```
!pip install mrjob
```

```
Collecting mrjob
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 439.6/439.6 kB 5.0 MB/s eta 0:00:00
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.10/dist-packages (from mrjob) (6.0.1)
Installing collected packages: mrjob
Successfully installed mrjob-0.7.4
```

```
%%writefile textfile.txt
Big Data, haven’t you heard this term before? I am sure you have. In the last 4 to 5 years, everyone is talking about Big Data. But do you re
Below are the topics which I will cover in this Big Data Tutorial:
    • Story of Big Data
    • Big Data Driving Factors
    • What is Big Data?
    • Big Data Characteristics
    • Types of Big Data
    • Examples of Big Data
    • Applications of Big Data
    • Challenges with Big Data

Story of Big Data
In ancient days, people used to travel from one village to another village on a horse driven cart, but as the time passed, villages became to
The same concept applies on Big Data. Big Data says, till today, we were okay with storing the data into our servers because the volume of th
Through this blog on Big Data Tutorial, let us explore the sources of Big Data, which the traditional systems are failing to store and proces
```

```
Writing textfile.txt
```

```
%%writefile worcount.py
from mrjob.job import MRJob
import re
class WordCount(MRJob):

    def mapper(self, _, line):
        words = re.findall(r'\w+', line.lower())
        for word in words:
            yield word, 1

    def combiner(self, word, counts):
        yield word, sum(counts)

    def reducer(self, word, counts):
        yield word, sum(counts)

if __name__ == '__main__':
    WordCount.run()
```

```
Writing worcount.py
```

```
!python worcount.py textfile.txt
```

```
"our"   2
"out"   3
"passed"        1
"people"        4
"pretty"        1
"problem"       2
"process"       2
"professionals" 1
"pull"  1
"pulling"       1
"really"        1
"refer" 1
"relying"       1
"said"  1
"same"  2
"says"  1
"server"        1
"servers"       1
"should"        1
"skills"        1
"smart" 2
"so"    2
"solution"      3
"solve" 1
"sources"       1
"speed" 1
"spread"        1
"store" 2
"storing"       1
"story" 2
"suggested"     1
Removing temp directory /tmp/worcount.root.20230821.171930.418402...
```

```
%%writefile WordStartCount.py
from mrjob.job import MRJob
import re
class WordStartCount(MRJob):

    def mapper(self, _, line):
        words = re.findall(r'\w+', line.lower())
        for word in words:
            for letter in word:
                yield letter, 1

    def combiner(self, letter, counts):
        yield letter, sum(counts)

    def reducer(self, letter, counts):
        yield letter, sum(counts)

if __name__ == '__main__':
    WordStartCount.run()
```

```
Writing WordStartCount.py
```

```
!python WordStartCount.py textfile.txt
```

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/WordStartCount.root.20230821.172339.700214
Running step 1 of 1...
job output is in /tmp/WordStartCount.root.20230821.172339.700214/output
Streaming final output from /tmp/WordStartCount.root.20230821.172339.700214/output...
"1"     1
"4"     2
"5"     1
"a"     172
"b"     41
"c"     43
"d"     63
"e"     171
"f"     27
"g"     59
"h"     78
"i"     134
"k"     12
"l"     79
"m"     30
"u"     46
"v"     22
"w"     33
```

```
"x"     3
"y"     27
"z"     1
"n"     88
"o"     146
"p"     32
"r"     85
"s"     112
"t"     189
Removing temp directory /tmp/WordStartCount.root.20230821.172339.700214...
```

```
%%writefile WordCountLength5.py
from mrjob.job import MRJob

class WordCountLength5(MRJob):

    def mapper(self, _, line):
        # Split the line into words
        words = line.split()

        # Emit each word as key and 1 as value if its length is 5
        for word in words:
            if len(word) == 5:
                yield word, 1

    def reducer(self, key, values):
        # Sum up the counts for each word
        yield key, sum(values)

if __name__ == '__main__':
    WordCountLength5.run()
```

```
Writing WordCountLength5.py
```

```
!python WordCountLength5.py textfile.txt
```

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/WordCountLength5.root.20230821.173358.775454
Running step 1 of 1...
job output is in /tmp/WordCountLength5.root.20230821.173358.775454/output
Streaming final output from /tmp/WordCountLength5.root.20230821.173358.775454/output...
"Below" 1
"Data," 3
"Data." 3
"Data?" 1
"Story" 2
"Types" 1
"about" 1
"along" 1
"blue," 1
"carry" 1
"cart," 2
"cart." 1
"check" 1
"cover" 1
"days," 1
"don\u2019t"    1
"fella" 1
"groom" 1
"have." 1
"heard" 1
"horse" 4
"think" 4
"towns" 1
"which" 3
"large" 1
"lives" 1
"more," 1
"okay." 1
"other" 1
"refer" 1
"said," 1
"says," 1
"smart" 2
"solve" 1
"speed" 1
"store" 2
Removing temp directory /tmp/WordCountLength5.root.20230821.173358.775454...
```

✓ 0s completed at 10:33 PM