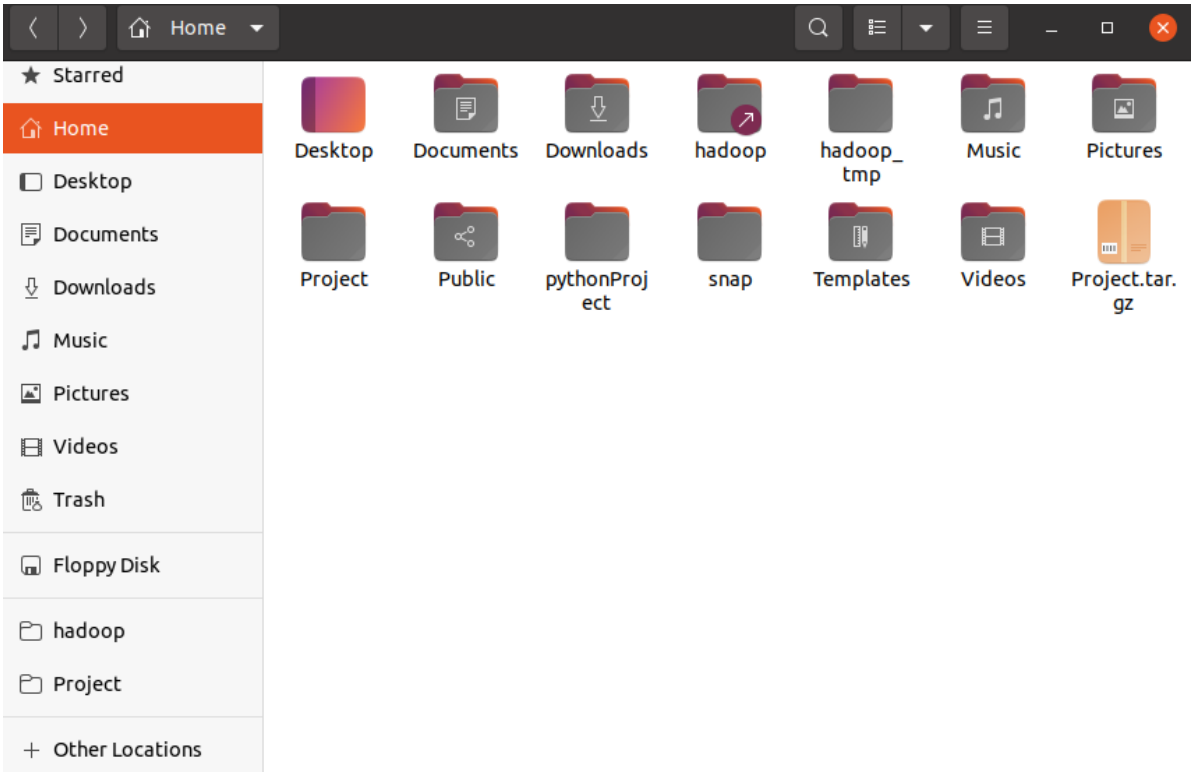
hadoop_streaming

hadoop_streaming

- 하둡을 파이썬이나 기타 다른 언어로 돌리는 방법
- 아래 참조문서들을 참고하면 좀더 쉽다

폴더구조

- 이 문서에서 대부분의 코드는 Project 폴더에서 실행
- pythonProject 내부에 pmapper.py와 preducer.py가 존재
- hdfs 폴더구조는 따로 표시하지 않음



python code

- 가장 상단에 `#!/usr/bin/env python`을 표시하여 우분투 파이썬 경로를 불러오는 것이 중요!
- excode: pmapper.py

```
#!/usr/bin/env python

import sys
for line in sys.stdin:
    line = line.strip()
    words = line.split()

    for word in words:
        print ('%s\t%s' % (word,1))
```

- excode: preucer.py

```
#!/usr/bin/python

from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

for line in sys.stdin:
    line = line.strip()
    word, count = line.split('\t',1)
    try:
        count = int(count)
    except ValueError:
        continue

    if current_word == word:
        current_count +=count
    else:
        if current_word:
            print ('%s\t%s' % (current_word, current_count))
        current_count = count
        current_word = word

if current_word == word:
    print ('%s\t%s' % (current_word, current_count))
```

find

- 아래 코드는 `hadoop-streaming*.jar`를 찾기위한 코드로 위치를 알면 생략해도 된다

```
find / -name 'hadoop-streaming*.jar'
```

- 실행코드

```
hadoop@ubuntu:~/Project$ find / -name 'hadoop-streaming*.jar'
find: ‘/snap/core18/1988/etc/ssl/private’: Permission denied
```

- 실행결과

```
/usr/local/hadoop/share/hadoop/tools/sources/hadoop-streaming-3.2.2-sources.jar
/usr/local/hadoop/share/hadoop/tools/sources/hadoop-streaming-3.2.2-test-sources.jar
/usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.2.jar
```

mapreduce 실행하기

- 사전 준비(실제 예시)
  - hdfs dfs -mkdir p_wordcount : hdfs에 데이터 저장폴더 생성하기
  - hdfs dfs -put data/wordcount-data.txt p_wordcount : input data 가져오기
  - hdfs dfs -rm -r p_wordcount_out : output파일 삭제하기 (이름은 본인이 만드는 것)

- maprduce 코드 양식

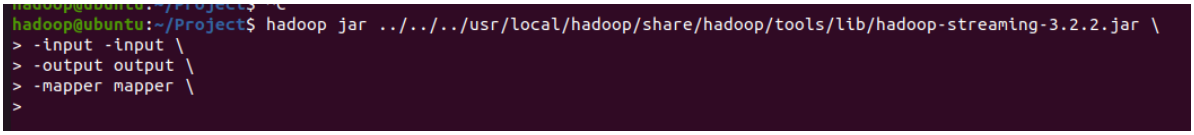
```
hadoop jar /hadoop-streaming-3.2.2.jar \
-input input \
-output output \
-mapper 'python3 mapper.py' \
-reducer 'python3 reducer.py'
```

- maprduce 실제 실행 코드

```
hadoop jar ../../../usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-
3.2.2.jar \
-D mapreduce.job.reduces=12 \
-input p_wordcount \
-output p_wordcount_out \
-mapper 'python3 ../pythonProject/pmapper.py' \
-reducer 'python3 ../pythonProject/preducer.py'
```

1. hadoop jar ..... : 하둡을 이용하여 jar파일을 실행시키겠다는 뜻, 이 코드에서는hadoop-streaming-3.2.2.jar를 실행시켰다
2. -D mapreduce.job.reduces : 파일결과를 몇 분할 할 것인지 이 예제에서는 결과를 12등분한다.
3. -input: p_wordcount : `p_wordcount`를 input파일로 사용한다
4. -output `p_wordcount_out`로 output파일을 저장한다. 이 때 이름은 자유롭게 지정가능하며 만약 같은 파일이 있다면 에러가 발생한다
5. -mapper 'python3 ...' : mapper로 다음 파일을 사용한다. 여기서 ''가 없거나 ''안에 python3를 명시해주지 않으면 default가 java기 때문에 에러가 발생한다.
6. -reudce 'python3 ...' : reducer로 다음 파일을 사용한다. 여기서 ''가 없거나 ''안에 python3를 명시해주지 않으면 default가 java기 때문에 에러가 발생한다.

cf). \(백슬래시) 는 명령어가 끝나지 않았다는 의미로 만약 치다 헷갈리거나 너무 길어질 경우 끊어칠 때 사용하면 된다.



결과

- 결과값은 wordcount_out 이라는 `dictionary`로 존재하여 기존 명령어인`hdfs dfs -cat p_wordcount_out/part-r-00000 |more`으로 할 경우 정확한 값을 알 수 없다

```
hadoop@ubuntu:~/Project$ hdfs dfs -ls p_wordcount_out
Found 13 items
-rw-r--r--   1 hadoop supergroup          0 2021-09-09 19:13 p_wordcount_out/_SUCCESS
-rw-r--r--   1 hadoop supergroup        383 2021-09-09 19:13 p_wordcount_out/part-00000
-rw-r--r--   1 hadoop supergroup        430 2021-09-09 19:13 p_wordcount_out/part-00001
-rw-r--r--   1 hadoop supergroup        449 2021-09-09 19:13 p_wordcount_out/part-00002
-rw-r--r--   1 hadoop supergroup        464 2021-09-09 19:13 p_wordcount_out/part-00003
-rw-r--r--   1 hadoop supergroup        438 2021-09-09 19:13 p_wordcount_out/part-00004
-rw-r--r--   1 hadoop supergroup        508 2021-09-09 19:13 p_wordcount_out/part-00005
-rw-r--r--   1 hadoop supergroup        420 2021-09-09 19:13 p_wordcount_out/part-00006
-rw-r--r--   1 hadoop supergroup        457 2021-09-09 19:13 p_wordcount_out/part-00007
-rw-r--r--   1 hadoop supergroup        548 2021-09-09 19:13 p_wordcount_out/part-00008
-rw-r--r--   1 hadoop supergroup        541 2021-09-09 19:13 p_wordcount_out/part-00009
-rw-r--r--   1 hadoop supergroup        424 2021-09-09 19:13 p_wordcount_out/part-00010
-rw-r--r--   1 hadoop supergroup        545 2021-09-09 19:13 p_wordcount_out/part-00011
```

- `hdfs dfs -cat p_wordcount_out/part-00000 |more`로 정확하게 지정해줘야한다.

```
hadoop@ubuntu:~/Project$ hdfs dfs -cat p_wordcount_out/part-00000 |more
"rejoicing      1
"undo   1
(and)   1
.       3
Administration, 1
All     1
Almighty        1
America 2
Americans       2
Americans--born 1
Americans:      1
Americas.       1
And     4
But     5
Can     1
Divided 1
East    1
Finally,        2
For     3
God     1
God's   1
God.    1
Hemisphere      1
His     2
I       3
If      1
In      2
Isaiah--to      1
Let     8
My      1
Nations,        1
Nor     1
North   1
Now     1
Since   1
So      1
South,  1
The     3
This    1
To      5
```

추가 공부할 점

- main 함수 적용방식이 조금 달라서 그 부분은 더 공부해봐야할 거 같다(공식문서 내용참조)

```
$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/hadoop-streaming.jar \
    -D
mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBase
dComparator \
    -D stream.map.output.field.separator=. \
    -D stream.num.map.output.key.fields=4 \
    -D map.output.key.field.separator=. \
    -D mapred.text.key.comparator.options=-k2,2nr \
    -D mapred.reduce.tasks=12 \
    -입력 myInputDirs \
    -출력 myOutputDir \
    -mapper org.apache.hadoop.mapred.lib.IdentityMapper \
    -reducer org.apache.hadoop.mapred.lib.IdentityReducer
```

참고문헌

- https://hadoop.apache.org/docs/r1.2.1/streaming.html
- https://www.youtube.com/watch?v=QNB1SZm2jS4&t=738s
- https://www.youtube.com/watch?v=rsMQ1Z3KZLM&t=47s
- https://www.youtube.com/watch?v=TcBkvCKE1rw&t=1831s