

```
In [1]: !hadoop fs -ls /
```

```
Found 21 items
drwxr-xr-x - CSE-41 supergroup 0 2022-09-15 12:18 /BDA
drwxr-xr-x - CSE-41 supergroup 0 2022-09-16 10:32 /Test
drwxr-xr-x - CSE-41 supergroup 0 2022-10-07 11:26 /input
drwxr-xr-x - CSE-41 supergroup 0 2022-09-26 15:23 /input1
drwxr-xr-x - CSE-41 supergroup 0 2022-09-23 06:56 /input2
drwxr-xr-x - CSE-41 supergroup 0 2022-09-27 13:08 /output
drwxr-xr-x - CSE-41 supergroup 0 2022-10-06 10:43 /output06102022
drwxr-xr-x - CSE-41 supergroup 0 2022-10-10 11:49 /output2050
drwxr-xr-x - CSE-41 supergroup 0 2022-09-28 14:37 /output_1
drwxr-xr-x - CSE-41 supergroup 0 2022-09-29 11:43 /output_20
drwxr-xr-x - CSE-41 supergroup 0 2022-09-30 11:41 /output_2022
drwxr-xr-x - CSE-41 supergroup 0 2022-10-06 11:24 /output_2023
drwxr-xr-x - CSE-41 supergroup 0 2022-09-29 11:38 /output_LAB
drwxr-xr-x - CSE-41 supergroup 0 2022-10-07 11:28 /output_friday
drwxr-xr-x - CSE-41 supergroup 0 2022-10-12 14:28 /output_wed
drwxr-xr-x - CSE-41 supergroup 0 2022-10-06 12:34 /sample
drwxr-xr-x - CSE-41 supergroup 0 2022-10-06 12:43 /sample_2
drwxr-xr-x - CSE-41 supergroup 0 2022-10-07 11:35 /sample_friday
drwxr-xr-x - CSE-41 supergroup 0 2022-09-12 06:20 /system
drwx----- - CSE-41 supergroup 0 2022-09-23 06:45 /tmp
drwxr-xr-x - CSE-41 supergroup 0 2022-09-26 15:27 /user
```

```
In [2]: !hadoop fs -mkdir /sample_friday
```

```
In [3]: !hadoop fs -ls /
```

```
Found 19 items
drwxr-xr-x - CSE-41 supergroup 0 2022-09-15 12:18 /BDA
drwxr-xr-x - CSE-41 supergroup 0 2022-09-16 10:32 /Test
drwxr-xr-x - CSE-41 supergroup 0 2022-10-07 11:26 /input
drwxr-xr-x - CSE-41 supergroup 0 2022-09-26 15:23 /input1
drwxr-xr-x - CSE-41 supergroup 0 2022-09-23 06:56 /input2
drwxr-xr-x - CSE-41 supergroup 0 2022-09-27 13:08 /output
drwxr-xr-x - CSE-41 supergroup 0 2022-10-06 10:43 /output06102022
drwxr-xr-x - CSE-41 supergroup 0 2022-09-28 14:37 /output_1
drwxr-xr-x - CSE-41 supergroup 0 2022-09-29 11:43 /output_20
drwxr-xr-x - CSE-41 supergroup 0 2022-09-30 11:41 /output_2022
drwxr-xr-x - CSE-41 supergroup 0 2022-10-06 11:24 /output_2023
drwxr-xr-x - CSE-41 supergroup 0 2022-09-29 11:38 /output_LAB
drwxr-xr-x - CSE-41 supergroup 0 2022-10-07 11:28 /output_friday
drwxr-xr-x - CSE-41 supergroup 0 2022-10-06 12:34 /sample
drwxr-xr-x - CSE-41 supergroup 0 2022-10-06 12:43 /sample_2
drwxr-xr-x - CSE-41 supergroup 0 2022-10-07 11:34 /sample_friday
drwxr-xr-x - CSE-41 supergroup 0 2022-09-12 06:20 /system
drwx----- - CSE-41 supergroup 0 2022-09-23 06:45 /tmp
drwxr-xr-x - CSE-41 supergroup 0 2022-09-26 15:27 /user
```

```
In [4]: local_file_path = "D:\BigData\hello_friday.txt"
!hadoop fs -put $local_file_path /sample_friday
```

```
2022-10-07 11:35:26,152 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localhostTrusted = false, remoteHostTrusted = false
```

```
In [5]: !hadoop fs -ls /sample_friday
```

```
Found 1 items
-rw-r--r--  1 CSE-41 supergroup          34 2022-10-07 11:35 /sample_friday/hel
lo_friday.txt
```

```
In [16]: !hadoop fs -cat /sample/hello_5.txt
```

```
cat: `/sample/hello_5.txt': No such file or directory
```

```
In [17]: pwd
```

```
Out[17]: 'C:\\Users\\CSE-41\\BDAProgram'
```

```
In [ ]: ls -ltr 'C:\\Users\\CSE-41\\BDAProgram'
```

```
In [ ]: cd Downloads
```

```
In [ ]: cd hadoop-3.1.0/
```

```
In [ ]: ls
```

```
In [ ]: #!/usr/bin/env python
        """mapper.py"""

import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    # increase counters
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        #
        # tab-delimited; the trivial word count is 1
        print_word = '%s\t%s' % (word, 1)
        print(print_word) # print '%s\t%s' % (word, 1)
```

```

In [ ]: #!/usr/bin/env python
        """reducer.py"""

        from operator import itemgetter
        import sys

        current_word = None
        current_count = 0
        word = None

        # input comes from STDIN
        for line in sys.stdin:
            # remove leading and trailing whitespace
            line = line.strip()

            # parse the input we got from mapper.py
            word, count = line.split('\t', 1)

            # convert count (currently a string) to int
            try:
                count = int(count)
            except ValueError:
                # count was not a number, so silently
                # ignore/discard this line
                continue

            # this IF-switch only works because Hadoop sorts map output
            # by key (here: word) before it is passed to the reducer
            if current_word == word:
                current_count += count
            else:
                if current_word:
                    # write result to STDOUT
                    print_count = '%s\t%s' % (current_word, current_count)
                    print(print_count)
                current_count = count
                current_word = word

            # do not forget to output the last word if needed!
            if current_word == word:
                word_count = '%s\t%s' % (current_word, current_count)
                print(word_count)


```

```

In [ ]: local_file_path = "D:\BigData\hello_2.txt"
        !hadoop fs -put $local_file_path /sample


```

```


In [ ]: !hadoop fs -ls /sample


```

In [6]: `!hadoop fs -ls /sample_friday`

```
Found 1 items
-rw-r--r--    1 CSE-41 supergroup          34 2022-10-07 11:35 /sample_friday/hello_friday.txt
```

In [2]: `!hadoop fs -cat /sample_friday/hello_friday.txt | C:\ProgramData\Anaconda3\python.exe`



```
hi      1
how     1
are     1
you     1
hi      1
today   1
is      1
friday  1
```

```
2022-10-12 14:41:51,186 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localhostTrusted = false, remoteHostTrusted = false
```

In [11]: `!hadoop fs -cat /sample_friday/hello_friday.txt | C:\ProgramData\Anaconda3\python.exe D:\BigData\dataStreaming\mapper.py | sort | C:\ProgramData\Anaconda3\python.exe D:\BigData\dataStreaming\reducer.py`

```
are      1
```

```
2022-10-07 11:47:56,080 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localhostTrusted = false, remoteHostTrusted = false
```

```
friday   1
hi        2
how       1
is        1
today     1
you       1
```

In [12]: `!hadoop fs -cat /sample_2/hello_5.txt`

```
hi how are you
hi today is thursday
```

```
2022-10-07 11:48:16,279 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localhostTrusted = false, remoteHostTrusted = false
```

In [2]: !hadoop fs -cat /sample_2/hello_5.txt | C:\ProgramData\Anaconda3\python.exe D:\B

```
are      1
hi       2
how      1
is       1
thursday      1
today      1
you       1
```

2022-10-10 12:01:38,926 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false

In [3]: !hadoop fs -cat /sample/hello_2.txt

```
hi how are you
today is thursday
```

2022-10-06 12:33:13,490 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false