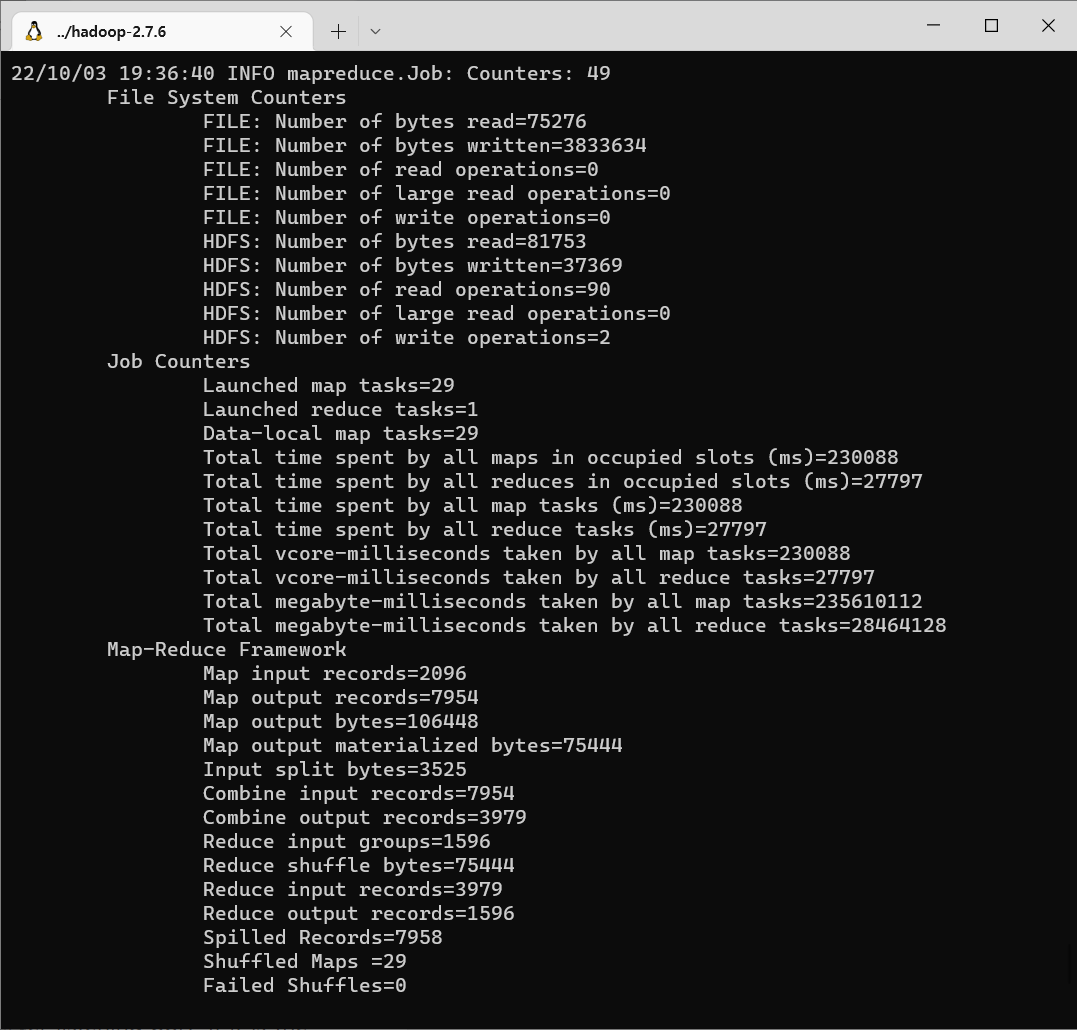
Report

컴퓨터공학과 2019102171 박다솔

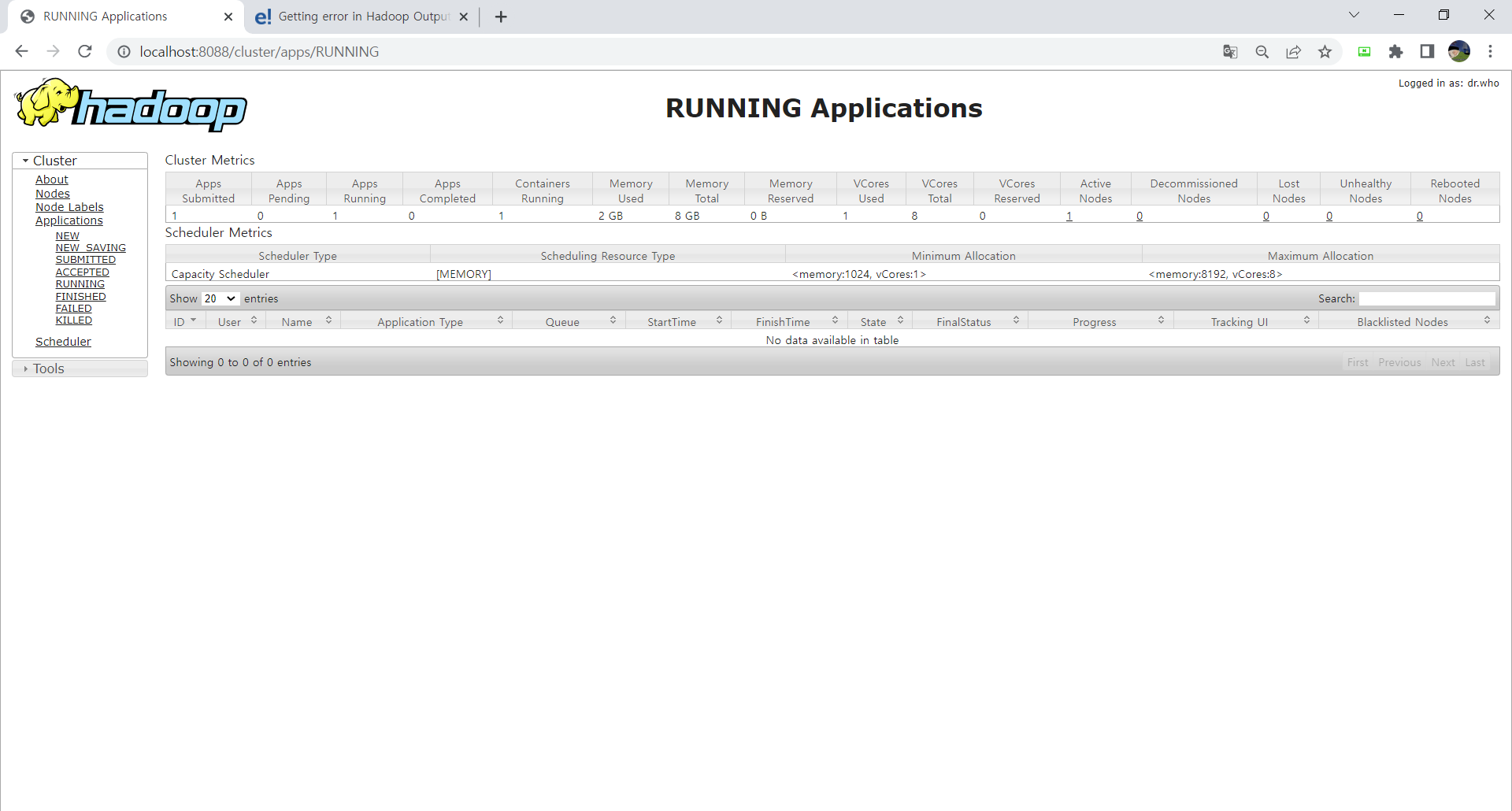
Wordcount 예제 실행 시 다음 로그가 출력된다.



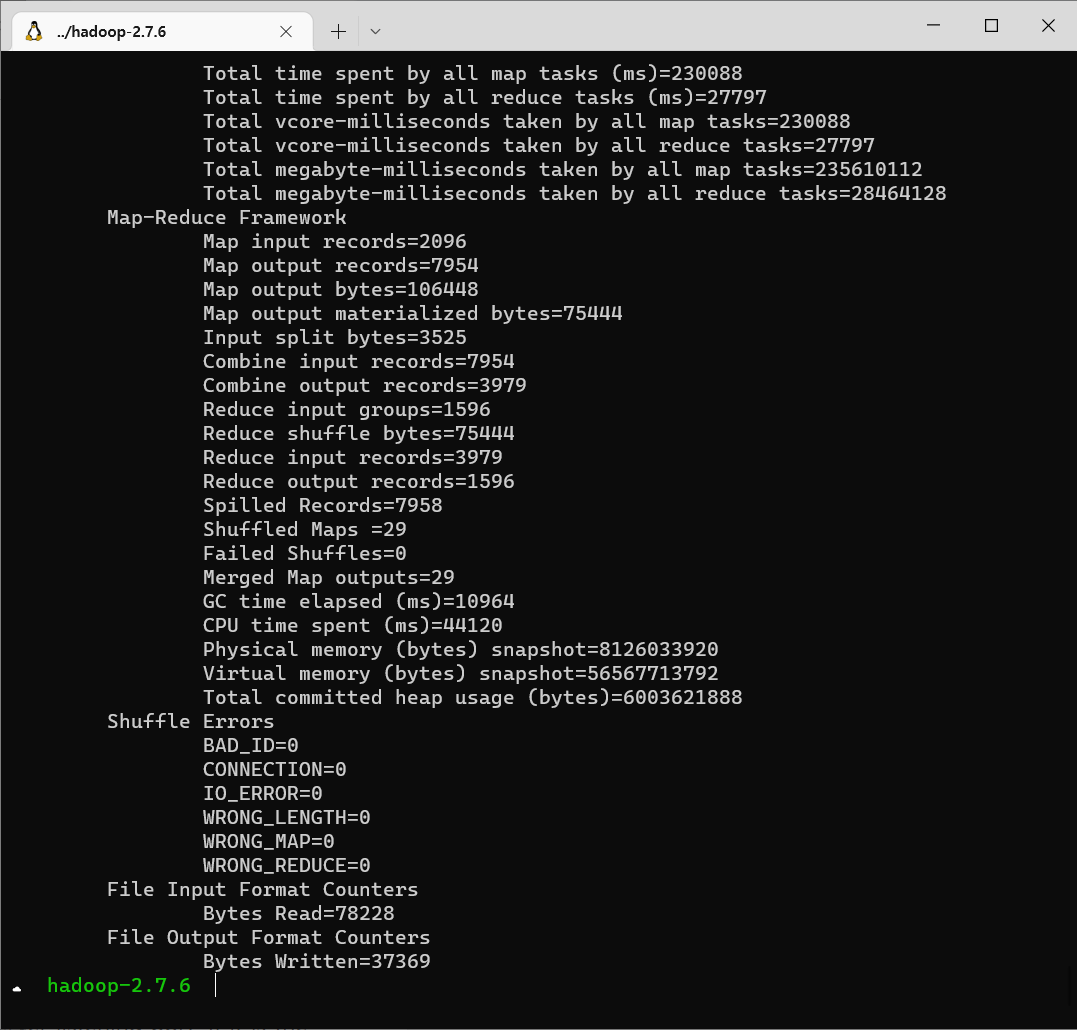
Hadoop에서 Counter는 MapReduce job에 대한 통계를 모으는 역할을 한다.

가장 위에 출력된 File System Counters는 파일 시스템에 의해 읽거나 쓰여진 byte의 수를 계산한다. FILE:은 로컬 파일 시스템을, HDFS:는 Hadoop Distributed File System 즉 분산 파일 시스템에 의해 읽거나 쓰여진 byte 수를 나타낸다.

Job Counters는 job-level을 측정한다. 작업이 running 상태일 때 변하는 값에 대해에서는 측정하지 않는다. 로그에 따르면 실행된 map task는 29개, reduce task는 1개이다. map task가 실행되는 데 걸린 총 시간은 230088ms이고 reduce는 27797ms동안 실행되었다. 컴퓨터가 실제로 작업을 처리하는 데 걸린 시간을 의미하는 wall time은 map과 reduce가 모두 전자와 동일하다. Total vcore-millisecons는 mapper 또는 reducer가 할당한 vCore의 수에 mapper가 실행된 시간을 곱한 값이다. 아래의 사진에서 vCore가 1개 할당됐음을 알 수 있다.



마지막으로 Total megabyte-milliseconds는 메모리의 양(MB)에 mapper 또는 reducer가 실행된 시간을 곱한 값을 의미한다.



Map records는 모든 mapper에서 처리된 총 record의 수이다. input은 RecordReader에서 mapper로 record가 넘어갈 때 업데이트되며, output은 OutputCollector로 넘어갈 때 업데이트 된다. Map output bytes는 mapper들에 의해 생산된 데이터의 양이며 동일하게 record가 Outputcollector로 넘어갈 때 업데이트 된다. Map output materialized bytes는 실제 디스크에 쓰여진 데이터의 양을 의미한다. Input split bytes는 split을 하면서 metadata 표현에 사용된 데이터의 양이다. Combine input records는 combiner가 처리한 총 record 수로 map output records와 동일한 7954개이다. combiner은 output에서 한 번 더 연산을 해서 reducer에게 넘기는 역할을 한다. 그러므로 Combiner에서 나온 output의 record 수는 Reduce input records 수와 3979로 동일하다. 이렇게 reducer로 들어온 record들의 unique key들의 수는 Reduce input groups에서 확인할 수 있다. 해당 unique key들로 reducer가 처리하면 groups 수와 동일한 1596개의 output이 나오게 된다. reducer로 전달되기 전 shuffling 과정을 거친 mapper는 29이며 failed한 mapper가 없으므로 output도 동일하게 29이다. GC(garbage collection)에 실제로 소요된 시간은 10964ms이고 모든 task에 대한 CPU 누적 시간은 44120ms이다. spilled Records, 즉 메모리가 초과됐을 때 디스크에 spill된 records는 7958개이며 이를 포함한 모든 task에 사용된 physical memory의 양은 Physical memory (bytes) snapshot에 나타난다. 가상 메모리 양, JVM에섯 사용 가능한 총 메모리의 양도 이어지는 로그에서 출력되었다.

Shuffle Errors는 발생한 에러 수를 나타내며 해당 로그에서는 모두 0으로 에러가 발생하지 않았다. 마지막으로 모든 파일 시스템에 대해 모든 task에서 읽은 데이터는 78228bytes이며 쓰여진 건 37369 bytes임을 확인할 수 있다.