

하둡(Hadoop)

1-1 HDFS(Hadoop Distribute File System)

File System

파일 시스템은 컴퓨터에서 파일이나 자료를 쉽게 발견 및 접근할 수 있도록 **보관 또는 조직하는 체제**를 가르키는 말이다.

(위키백과)

리눅스

(ext, ext2, ext3, ext4, ZFS,..)

맥 OS X : HFS 플러스

윈도우 : FAT, NTFS

1-1 HDFS(Hadoop Distribute File System)

분산 파일 시스템(DFS)

컴퓨터 네트워크를 통해 공유하는
여러 호스트 컴퓨터 파일에
접근할 수 있게 하는
파일 시스템이다.

1-1 HDFS(Hadoop Distribute File System)

데이터에 대한 어떤 고민...?

- (1) 좀 더 **경제적**으로 데이터를 저장할 수 있을까?
- (2) 실시간으로 **데이터를 저장**할 수 있을까? 정형, 비정형 데이터..
- (3) 데이터가 많아지면 **쉽게 간편하게 확장**할 수 있을까?
- (4) 데이터가 많은데, 몇 대의 PC로는 저장하기가 어려운데,
여러 대로 나누어 많은 용량을 저장하게 할 수 있을까?
- (5) 만약 한곳에 데이터를 저장했는데, 한 곳의 데이터가
없어지면 어떻게 해야 하나? 여러 곳에 데이터를 분산시켜야 하지 않나?

1-1 HDFS(Hadoop Distribute File System)

그러면 어떻게 해야 하지?

1-1 HDFS(Hadoop Distribute File System)

- (1) 좀 더 싼 비용으로 여러 대의 컴퓨터를 사용해 보면 어떨까?
- (2) 그리고 어떤 처리를 분산시켜서 동시에 처리하게 하면 어떨까?
- (3) 정형화된 데이터 형태가 아닌 (키, 값) 형태로 만들어서 저장해 보자.
- (4) 데이터를 나눠서 놓다. 만약의 사태를 위해 분산시켜 두자.

1-1 HDFS(Hadoop Distribute File System)

그러면 뭐부터 시작해 볼까?

(1) 자 그럼 먼저 저장시킬 저장소를 준비해 보자.

1-1 HDFS(Hadoop Distribute File System)

그러면 뭐부터 시작해 볼까?

(1) 자 그럼 먼저 저장시킬 저장소를 준비해 보자.

◆ HDFS 정의

HDFS는 Hadoop 어플리케이션에 의해 사용되는 기본 분산 저장장소이다. (비정형 데이터)

1-1 HDFS(Hadoop Distribute File System)

◆ HDFS 정의

HDFS는 Hadoop 어플리케이션에 의해 사용되는 기본 분산 저장장소이다.



자 이제 여러 군데의 저장장소를 (PC 또는 물리적 분리된 공간) 준비 OK

1-1 HDFS(Hadoop Distribute File System)

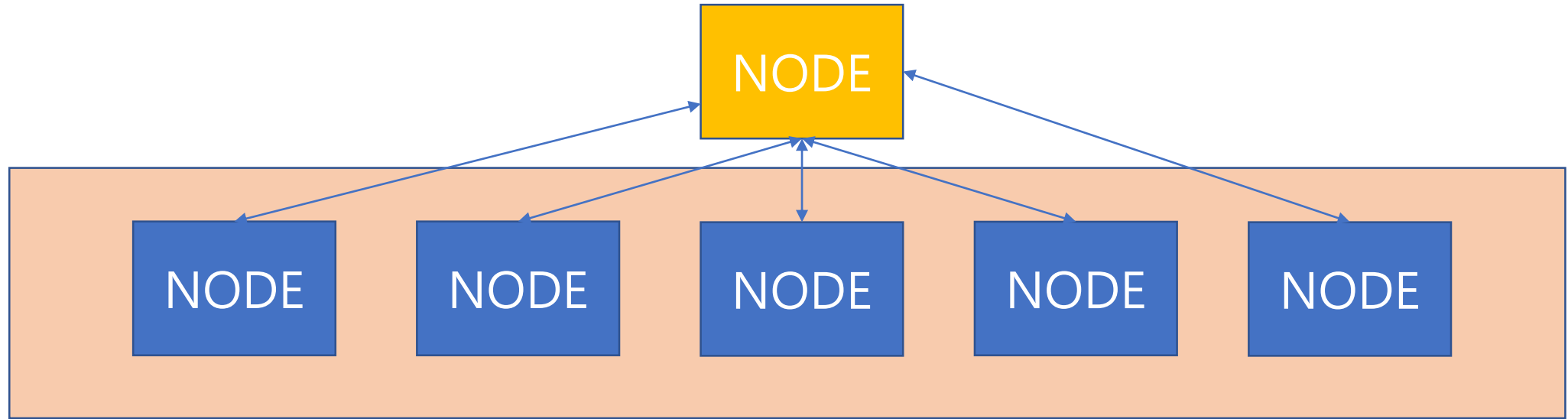
◆ HDFS 정의

HDFS는 Hadoop 어플리케이션에 의해 사용되는 기본 분산 저장장소이다.



하나의 NODE를 전체 친구들을 관리하는 친구들로 하자. 이를 우리는 **마스터 노드**, 마스터 노드에는 친구들의 이름과 특징을 기억하고 있어서, 우리는 **namenode**라 한다.

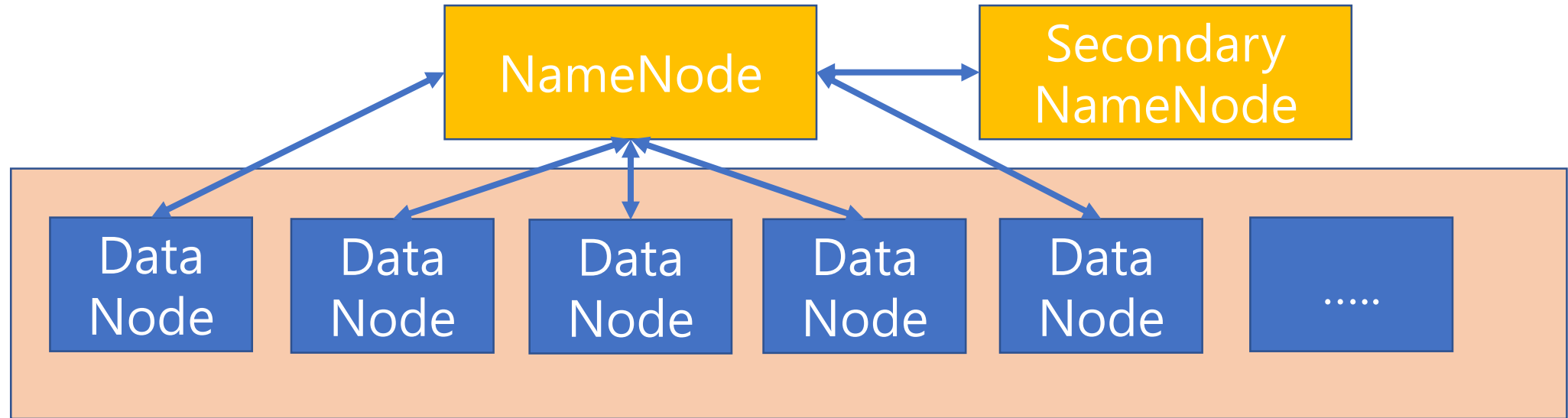
1-1 HDFS(Hadoop Distribute File System)



하나의 NODE를 전체 친구들을 관리하는 친구들로 하자. 이를 우리는 마스터 노드, 즉 **네임노드(namenode)**라 한다.

다른 친구들은 슬레이브(slave) 노드이다. 여기에는 데이터가 저장되기 때문에 **데이터 노드(data node)**라고 한다.

1-1 HDFS(Hadoop Distribute File System)



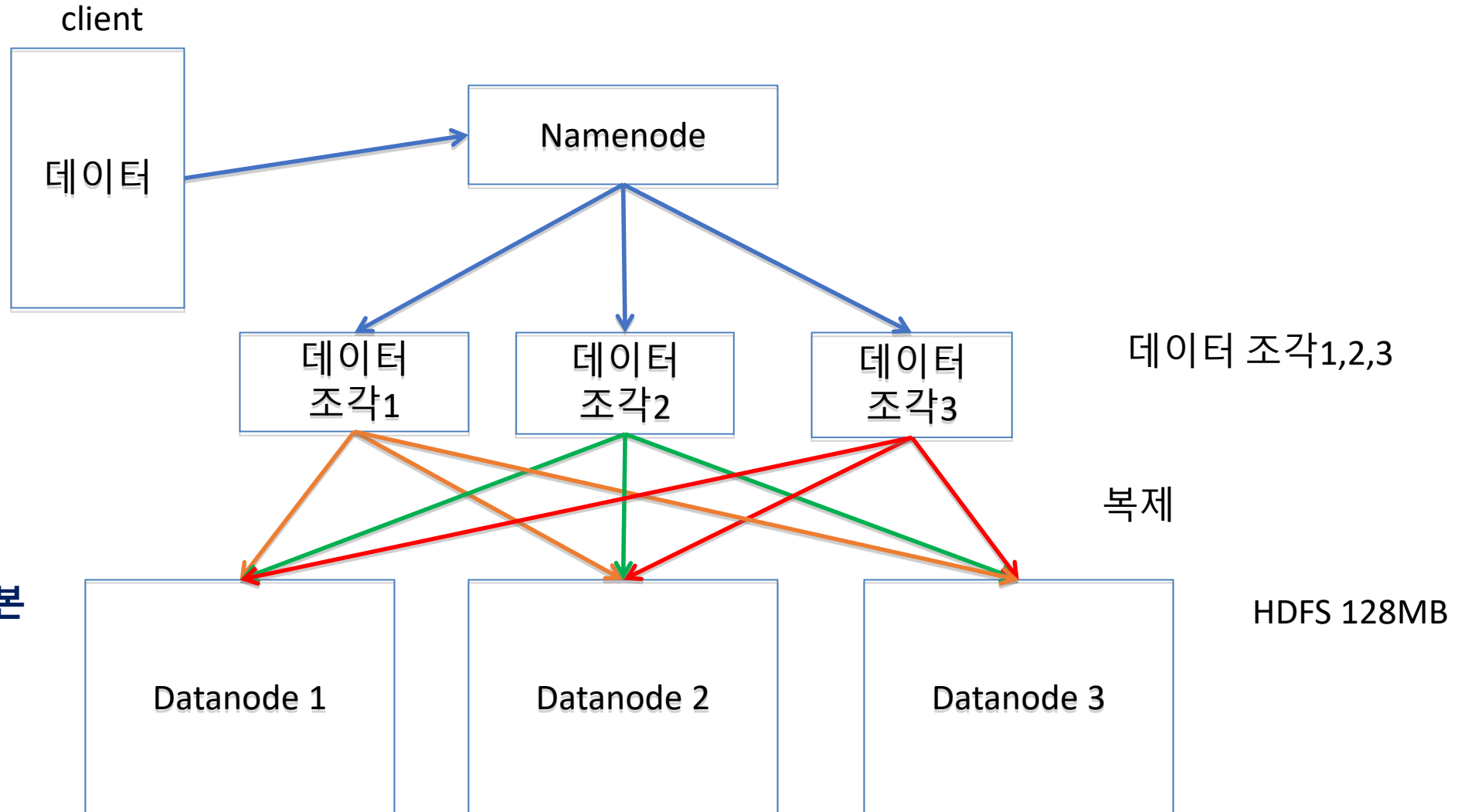
만약 네임노드(NameNode)가 문제가 생기면 이를 대체해 줄 친구가 필요해, 그렇다면 Secondary NameNode를 준비해 두자.

1-1 HDFS(Hadoop Distribute File System)

◆ 분산환경에서 데이터 처리

분산 환경에서는 어떻게 데이터를 처리해야 할까?

1-1 HDFS(Hadoop Distribute File System)



기본 한 블록당 3개 복사본

하나 : 자신위치

둘 : 같은 랙의 다른노드

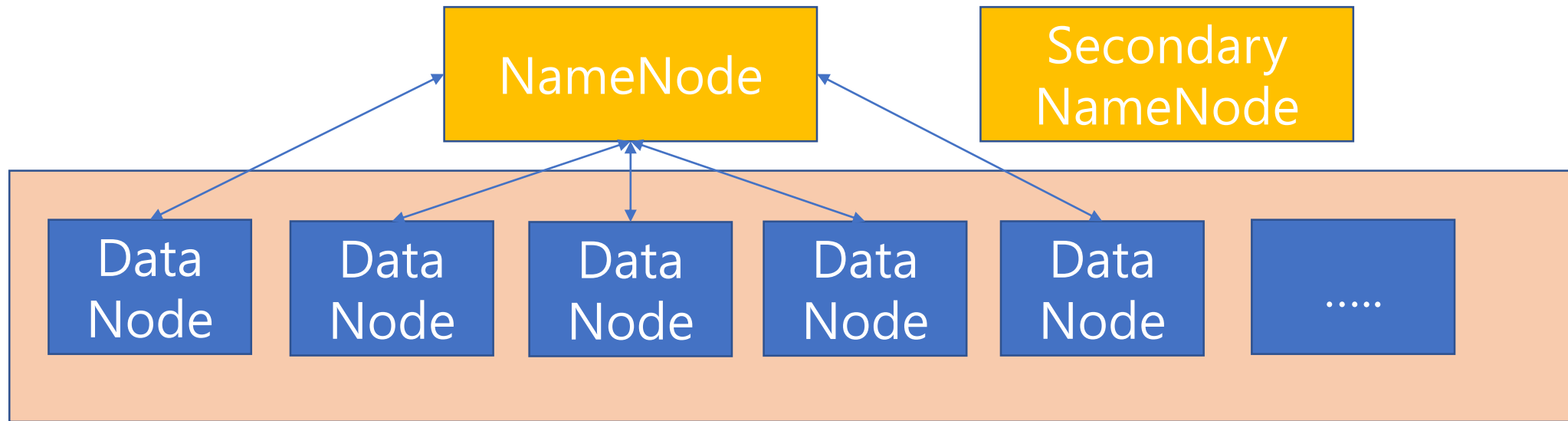
셋 : 다른 랙의 다른 노드

1-1 HDFS(Hadoop Distribute File System)

◆ Summary

- 신뢰성과 내고장성(Fault tolerant)
- 큰 데이터를 처리 가능함
- 마스터(Master) 슬레이브(Slave) 구조
- 한 번의 파일 접근 쓰기만 가능

1-2 MAPREDUCE

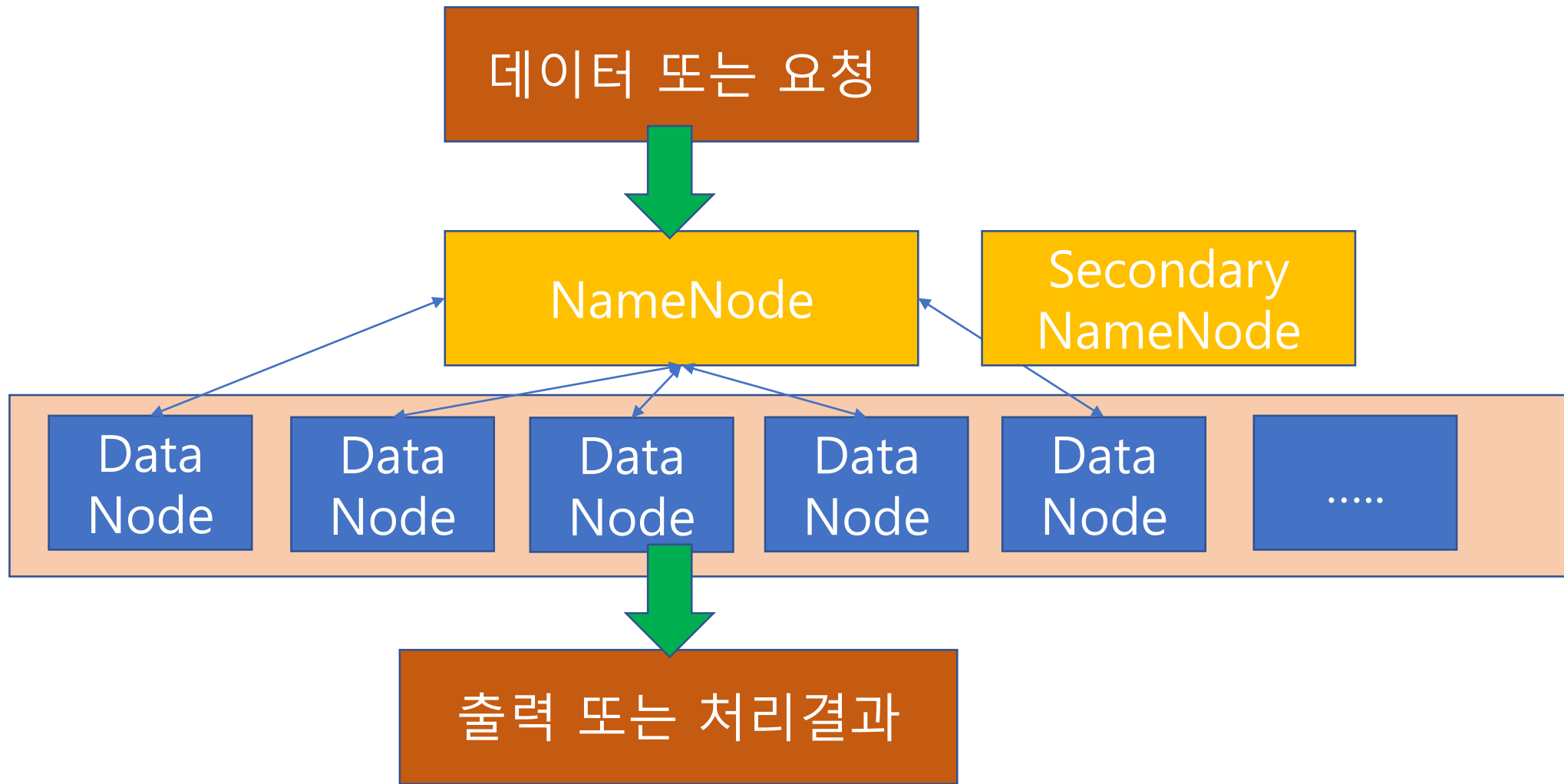


자, 그럼 우선 어느정도 형태를 갖췄다.

시스템 요소는 준비 완료!

이제 뭘 해야 하나?

1-2 MAPREDUCE



데이터에 대해서 어떤 요청을 처리해 줄 수 있는 무언가의 프로그래밍이
필요해..

1-2 MAPREDUCE

◆ MAPREDUCE를 이해하기

- 시스템에 구체적인 처리하는 **동작**을 넣어보자. **프로그래밍 기법**을 담아보자.

1-2 MAPREDUCE

◆ MAPREDUCE 정의

-> MapReduce는 클러스터에서 병렬 분산 알고리즘을 사용하여 빅데이터 세트를 처리하고 생성하기 위한 프로그래밍 모델 및 관련 구현이다.

클러스터(Cluster) : 여러대의 컴퓨터들이 연결되어 하나의 시스템처럼 동작하는 컴퓨터의 집합

알고리즘 : 문제를 푸는 방법

1-2 MAPREDUCE

◆ MAPREDUCE 시스템

-> 방대한 양의 데이터를 처리하는 작업을 작성하기 위한 **소프트웨어 프레임워크**이다.

-> 분산 방식으로 동작하는 **프로그래밍 모델**이다.

MapReduce외에도 있다. 기타 모델도 있다.(MPI, BSP)

-> **분산 처리를 한다는 것은 여러 대의 컴퓨터가 동시 처리를 한다는 것을 의미한다. → 성능 향상**

1-2 MAPREDUCE

◆ 처리 절차

- (1) 시스템은 **입력 데이터를 처리**하고,
- (2) **데이터**를 컴퓨터 네트워크상에 **분산해 병렬로 처리**한다.
- (3) 최종적으로 **출력 결과**들을 차후에 **취합할 하나의 파일로 결합**한다.

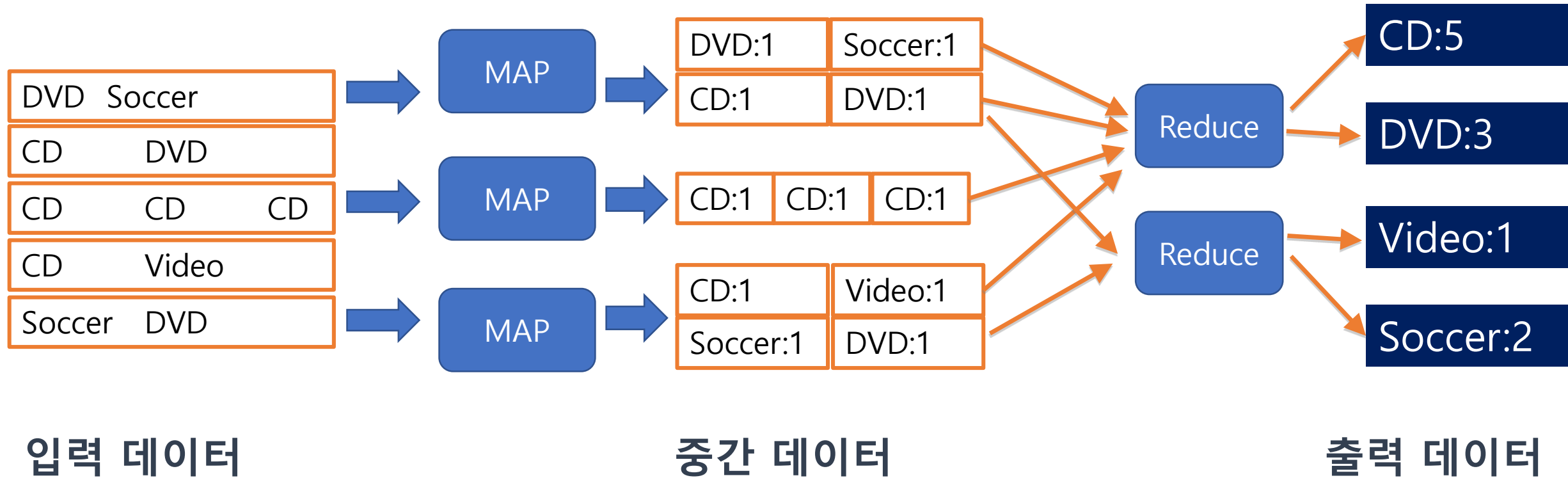
1-2 MAPREDUCE

◆ 2개의 과정으로 한다면...

- (1) 키-값 쌍의 데이터를 처리하는 맵(MAP)
- (2) 리듀스(Reduce)

1-2 MAPREDUCE

◆ MapReduce



(1) Mapper는 한 라인씩 읽어들이는 작업

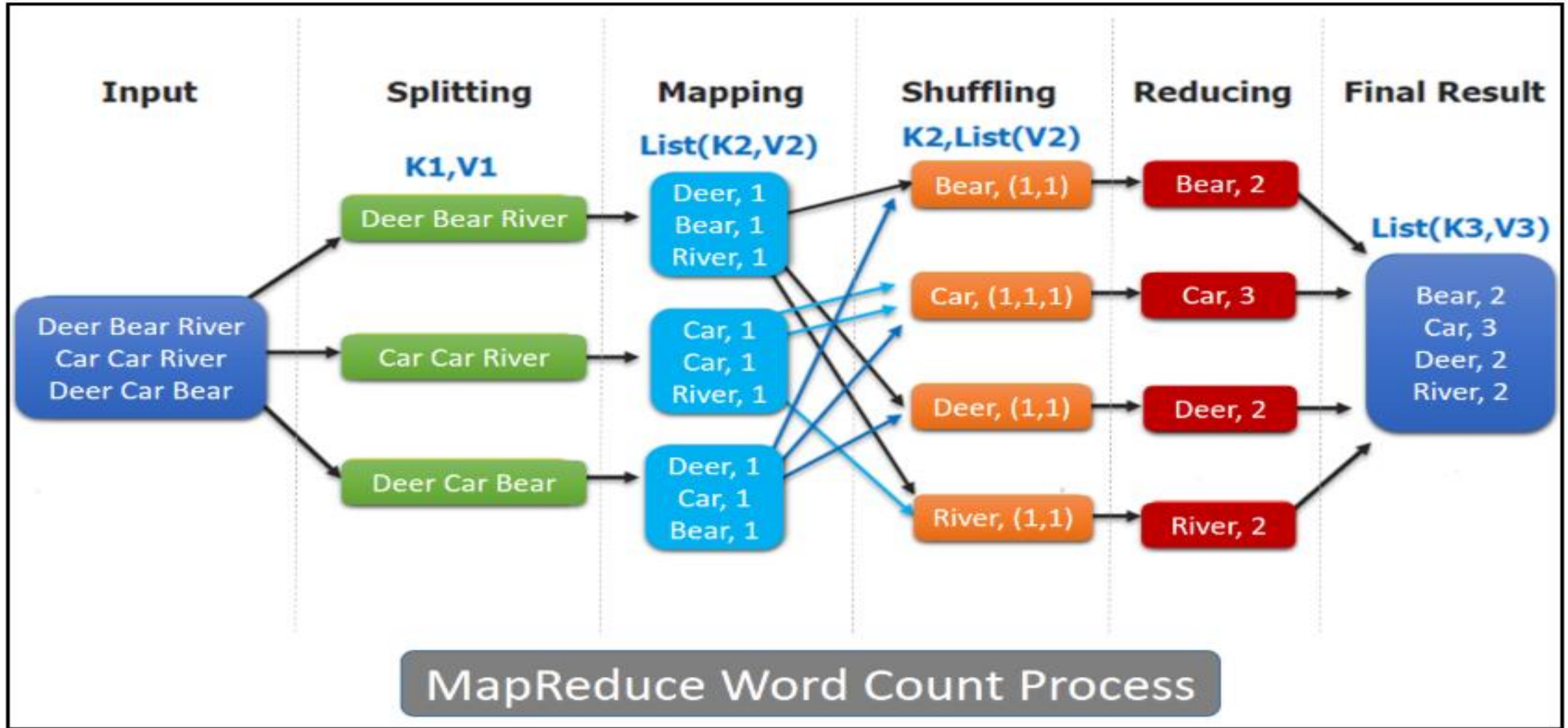
(2) Mapper의 출력을 모아서 단어의 개수를 합산하는 작업을 Reducer가 수행.

1-2 MAPREDUCE

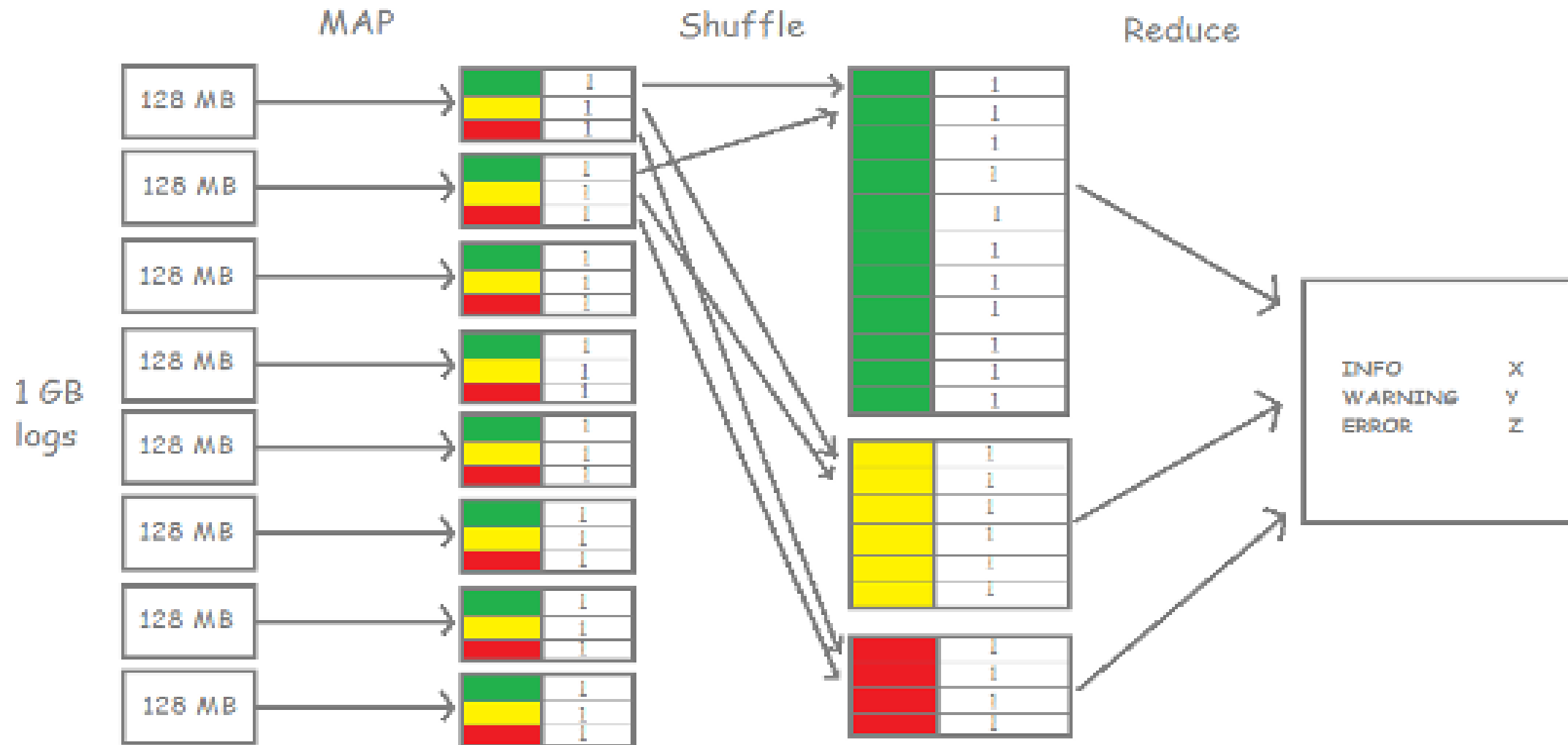
◆ 3개의 과정으로 한다면...

- (1) Map : 각 작업자 노드는 로컬 데이터에 맵 기능을 적용한다. 키-값 쌍의 데이터를 처리하는 맵(MAP)
- (2) Shuffle(셔플) : 작업자 노드는 출력키를 기반으로 재배포한다. 한 키에 속한 모든 데이터가 동일한 작업자 노드에 있다.
- (2) 리듀스(Reduce) : 작업자 노드가 키 별로 각 출력 데이터 그룹을 병렬로 처리

1-2 MAPREDUCE

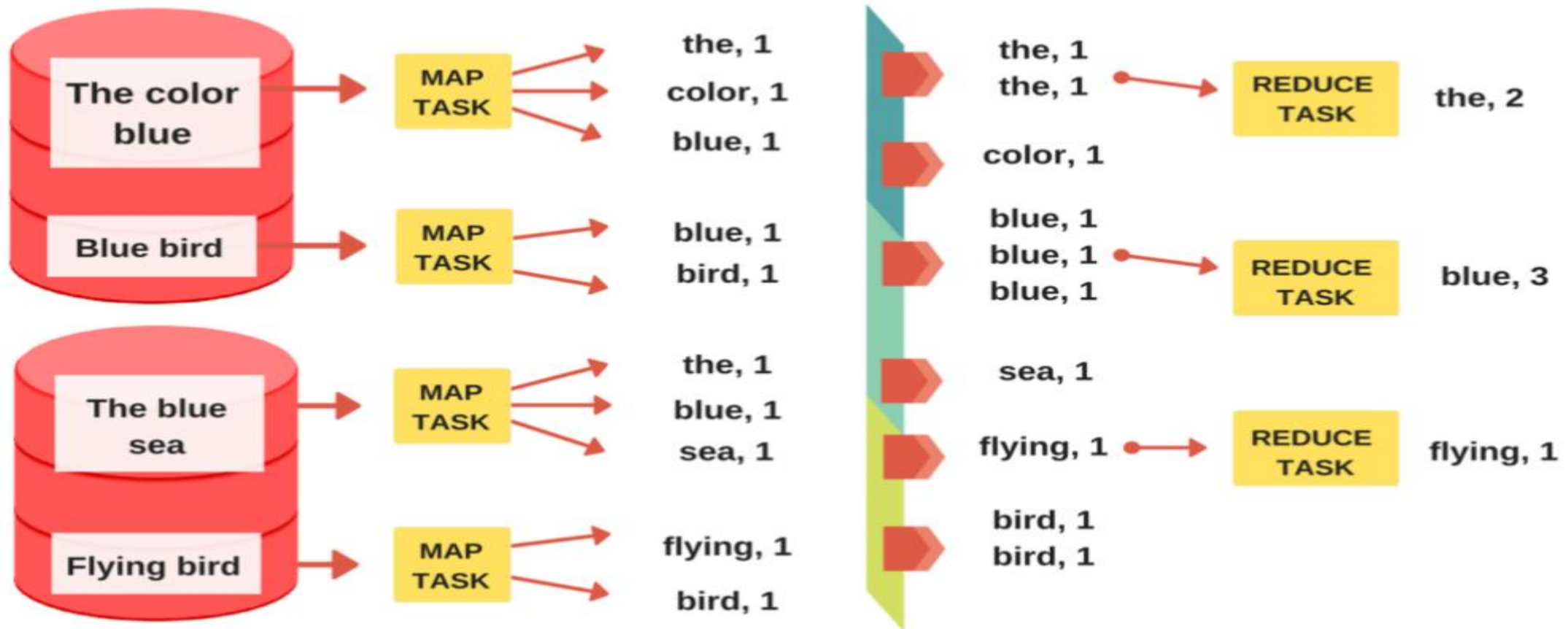


1-2 MAPREDUCE



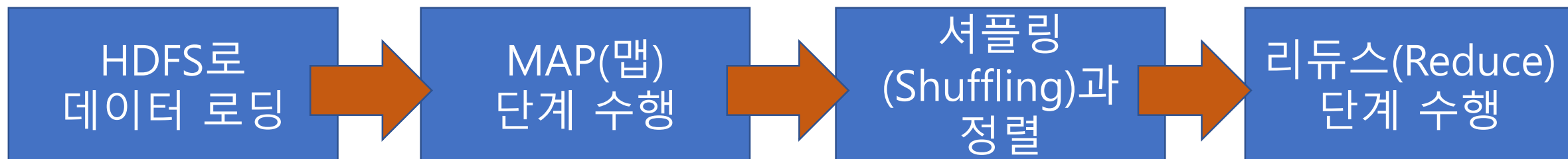
1-2 MAPREDUCE

SORT and SHUFFLE



1-2 MAPREDUCE

◆ 하둡 맵리듀스 데이터를 처리하는 4단계



1-2 MAPREDUCE

◆ 맵 리듀스의 구조(Hadoop 1.0 버전)

- (1) 전형적인 맵리듀스는 Job 제출, Job 초기화, Task 할당, Task 실행, 진행 상황 갱신, 잡 완료 연관활동 포함한다.
- (2) 잡 트래커에 의해 관리.
- (3) 태스트 트래커에 의해 실행.

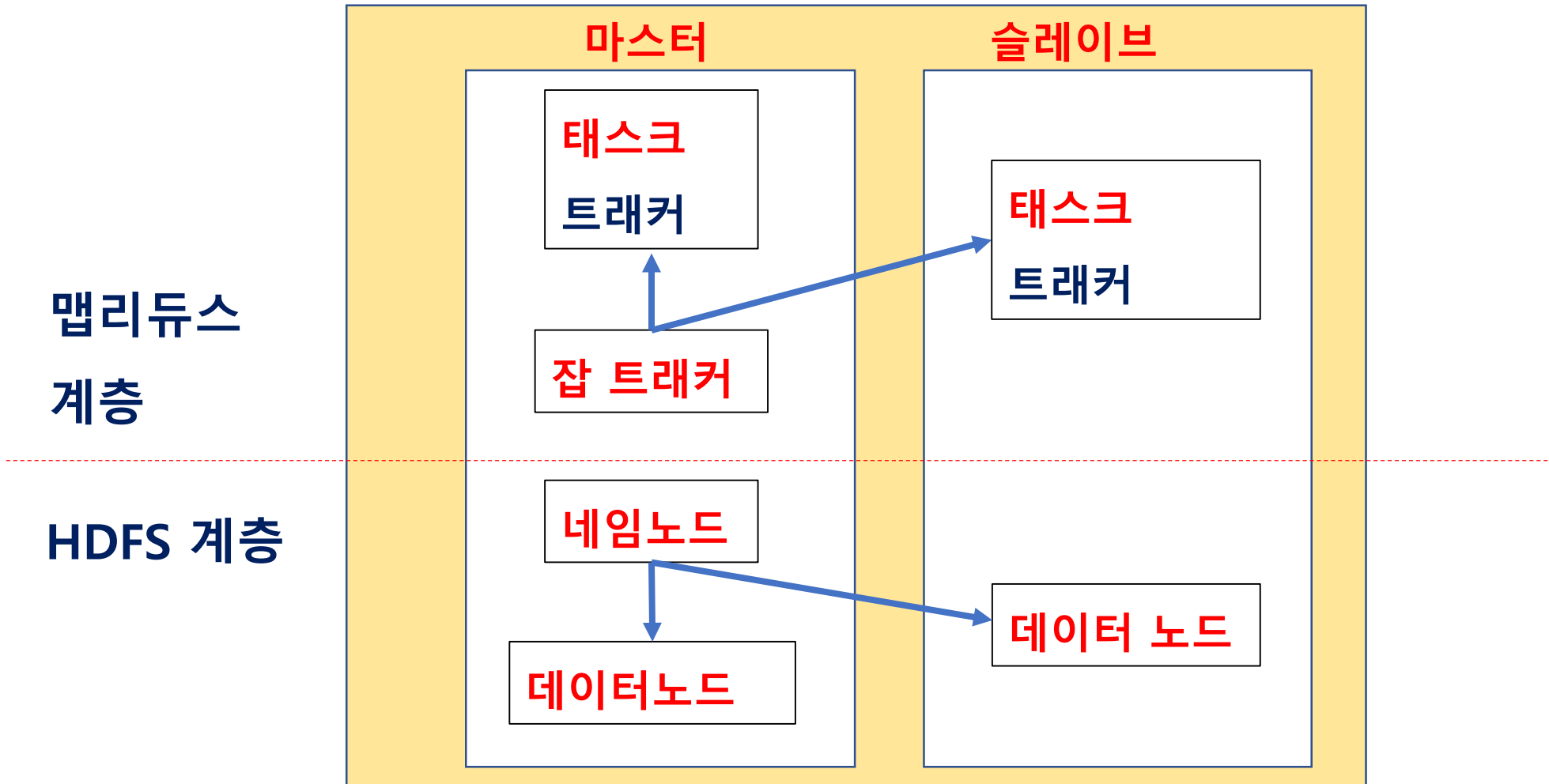
1-2 MAPREDUCE

◆ 하둡 맵리듀스 개체 나열(Hadoop 1.0)

- (1) 클라이언트(Client)
- (2) 잡 트래커(job tracker)
- (3) 태스크 트래커(Task tracker)
- (4) HDFS : 입력과 출력 데이터를 저장

1-2 MAPREDUCE

◆ MapReduce, HDFS



1-3 HIVE

◆ HIVE



Painted wooden beehives with active honey bees



◆ HIVE

점점 기존의 RDBMS를 하둡(Hadoop)이 대체하고 있다.

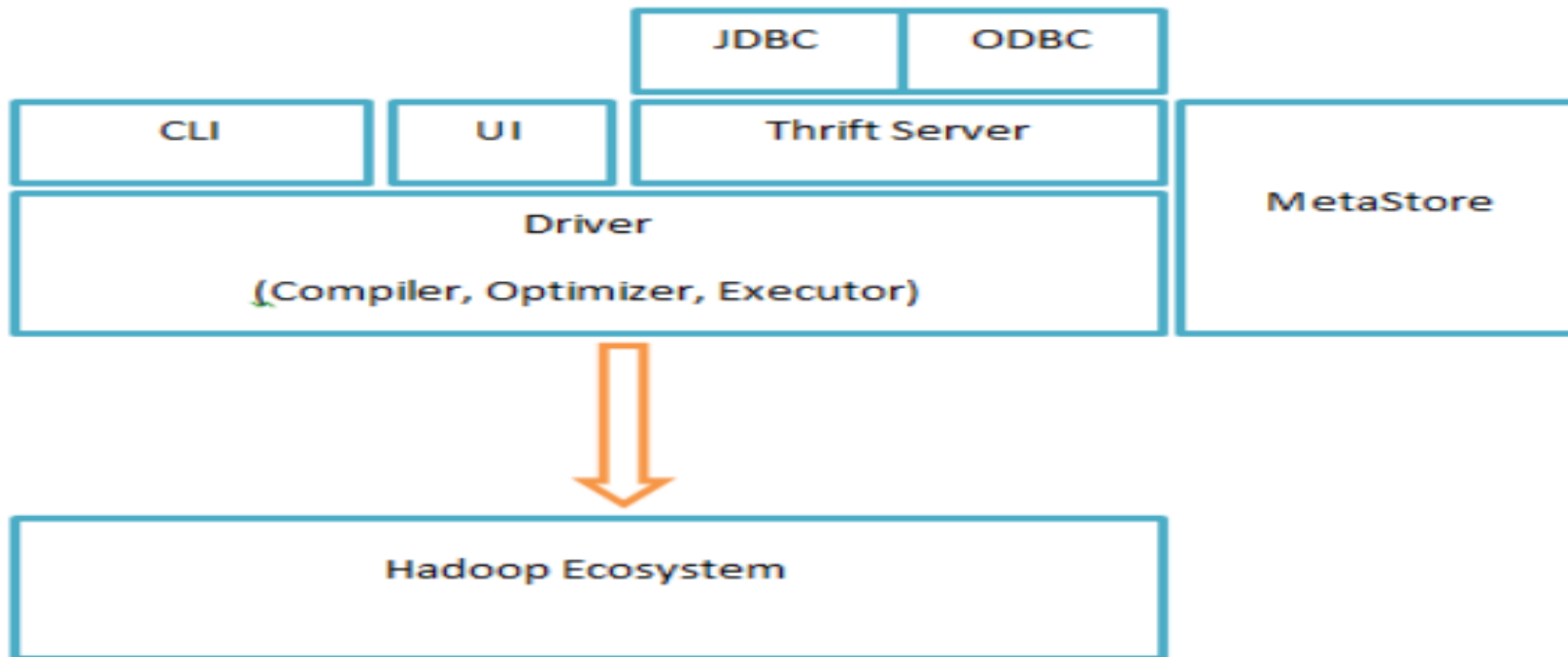
그렇다면 우리가 기존에 쓰던 **SQL문**을 Hadoop에 사용할 수 있을까?

◆ HIVE

- ◆ Apache Hadoop위에 구축된 데이터 웨어 하우스 소프트웨어 프로젝트
- ◆ Hadoop과 통합되는 다양한 데이터 베이스 및 파일 시스템에 저장된 데이터를 쿼리(Query)하는 SQL과 유사한 인터페이스 제공.
- ◆ FaceBook에서 처음 개발, 넷플릭스(Netflix), 금융 산업 규제 당국(FINRA)에서 사용하고 개발.

1-3 HIVE

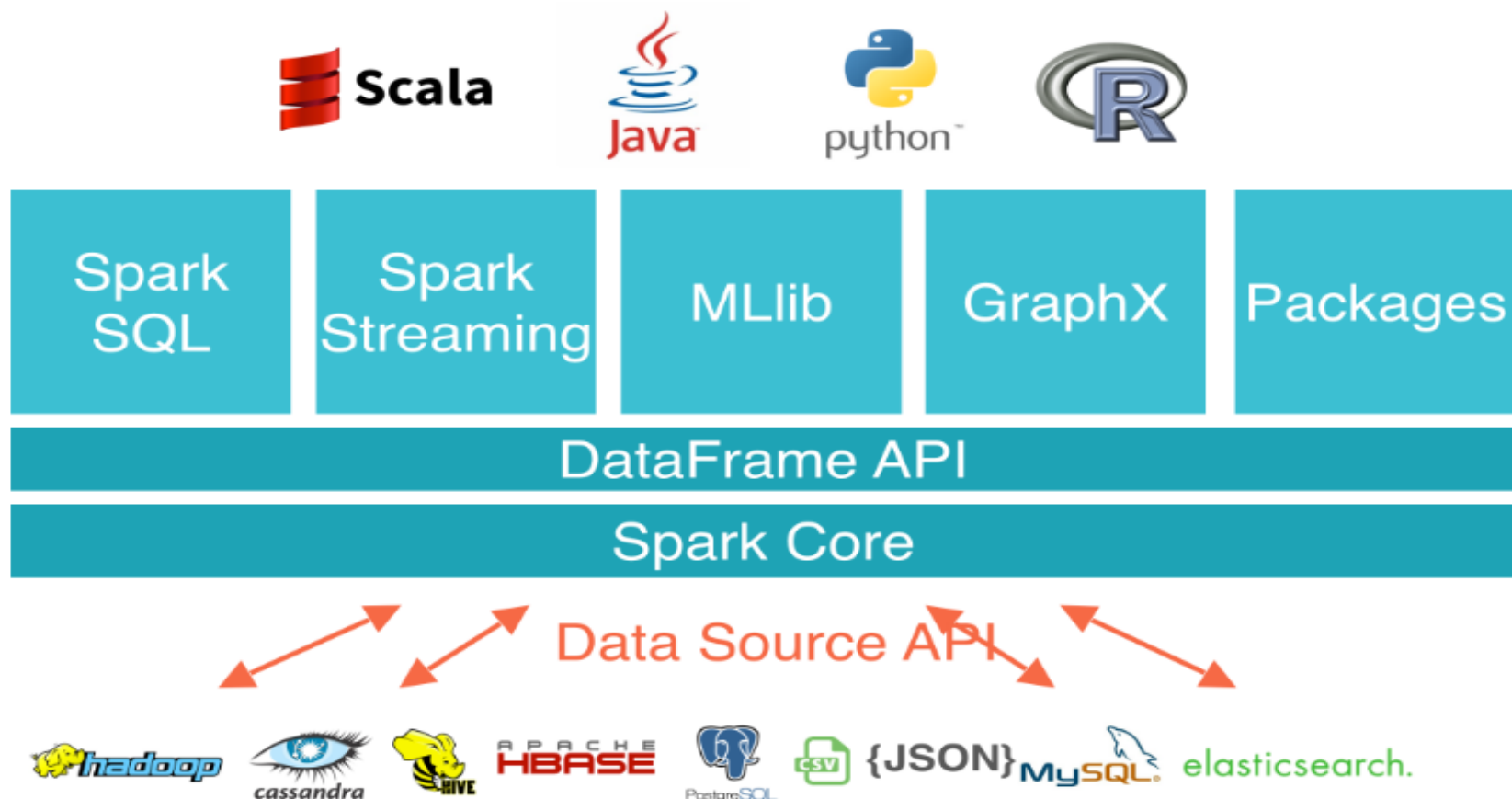
◆ HIVE 구조



MetaStore : Hive tables와 RDBMS의 분할 metadata를 저장하는 서비스

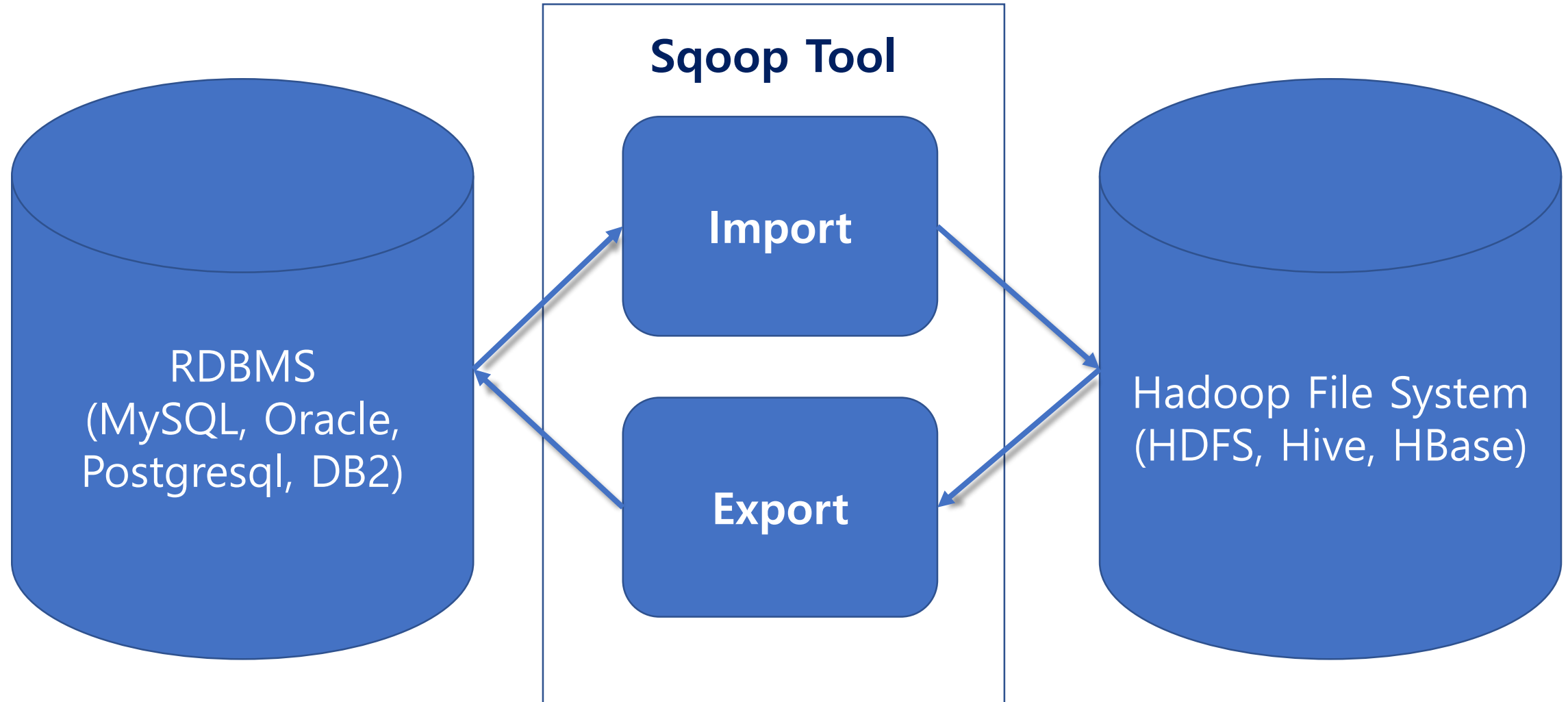
Thrift Server : HiveServer는 원격 클라이언트가 다양한 프로그래밍 언어를 사용하여 요청을 제출하고 결과를 검색할 수 있게 해주는 서비스입니다.

1-4 Spark



 databricks

1-5 Sqoop



1-6 RHadoop

RHadoop

RHadoop

RHadoop

RHadoop

REFERENCE

◆ MAPREDUCE

<http://bitly.kr/2F7fk> 참조

<https://datawhatnow.com/batch-processing-mapreduce/> 참조