데이터분석 전문가 가이드

과목 2. 데이터 처리 기술 이해
제 2장 데이터 처리 기술

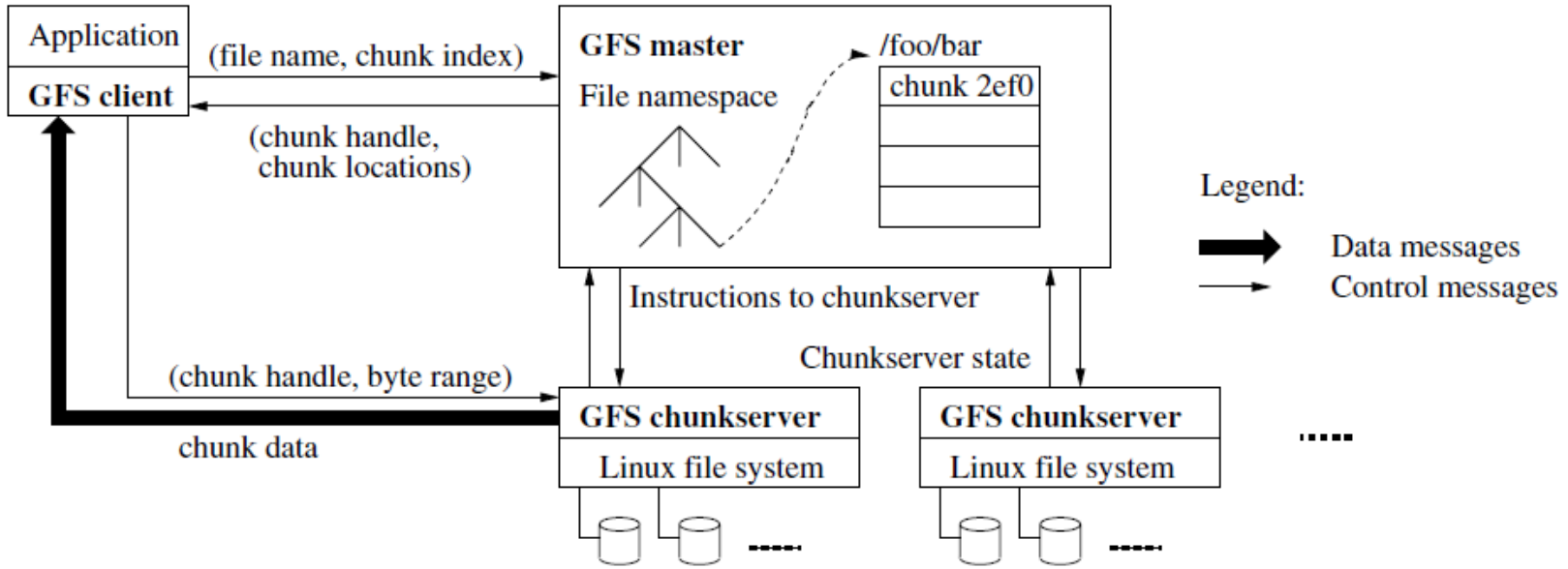
출처 : 데이터분석 전문가 가이드, 한국데이터베이스진흥원

제 1절 분산 데이터 저장 기술

1. 분산 파일 시스템

가. 구글파일 시스템(GFS, Google File System)

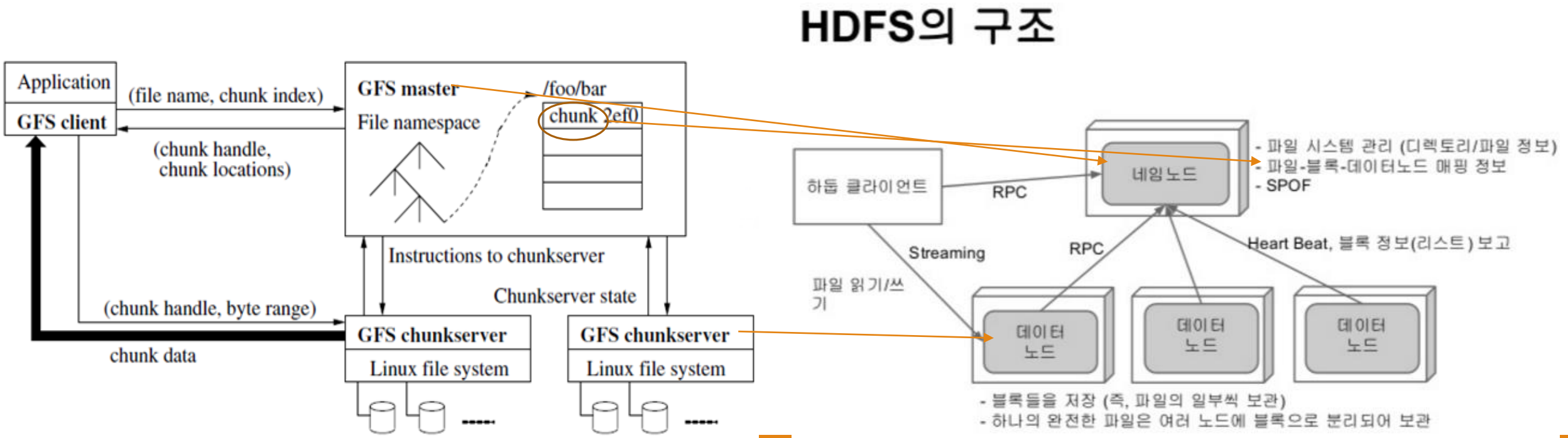
- 저가형 서버로 구성된 환경에서 고장이 빈번히 발생을 가정, 파일은 대용량이라고 가정
- 주로 연속적으로 많은 데이터를 읽는 연산, 순차적 데이터 추가
- 낮은 응답 지연 시간보다 높은 처리율이 보다 중요



제 1절 분산 데이터 저장 기술

나. 하둡 분산 파일 시스템

- 구글 파일 시스템(오픈소스 X)과 아키텍쳐가 동일, 오픈소스임
- 하나의 네임노드(NameNode)와 다수의 데이터노드(DataNode) 구성



제 1절 분산 데이터 저장 기술

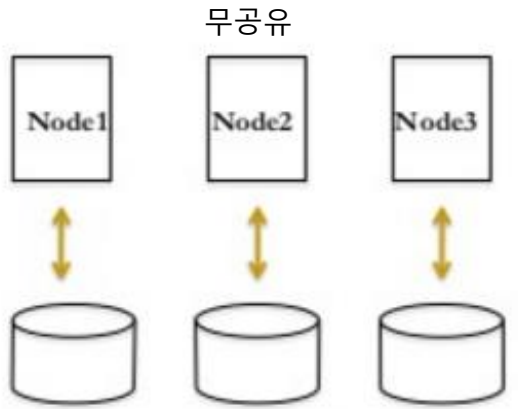
다. 러스터

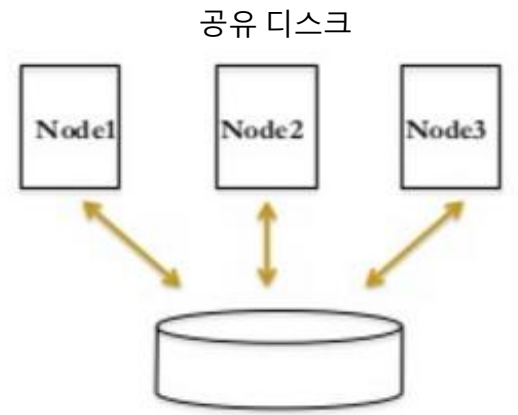
- 러스터(Lustre)는 병렬 분산 파일 시스템으로서 주로 고성능 컴퓨팅의 대용량 파일 시스템으로 사용되고 있다. 러스터(Lustre)의 이름은 Linux와 cluster의 혼성어이다.
- 러스터는 GNU GPL 정책의 일환으로 개방되어 있으며 소규모 클러스터 시스템부터 대규모 클러스터 시스템용 고성능 파일 시스템이다.
- 높은 성능과 코드가 개방되었기 때문에 슈퍼컴퓨터에 자주 사용된다.
- 대한민국에서는 한국과학기술정보연구원과 대한민국 기상청이 러스터 파일 시스템 기반 스토리지를 사용하고 있다.

제 1절 분산 데이터 저장 기술

2. 데이터베이스 클러스터

- 데이터를 통합할때, 성능향상과 가용성을 높이기 위해서 파티셔닝 또는 클러스터링을 이용
- 데이터베이스 파티셔닝의 장점
  - 파티션 사이의 병렬 처리를 통한 빠른 데이터 검색 및 처리 성능
  - 성능의 선형적인 증가 효과와 특정 파티션이 장애가 발생해도 서비스가 가능한 고가용성을 확보
- 데이터베이스 클러스터링 구현 방법
  1) 무공유(Shared Nothing)
     - 장점 : 노드 확장 제한이 없음.
     - 단점 : 장애를 대비해서 별도의 Fault-tolerance 구성 필요
     - 대부분의 데이터베이스 클러스터에서 채택
  2) 공유 디스크(Shared Disk)
     - 장점 : 높은 수준의 Fault-tolerance
     - 단점 : 클러스터가 커지면 디스크 영역에서 병목
     - Oracle RAC에서 채택





제 1절 분산 데이터 저장 기술

3. NoSQL

NoSQL은 Key와 Value와 같은 단순한 자료 저장, 아주 빠른 조회와 저장이 가능함.

Join연산기능이 지원되지 않지만, 대용량과 확장성을 제공

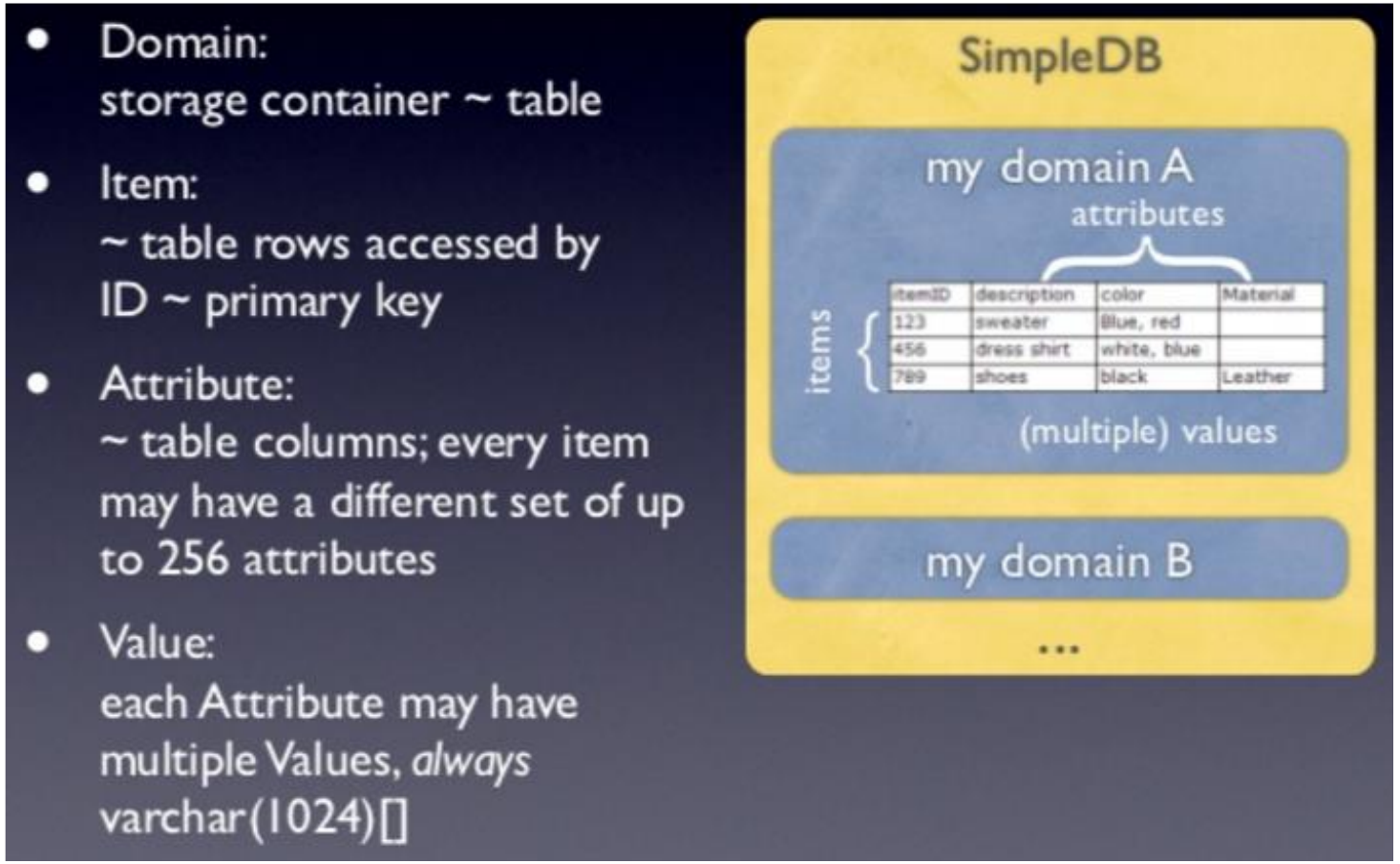
가. 구글 빅데이블과 HBASE

	Column Family		
rowkey	cf1	cf2	cf3
prefix-001	col00 -> value col01 -> value	cola -> value colb -> value	coli -> value colii -> value
prefix-002	col02 -> value col03 -> value	colc -> value cold -> value	coliii -> value coliv -> value

제 1절 분산 데이터 저장 기술

3. NoSQL

나. 아마존 SimpleDB



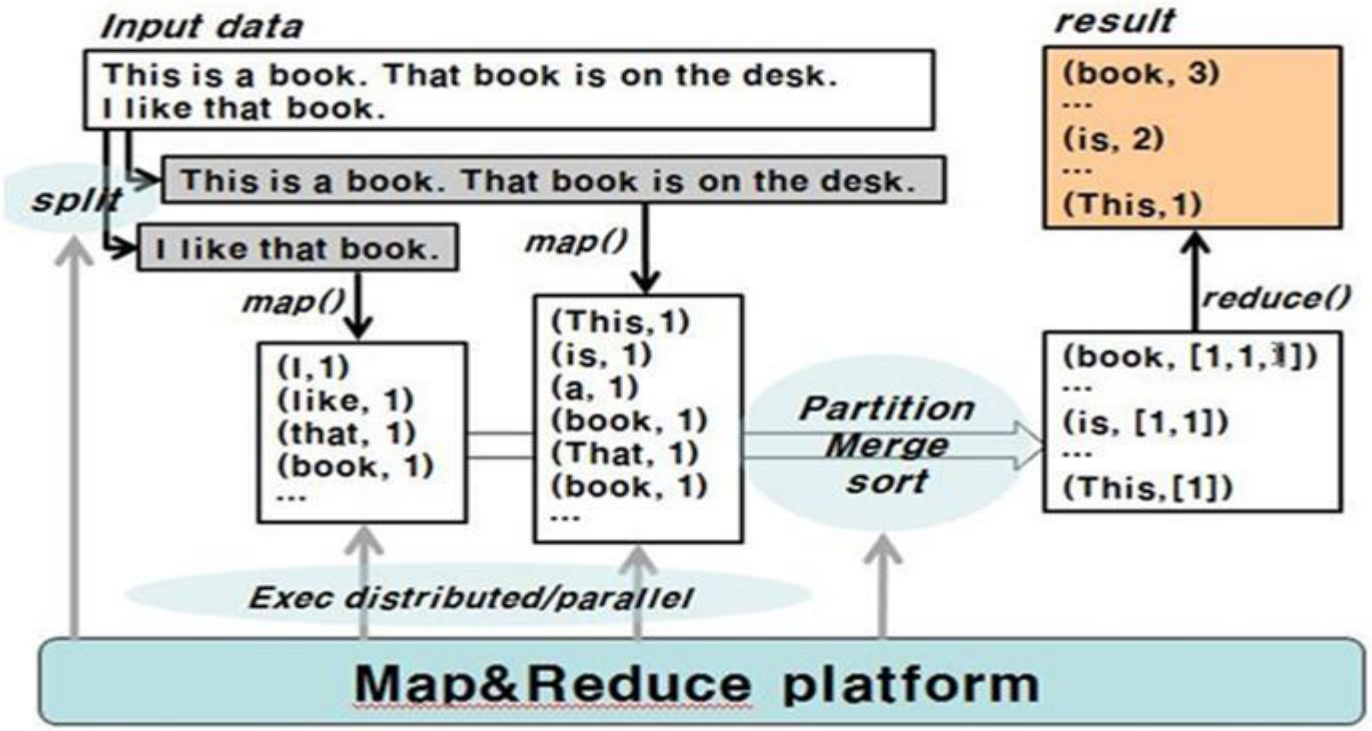
제 2절 분산 컴퓨팅 기술

1. MapReduce

가. 구글 MapReduce

나. Hadoop MapReduce

Map(k1, v1) -> list(k2, v2)
Reduce(k2, list(v2)) -> list(v2)



제 2절 분산 컴퓨팅 기술

2. 병렬 쿼리 시스템

가. 구글 Swazall

나. 아파치 Pig(스크립트)

다. 아파치 Hive(SQL)

제 2절 분산 컴퓨팅 기술

3. SQL on Hadoop

하둡의 MapReduce와 Hive는 대용량 데이터를 배치 처리에 최적화

실시간 처리를 위해서 “SQL on Hadoop” 이라는 실시간 SQL 질의 분석 기술 주목됨.

대표적인 예) Impala , Tajo 등

제 3절 클라우드 인프라 기술

- 클라우드 컴퓨팅은 동적으로 확장할 수 있는 가상화 지원들을 인터넷으로 서비스하는 기술
- 3가지 유형 : SaaS, PaaS, IaaS
- 클라우드 컴퓨팅의 근간이 되는 인프라 기술은 서버 가상화 기술임.
  - 상용솔루션 : vmware, MS Hyper-V
  - 오픈소스 : KVM, Xen
- 서버 가상화의 효과
  - 가상머신 사이의 데이터 보호 => 공공기관, 금융회사, 의료 기관에서 많이 도입.
  - 예측하지 못한 장애로부터 보호
  - 공유 자원에 대한 강제 사용의 거부
  - 서버 통합
  - 자원 할당에 대한 증가된 유연성
  - 시스템 관리

제 3절 클라우드 인프라 기술

가. CPU 가상화

- 물리적 서버 위에 가상화 레이어를 통해 운영체제가 수행하는 필요한 하드웨어 환경을 가상으로 만들어 것을 하이퍼바이저(Hypervisor)라 함.
- 하이퍼바이저 분류
  - 운영체제 수정 여부
    - 완전가상화(수정X , OS와 독립적), 하드웨어 지원 완전가상화 (수정X) , 반가상화(수정O , OS와 의존적)
  - 하드웨어 드라이버의 계층에 따라
    - Monolithic(Hypervisor에 포함), Microkernel(가상머신에 포함)
  - 하드웨어 가상화 여부
    - 호스트 기반 가상화, 컨테이너 기반 가상화(docker)

제 3절 클라우드 인프라 기술

나. 메모리 가상화

다. I/O 가상화

- 가상 이더넷
  - 물리적으로 존재하지 않는 자원을 만들어 내는 에뮬레이션 기능
- 공유 이더넷 어댑터
  - 여러 개의 가상머신이 물리적인 네트워크 카드를 공유하게 하는 기술
- 가상 디스크 어댑터
  - 내장 디스크 또는 외장 디스크를 가상머신에 할당하는 기술