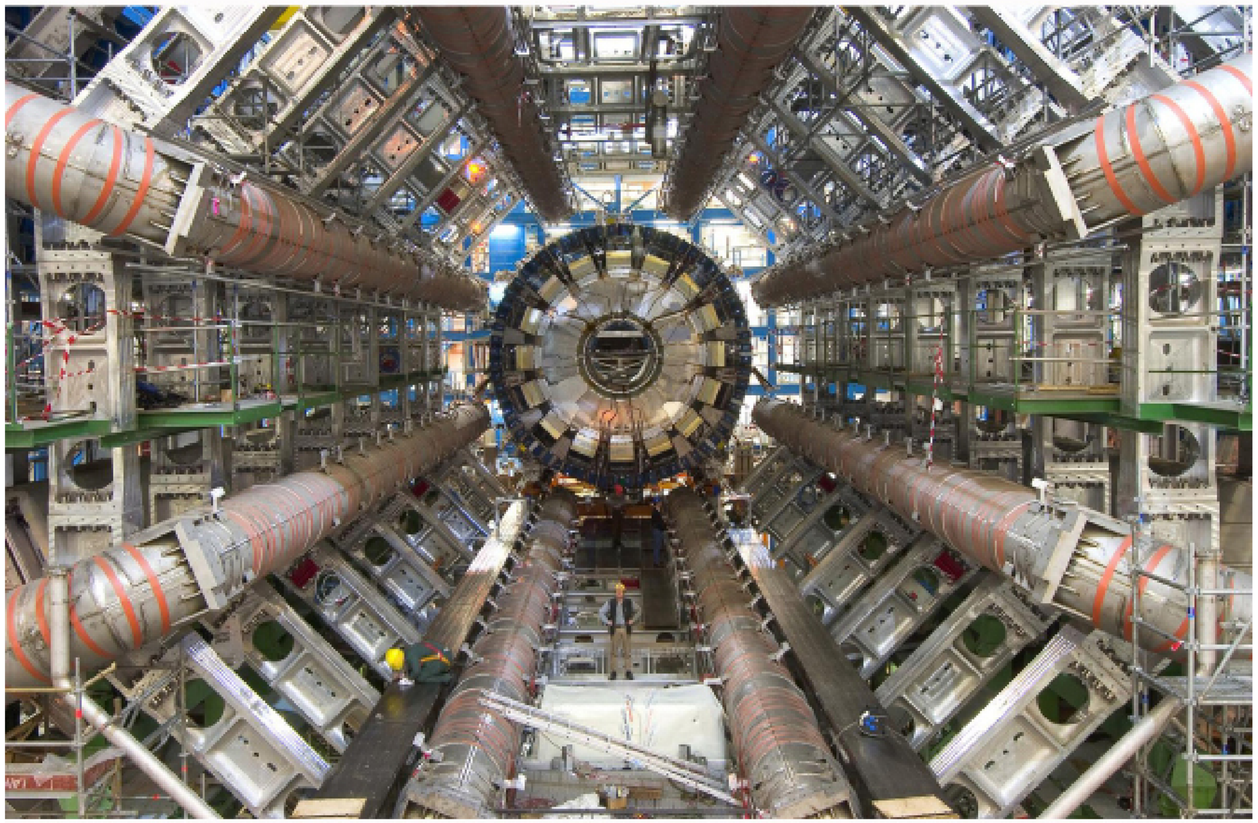
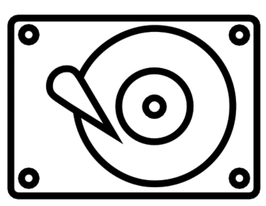
Yandex

HDFS

Namenode Architecture



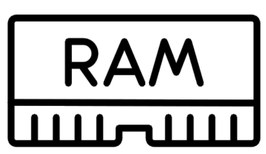
1 year ~ 10 PB



replica – 3개

10 PB / 2 TB * 3 ~ 15 k

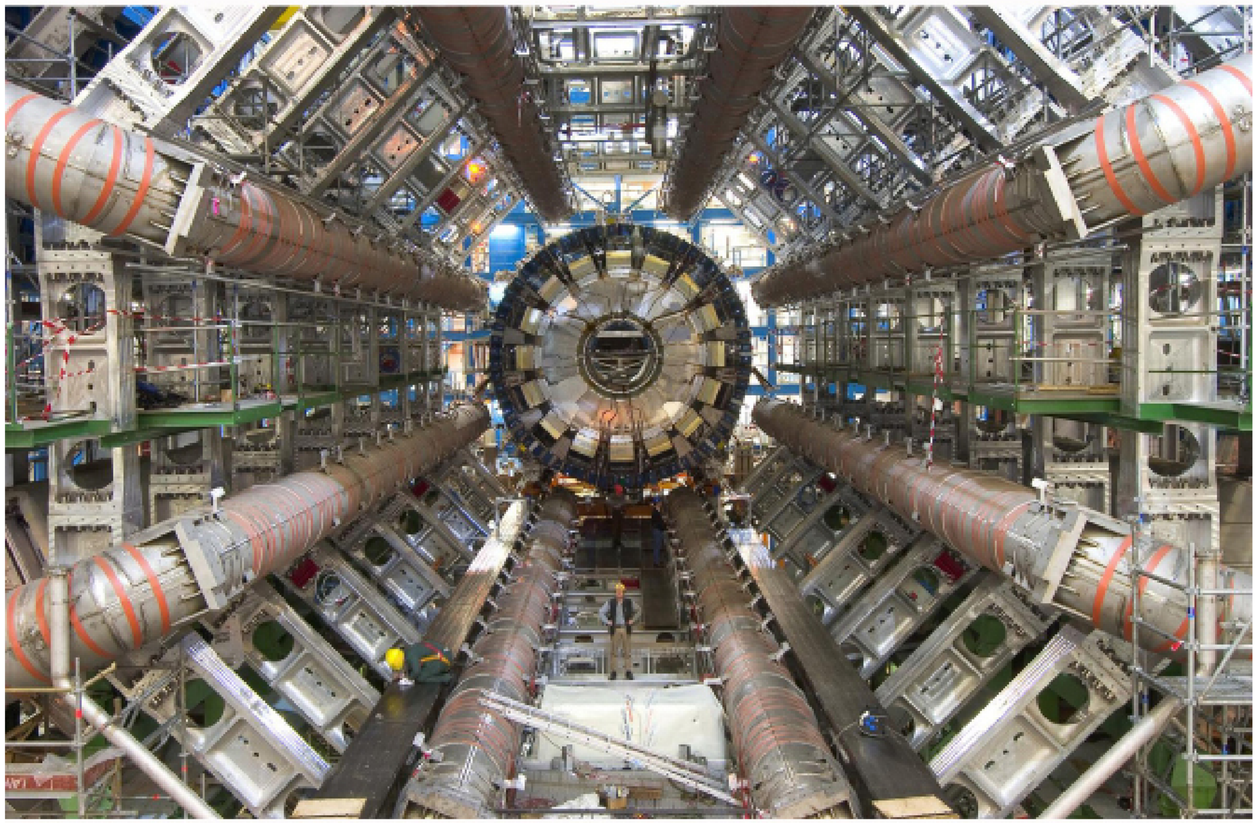
10PB 데이터를 저장하기 위해 15k개의 2TB 드라이브가 필요



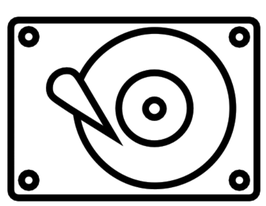
150 B - average block size on Namenode

일반적인 블록의 크기

https://issues.apache.org/jira/browse/HADOOP-1687

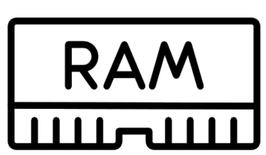


1 year ~ 10 PB



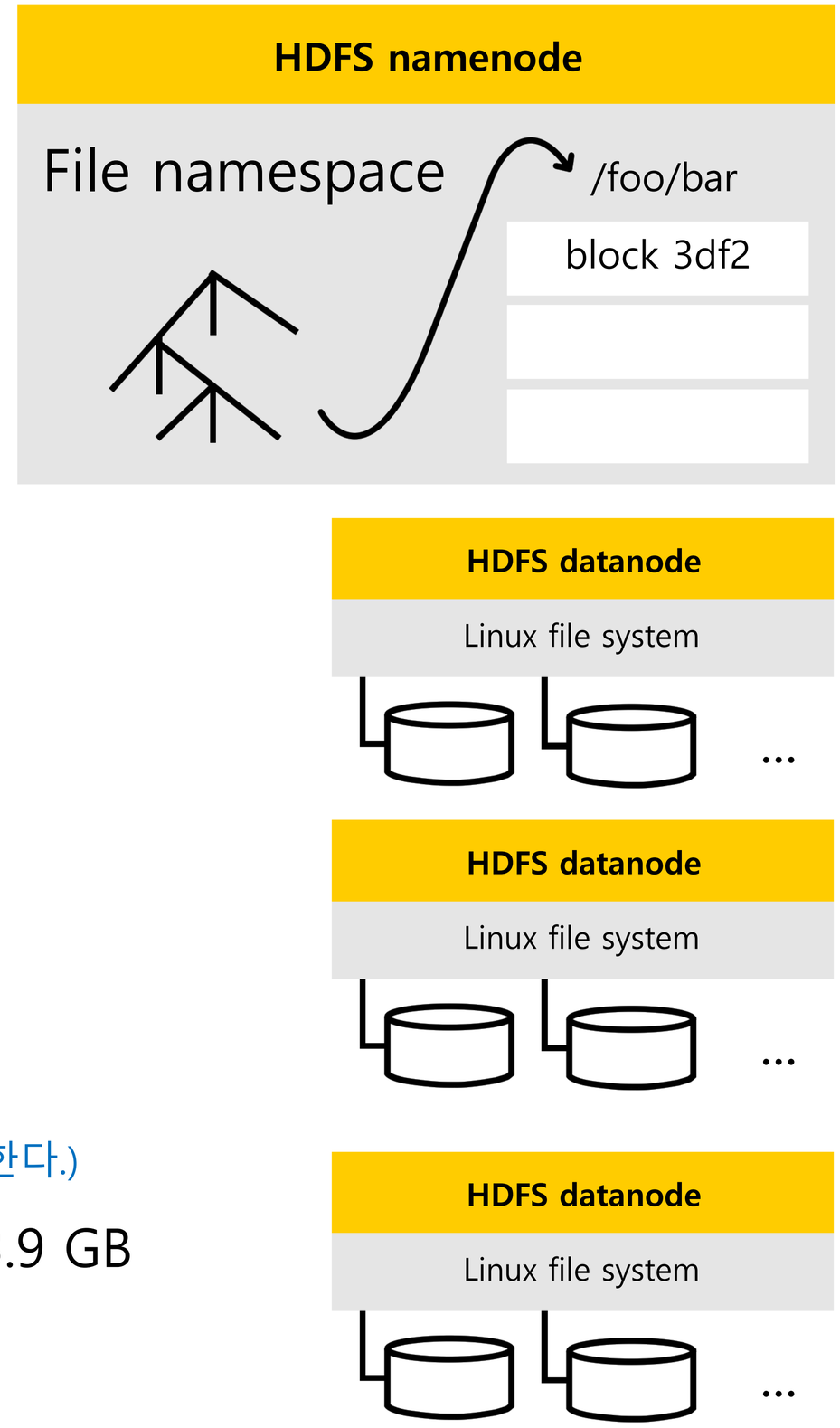
10 PB / 2 TB * 3 ~ 15 k

10PB의 데이터를 저장하기 위한 램 용량
(여기서 10PB는 replica를 포함한 데이터를 말한다.)

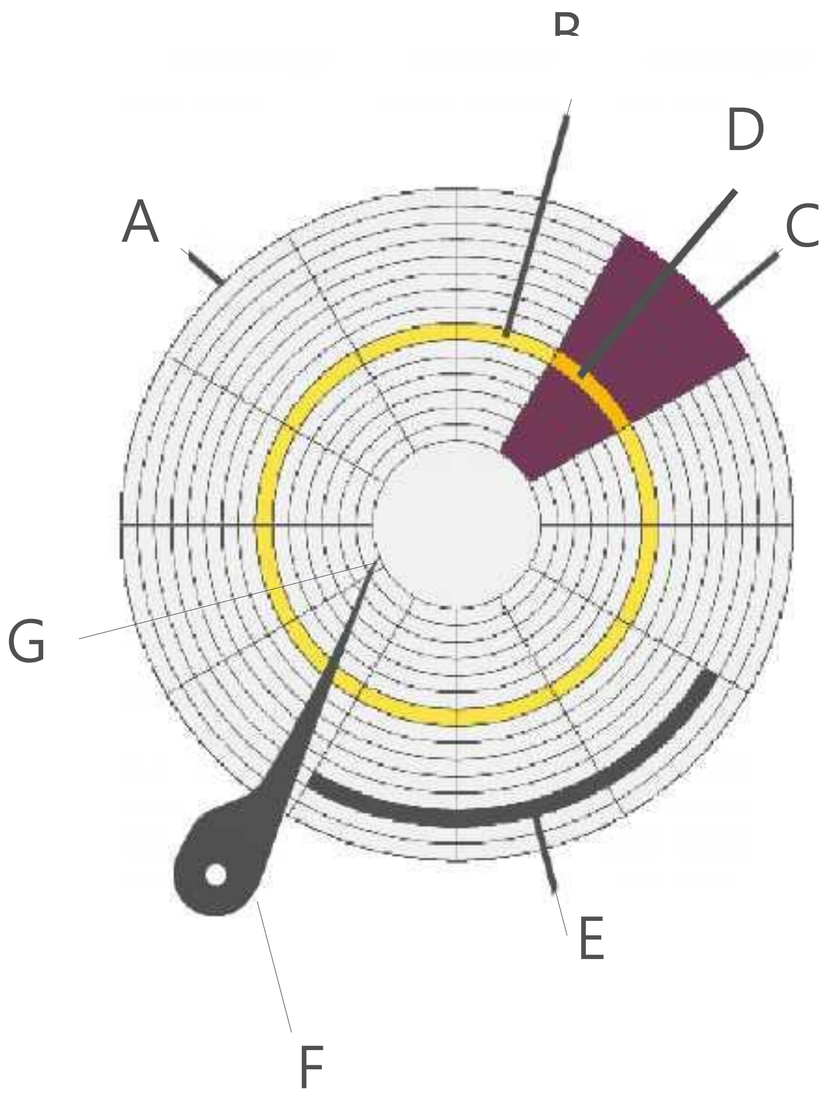


10 PB / (128 MB * 3) * 150 B ~ 3.9 GB

128MB – default block size



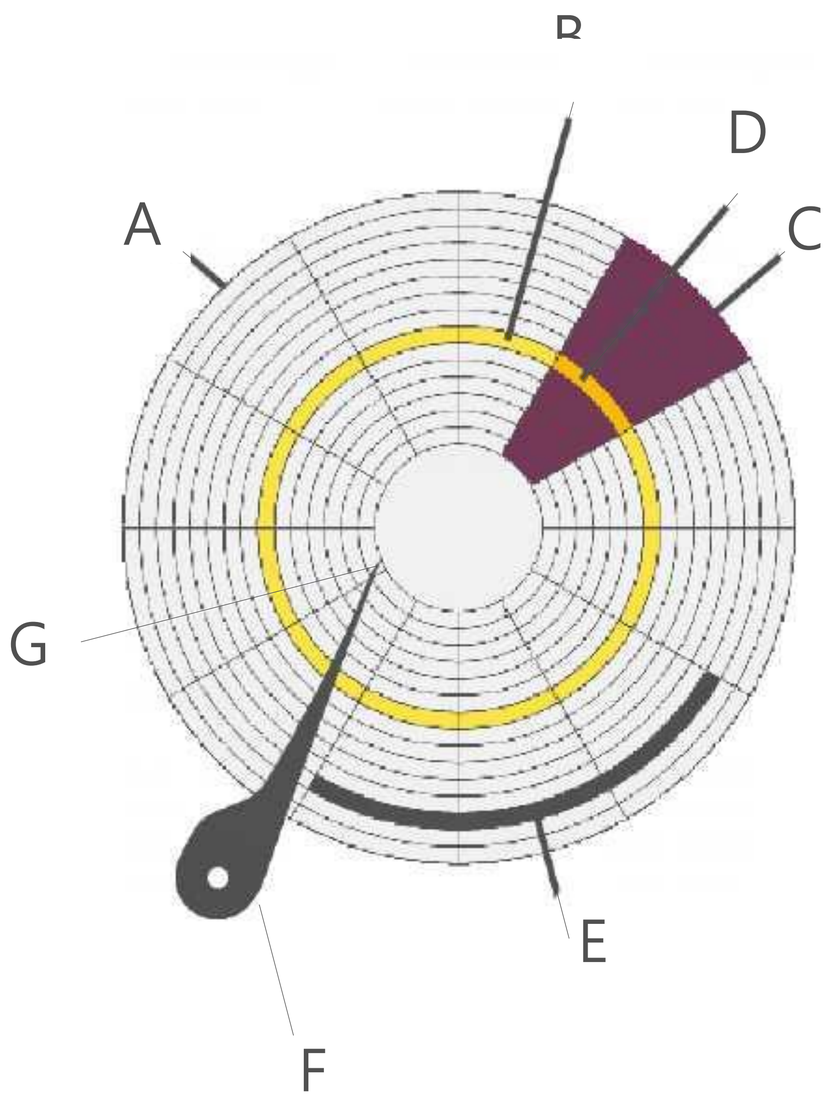
Default Block Size



A — Platter
B — Track
C — Disk Sector
D — Track Sector
E — Cluster
F — Actuator Arm
G — Head

Default Block Size

Default 설정은 64MB이지만 보통 128MB로 사용한다.



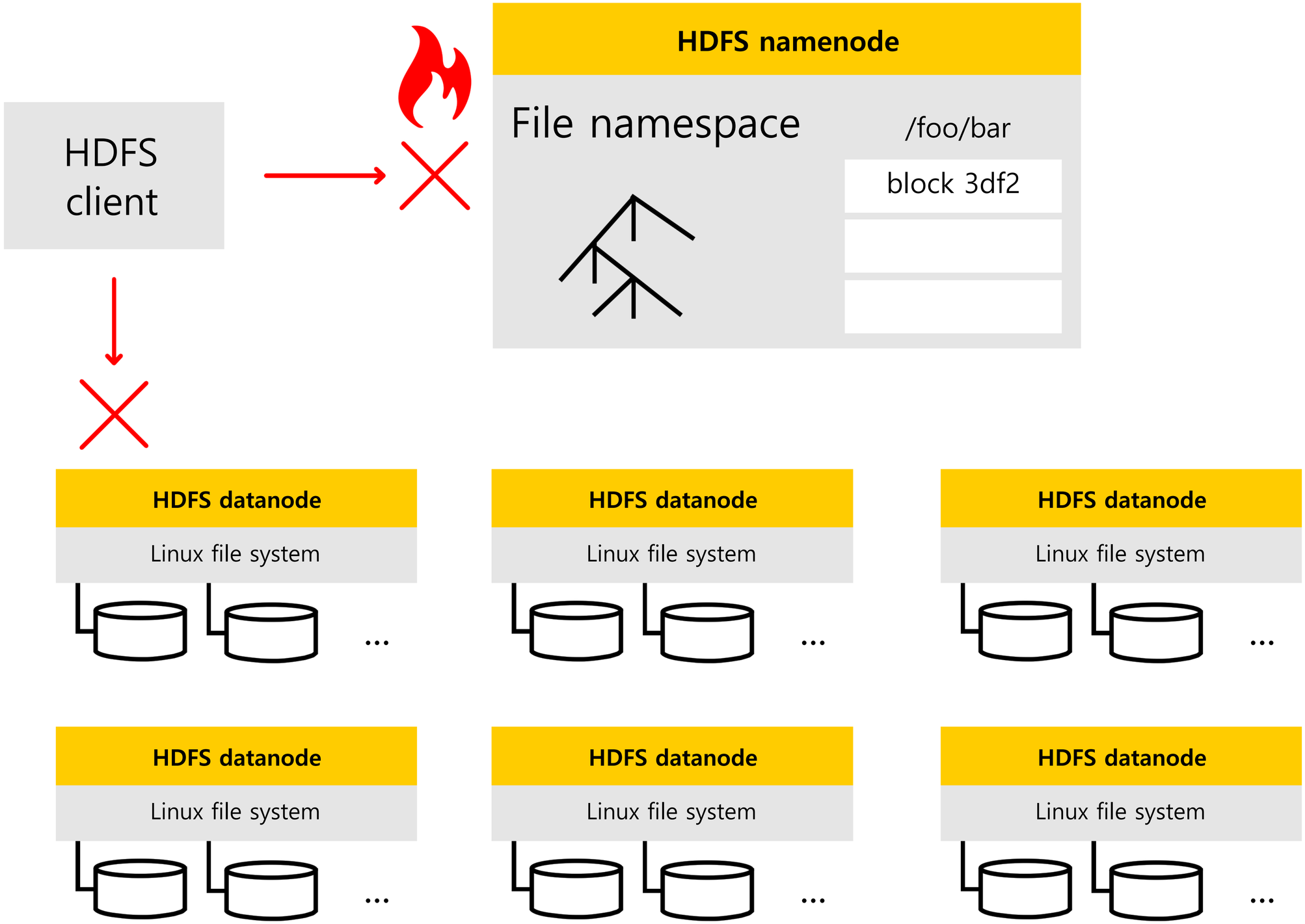
Samsung 940 PRO SSD:
* reading speed - 3.5 GB/sec
* 128 MB reading time - 30-40 ms
seek time: 0.2-0.8 ms

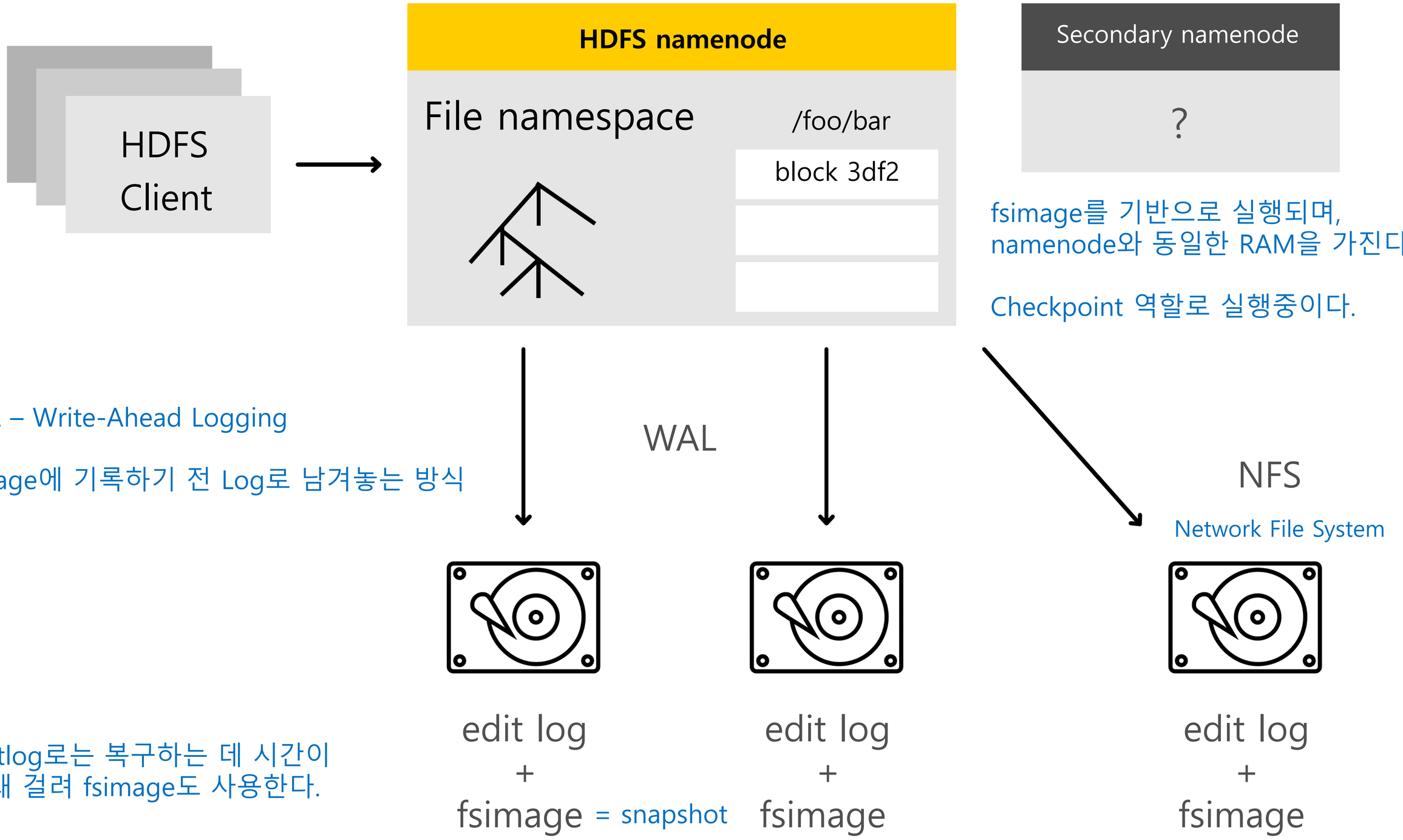
Block size 선정 기준
: seek time이 data block을 읽는 시간의 1%보다 빨라야 한다.
e.g.) 30-40ms x 0.01 = 0.3-0.4ms

small files problem

데이터를 잘게 나누면 프로세스가 많아져 bottleneck 문제와 Block 개수가 늘어나 램 용량 문제가 발생한다.

Namenode와 datanode가 동시에 문제가 생길 경우, 문제가 심각하다.





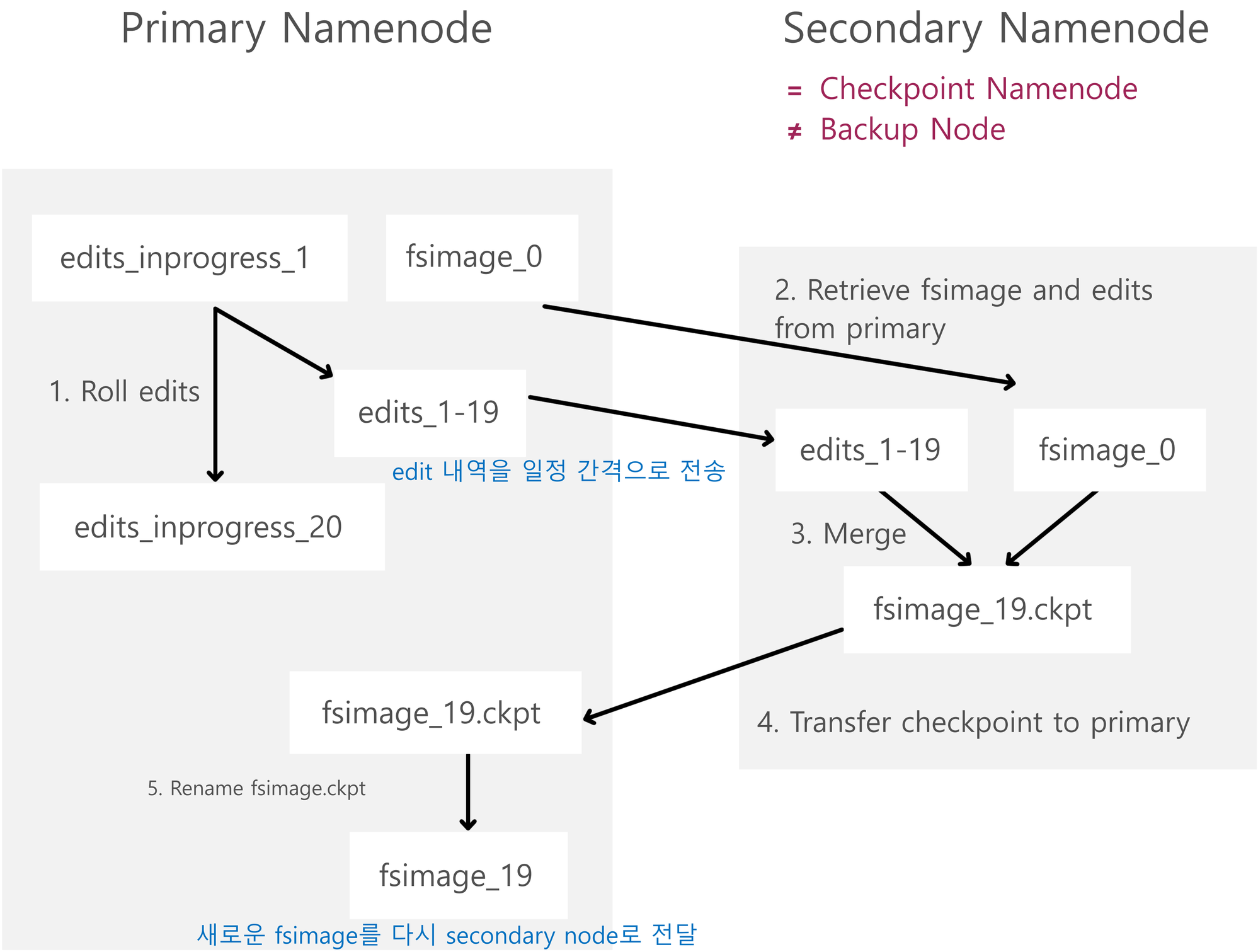
fsimage를 기반으로 실행되며,
namenode와 동일한 RAM을 가진다.

Checkpoint 역할로 실행중이다.

WAL – Write-Ahead Logging

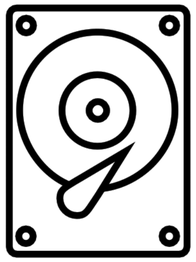
Storage에 기록하기 전 Log로 남겨놓는 방식

Editlog로는 복구하는 데 시간이
오래 걸려 fsimage도 사용한다.



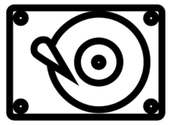
https://issues.apache.org/jira/browse/HADOOP-4539

분산 저장의 장점

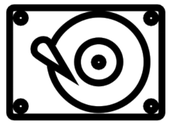


2 TB

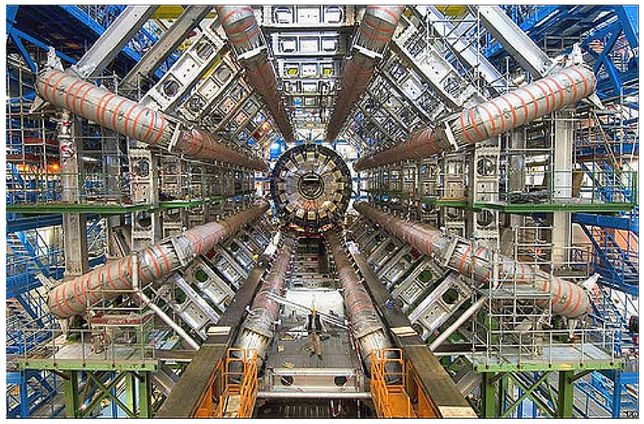
vs

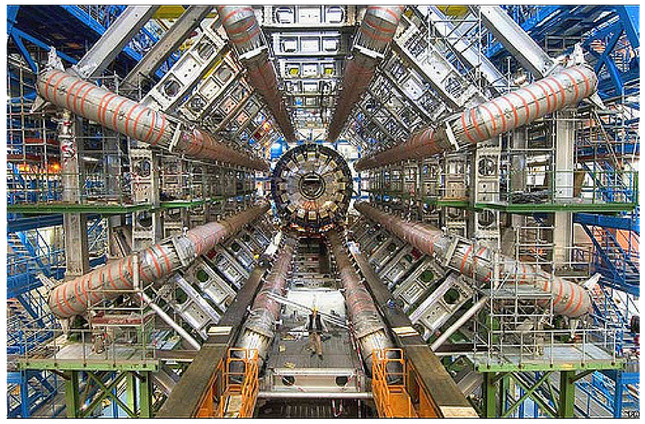


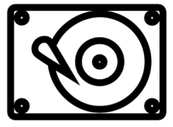
1 TB

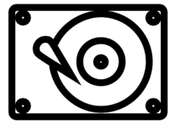
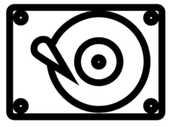


1 TB





1 year ~ 10 PB 

1 year ~ 5 PB  + 5 PB 

35 days

데이터를 읽는 시간

17.5 days

Summary

- you can explain and reason about HDFS Namenode architecture (RAM; fsimage + edit log; block size)
- you can estimate required resources for a Hadoop cluster
- you can explain what small files problem is and where a bottleneck is
- you can list differences between different types of Namenodes (Secondary / Checkpoint / Backup)

BigDATAteam