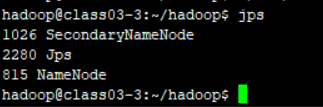
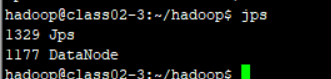
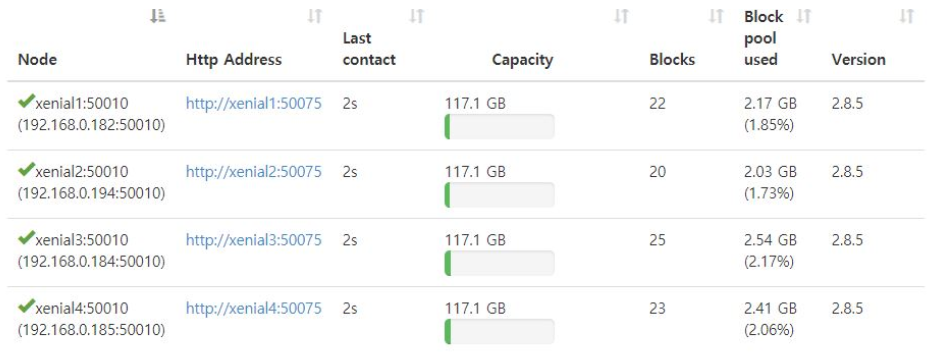
Cluster Setting review



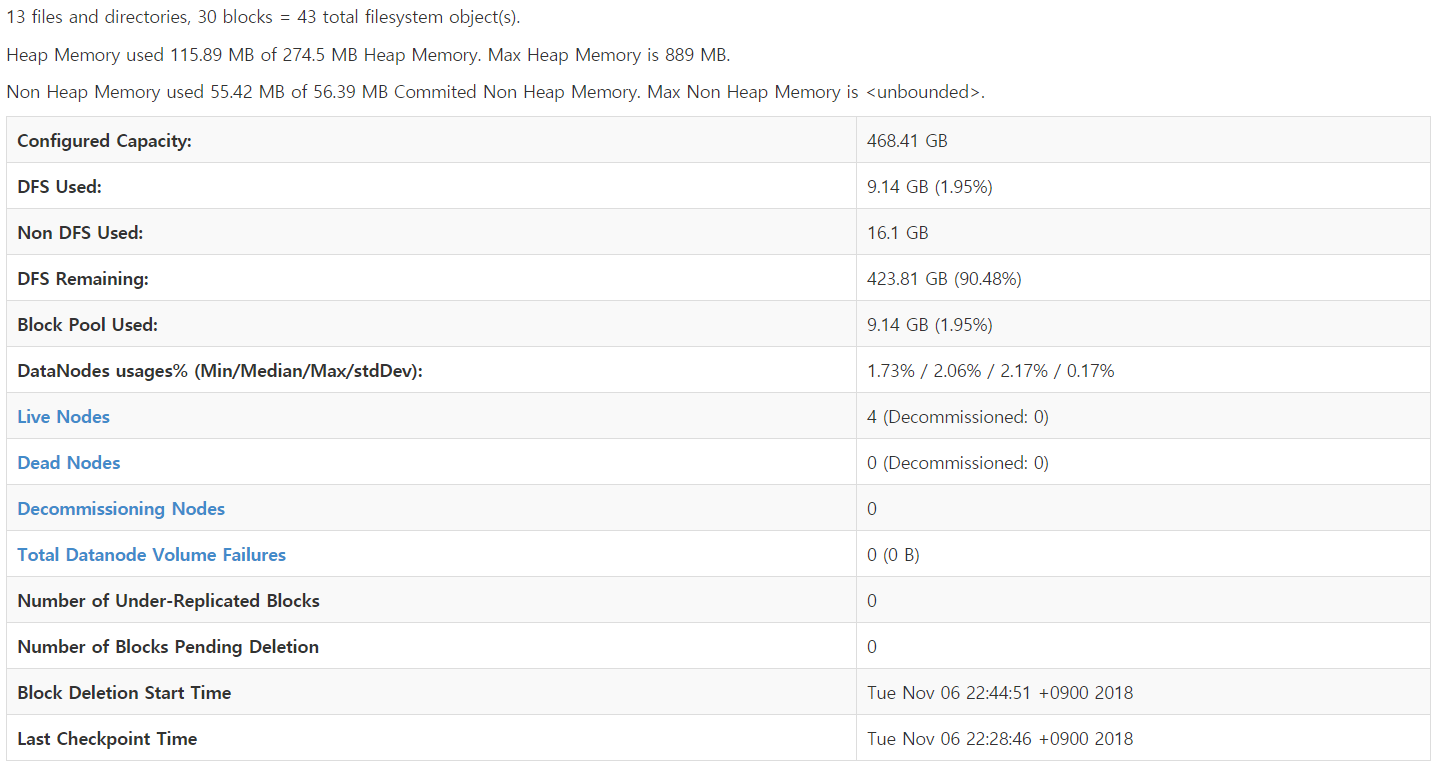
< Slave jps > < Master jps >

각 노드에서 실행 데몬을 출력결과 입니다. yarn관련 데몬인 NodeManager, ResourceManager을 끈 실습환경입니다. secondaryNameNode는 따로 설정하지 않아 master에 있습니다.



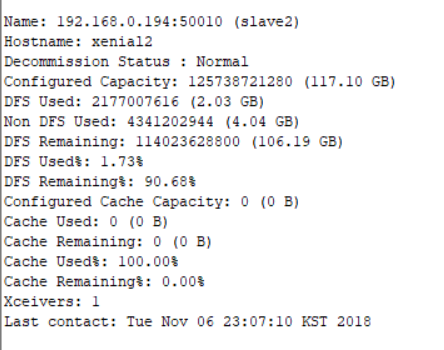
< 그림 1. master:50070 Web UI의 datanodes >

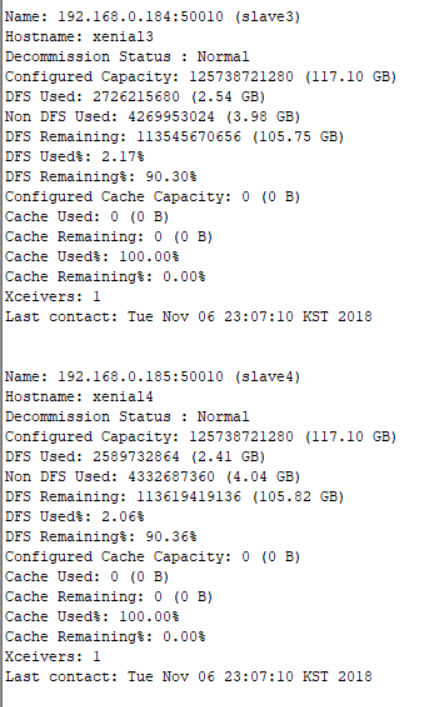
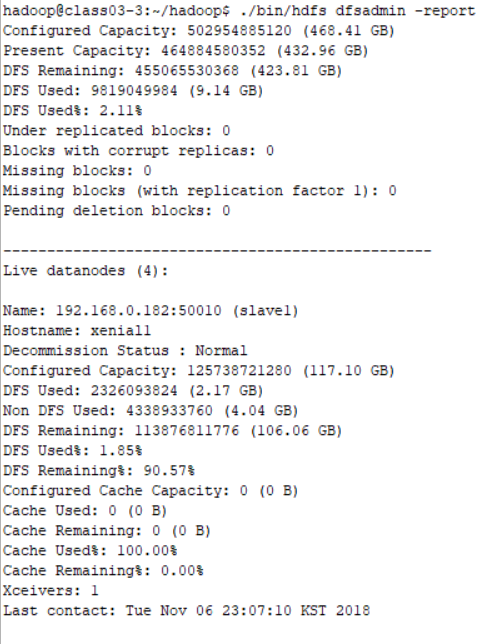
조는 총 5명이며 1명이 master, 4명이 slave node가 되었으며 master:50070의 datanodes 탭에서 그 결과를 볼 수 있습니다. slave가 총 4명이고 각각의 내부 아이피, 포트번호까지 출력됩니다. last contact 시간이 상대적으로 표시 되고 있고 각 slave의 수용 용량과 각각 몇 개의 block이 있는지 그 block들이 용량의 몇 %를 차지하는지 보여주고 있습니다. Slave의 총 block은 90(22+20+25+23)개인데 replication 수를 3으로 설정되어 있으므로 중복되지 않는 데이터가 차지하는 블록은 3으로 나누어 총 30개인 것을 유추 가능합니다.



< 그림 1. 상태에서의 master:50070에서 overview의 summary >

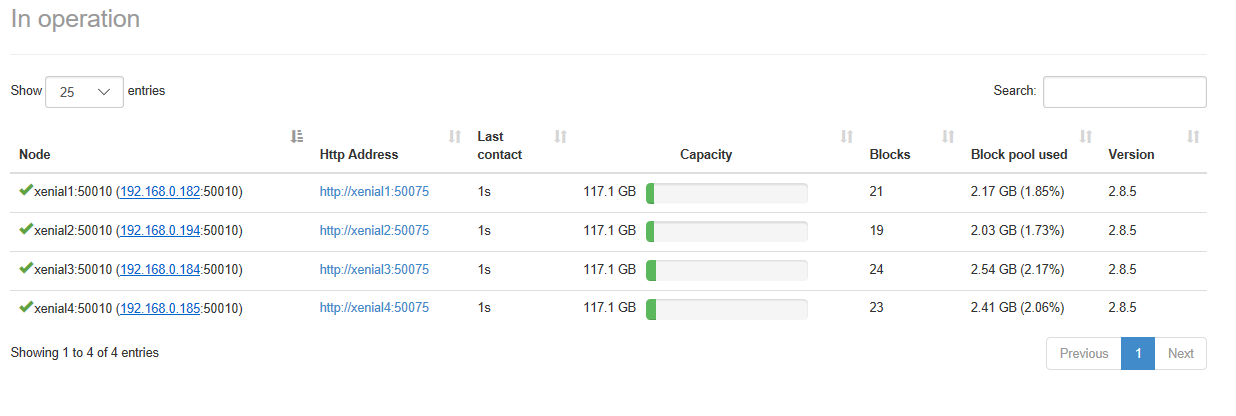
Overview의 summary tap에서는 이 클러스터의 핵심 정보를 요약해 table화 되어있습니다. 그림 첫 3줄은총 13개의 file 그리고 dir, 30개의 block이 있으므로(그림1의 설명에서 계산 됨) 총 filesystem의 object는 43개가 있고 memory 사용정보를 나타냅니다. Configured capacity는 각각의 slave의 capacity인 117.1GB \* 4로 계산되며 HDFS으로 사용되는 용량은 9.14GB 그 외의 클러스터의 기본적인 프로그램의 용량(운영체제, 기존파일)으로 slave의 disk를 온전히 사용할 수 없으므로 HDFS가 사용하지 못하는 부분이 16.1GB 입니다. 사용 가능한 총 capacity은 423.81GB입니다. Slave의 usages 정도를 최소, 최대, 평균, 표준편차값으로 계산되며, Live Nodes에서 slave가 4임을 확인합니다. 그 외에는 dead nodes, failures, 과복제 블록, 삭제 대기블록, last checkpoint time 등을 확인가능 합니다.



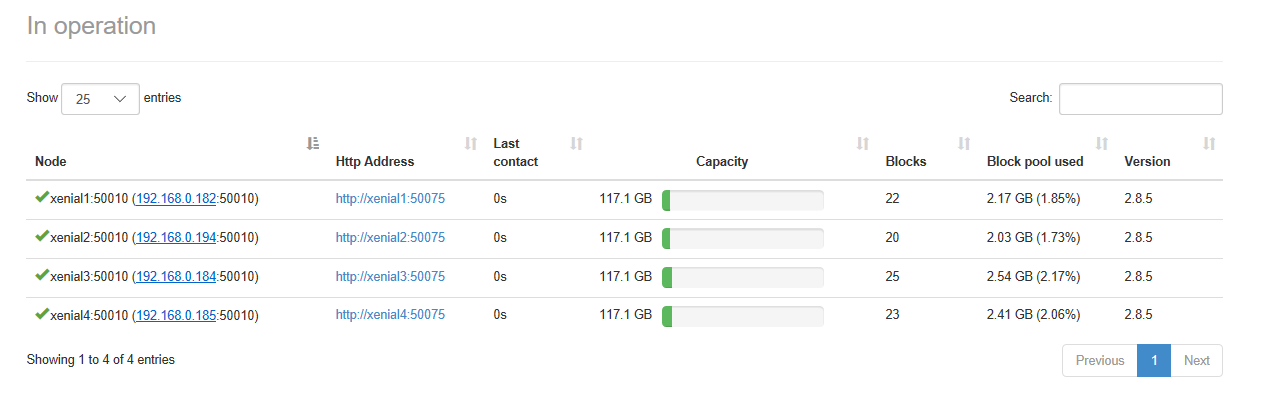


< hdfs dfsadmin –report command 입력 결과 >

hdfs dfsadmin –report command 입력 시에는 master:50070의 overview의 summary탭을 전체 cluster에서 상황을 보여준 뒤 추가적으로 각각 slave의 summary를 출력합니다.

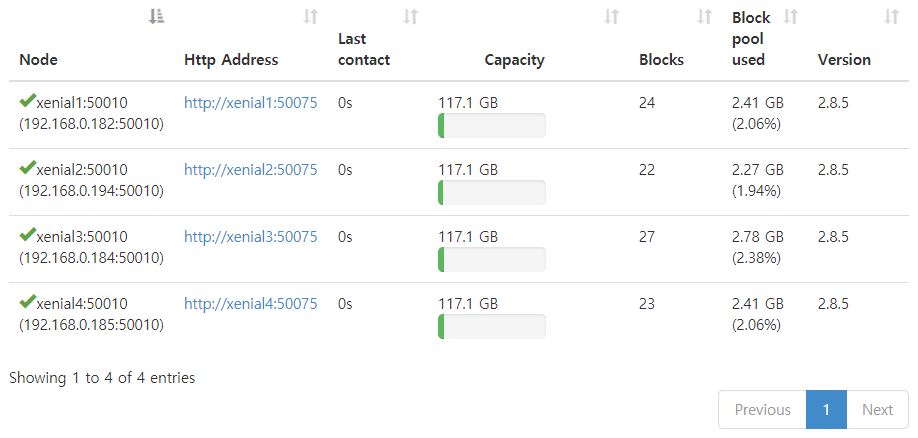
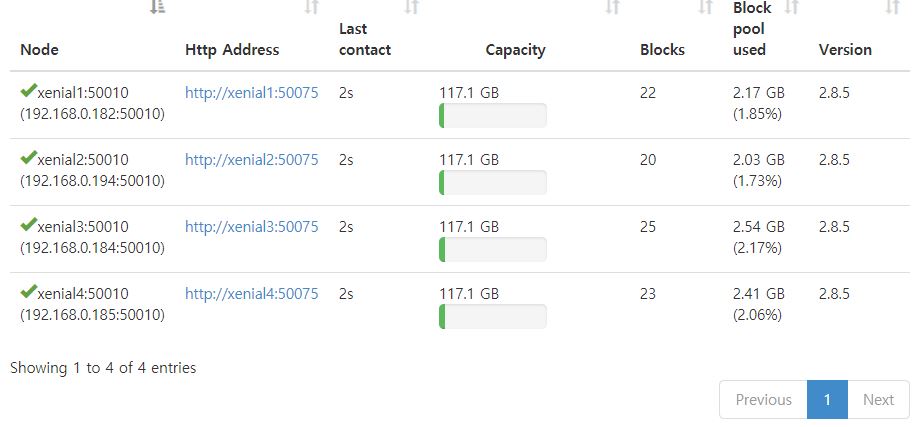


< 그림 2. 1byte file hdfs 추가 전 >



< 그림 3. 그림 2.에서 1byte file hdfs에 put한 결과 >

Cluster 구성 후 다양한 실험을 해보았는데, 임의로 1byte file 생성후 hdfs에 put한 결과입니다. replica 복제계수가 3이므로 3개의 slave node에서 각각 블록이 1씩 증가한 것을 확인 할 수 있습니다. 하지만 이 경우 파일 크기가 매우 작으므로 블록 효율이 매우 떨어집니다.



< gzip을 hdfs에 put한 결과 >

또 다른 실험은 blocksize(=128Mb)를 넘는 regular file과 compression file의 put한 결과의 차이인데, 결론은 똑같이 block 단위로 input되었고 다만 gzip파일은 split 할 수 없어 mapper 과정에서 하나의 mapper만 돌아가므로 hadoop에서 비효율적인 compression codec임을 알 수 있습니다.