보 처리 기술의 이슈

1-1. 대용량 데이터의 입력과 출력에 대한 요구사항을 만족하는 환경이 필요하다.

- 
- ```
graph LR; A["1. Office 툴 이용<br/>2. 프로그래밍 파일 입출력<br/>3. 데이터 변환 및 출력<br/>4. 입력 요소들의 관리"] --> B["1. 데이터베이스 이용<br/>2. JDBC를 통한 서비스<br/>3. 정형 데이터 관리"]; B --> C["물리적인 시스템<br/>환경의 한계극복<br/>(H/W, N/W 등)"]
```
1. Office 툴 이용
  2. 프로그래밍 파일 입출력
  3. 데이터 변환 및 출력
  4. 입력 요소들의 관리

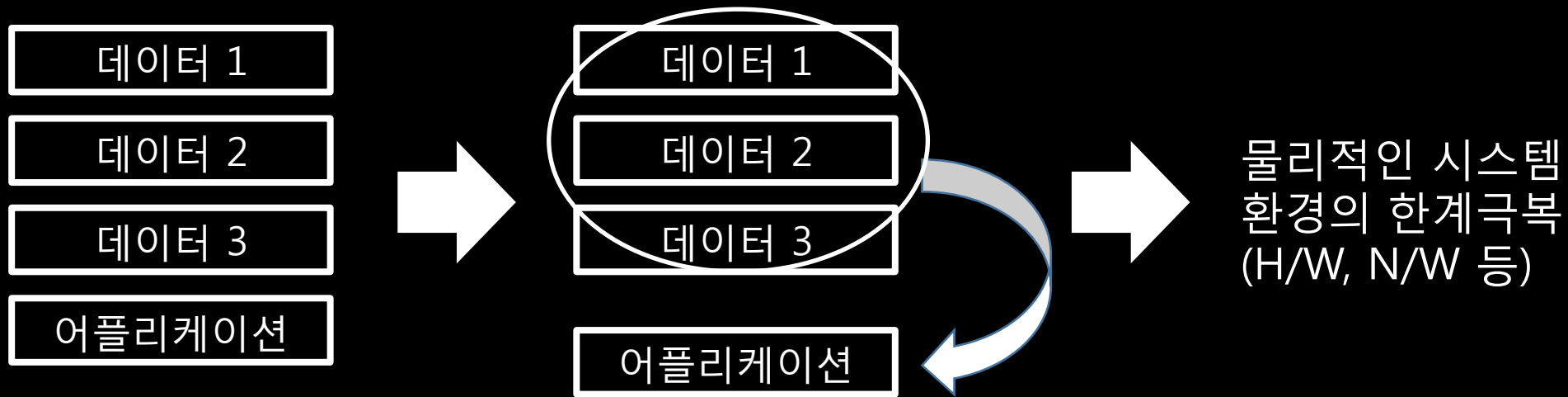
1. 데이터베이스 이용
2. JDBC를 통한 서비스
3. 정형 데이터 관리

물리적인 시스템  
환경의 한계극복  
(H/W, N/W 등)

### 왜 새로운 시스템 환경을 요구하는가?

#### 1. 정보 처리 기술의 이슈

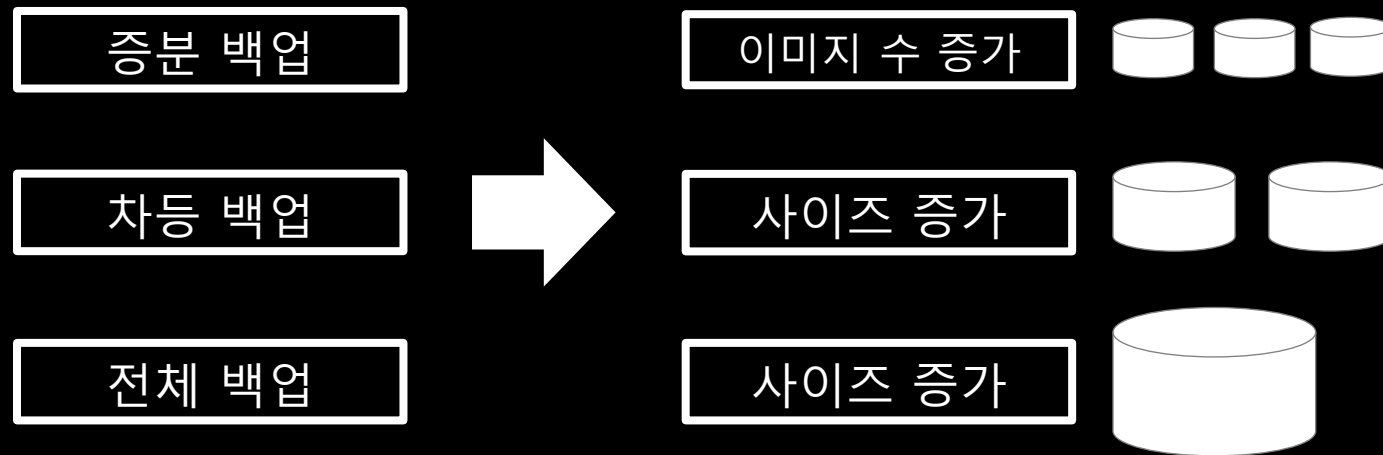
1-2. 데이터 처리를 위해 어플리케이션 수준에서 데이터를 한 곳에 모아서 처리가 이루어진다.



### 왜 새로운 시스템 환경을 요구하는가?

#### 1. 정보 처리 기술의 이슈

##### 1-3. 데이터 량이 증가함에 따라 백업과 복원에 처리 속도 지연 문제



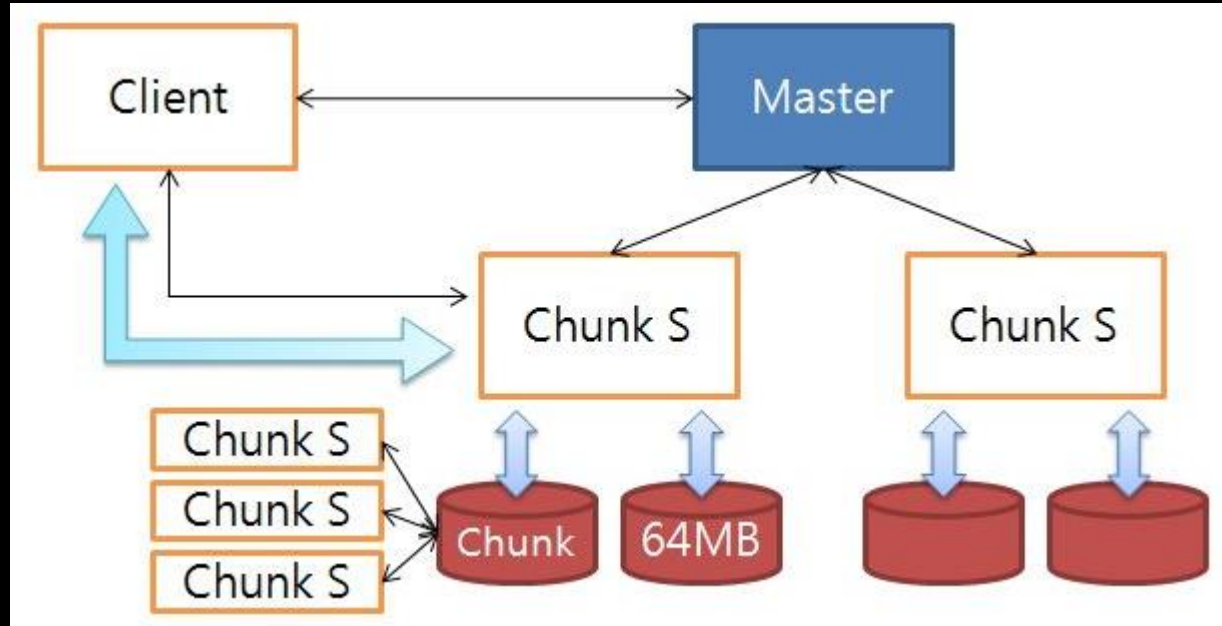
### 분산 환경의 구글 파일 시스템(GFS, Goole File System)

구글에 의해 자기 회사 사용 목적으로 개발된 분산파일 이시스템으로 일반 사용 하드웨어를 이용하여 여러대의 서버를 연결하고 데이터에 대한 접근을 효율적이고 안정적으로 구현한 파일 시스템 모델로써 근래 대용량 데이터를 저장하고 관리하는 S/W의 롤 모델로 유명하다.

#### 설계 목표

1. 증가하는 대용량 데이터를 저장하기 위한 구글의 핵심 데이터 스토리지와 구글 검색 엔진을 위해 최적화 되어 있다.
2. 구글 초창기에 레리 페이지와 세르게이 브린이 개발한 빅파일에서 개선되었다.
3. 데이터가 덮어쓰거나 삭제하는 경우가 극히 드물며 추가되거나 읽혀지는데 유리하다.

### 분산 환경의 구글 파일 시스템(GFS, Goole File System)



### GFS의 동작 설명


1. 클라이언트에서 마스터에 읽기와 쓰기에 대한 요청 전달
2. 마스터는 클라이언트와 제일 가까운 청크 서버의 정보 확인 후 전달
3. 클라이언트는 마스터에서 제공한 청크 서버의 정보를 바탕으로 읽고 쓰기 실행

# 오픈 소스 분산 파일 시스템

---

## 오픈 소스 하둡

1. 2003년 구글의 구글 파일 시스템을 모델로 시작
2. 야후의 검색 엔진 로그를 수집하기 위한 파일 시스템
3. Doug Cutting에 의해 2005년도 개발된 분산 파일 시스템
4. 분산 저장과 분산 처리를 지원
5. 마스터와 슬레이브 구조를 가짐



감사합니다.