Hadoop Project (HW#2)

1. 올려드린 하둡 매뉴얼대로 싱글노드 하둡 클러스터를 세팅합니다.
WordCount.java 소스 파일은 첨부되어 있는 파일을 그대로 사용하셔도 됩니다.
Note: 리눅스를 설치하실 때 익숙하지 않으셔서 걱정이 되시면 오라클 VirtualBox 같은 VM 프로그램을 윈도우즈에 설치하시고 그 위에다 리눅스를 설치하시는 걸 추천합니다. 이 방법이 제일 안전한 방법입니다.

2. 과제에 사용할 실제 데이터 (2006.csv)는 게시판에 upload되어 있습니다. 첨부 파일 용량이 100메가 제한이 있어서 tar.gz 으로 묶은 하나의 파일을 split해서 두개 (2006.tar.gzaa, 2006.tar.gzab) 로 쪼개서 올려드렸고 두 개 파일을 합쳐서 압축을 푸신 후 사용하시면 됩니다. 압축을 풀고 나면 672MB 정도가 됩니다.

3. 하둡에서 WordCount 를 실행합니다.
실행하실 때 화면에 나타나는 progress 를 숙제 제출용 파일에다 redirect (2>)해서 반드시 저장하시기 바랍니다.
예) hadoop jar ./wordcount.jar WordCount /user/dpark/wordcount/input
/user/dpark/wordcount/output **2>&1 | tee** output_박동철_1234567.txt

4. 하둡이 실행되고나서 최종 wordcount output 데이터를 위에서 만든 숙제 제출용 기존 파일 끝에다 append (>>) 해서 붙입니다.
예) hadoop fs -cat /user/dpark/wordcount/output/* **>>** output_박동철_1234567.txt

이렇게 하면 위 두 가지 내용을 모두 담은 하나의 파일이 만들어집니다.

5. 이렇게 리눅스에서 만들어진 output 파일은 윈도에서 보면 줄바꿈 문자형식의 차이로 인해 양식이 다 깨어져 있으므로 아래와 같이 변환처리를 거친 후 제출해 주시기 바랍니다 (변환처리 안할 경우 감점 -5).
예) todos output_박동철_1234567.txt

만약 tofrodos 패키지가 우분투에 설치되어 있지 않을 경우 아래와 같이 패키지를 먼저 설치한 후 위 명령어 대로 변경해주시면 됩니다.
예) sudo apt-get install tofrodos

[제출물]
1. 위에서 실행결과를 redirect 한 **output 파일 (파일명: output_이름_학번.txt)**
2. 온라인 제출시 위 첨부파일과 함께 제출란에 다음 두 가지를 반드시 명시할 것.
    2.1) 하둡 매뉴얼대로 세팅을 마친 후, 예제에 나오는 작은 사이즈 (few KB ~ few MB)의 데이터를 이용하여 wordcount 를 실행하면 아무런 문제가 없습니다. 그렇지만 2006.csv 데이터처럼 대용량의 데이터를 이용할 경우 하둡에 문제가 생기면서 제대로 실행이 안됩니다. 이때 나타나는 **에러가 무엇인지 간략하게 스노보드에 명시**하시기 바랍니다. (그럴 가능성은 희박하지만 혹시라도 아무런 문제없이 잘 실행이 된다면

'문제없음'이라고 명시하시면 됩니다. 대신 문제없이 실행된 실행결과를 위 output 파일을 통해서 보여주셔야 합니다.). 다시 말해서 매뉴얼에 나오는 작은 데이터가 문제없이 돌아갔더라도 2006.csv 같이 큰 데이터를 돌리면 에러가 날 확률이 아주 높습니다. 아마 거의 100% 에러가 날겁니다. 따라서 에러가 나는 것은 너무나 당연하고 이때 에러가 무엇때문에 발생했는지 명시를 하고, 그 문제를 해결하셔서 2006.csv가 돌아가도록 만들고 어떻게 해결했는지 명시를 하시는 게 이번 과제의 최종목표입니다.

2.2) 위에 생긴 에러를 어떻게 해결했는지 간략하게 **해결 방법을 과제제출시 스노보드에 같이 명시**해주시기 바랍니다.

[제출기간] 11월 17 일 (목) 수업 전까지 (13:30). 제출 기한 엄수 (**지각 제출은 0 점 처리**)

온라인으로만 제출하시면 됩니다.