mid

Week 2: Word Count Program: Using Map Reduce (Python)

1. vi ri.txt

**Content:**

Hello

Hi

Good morning

Hello

Hi

Hi

1. sudo nano mapper.py / vi mapper.py

**Content :**

#!/usr/bin/python3

"mapper.py"

import sys

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:

        print ('%s\t%s' % (word, 1))

1. sudo nano reducer.py

Content: #!/usr/bin/python3

"reducer.py"

import sys

current\_word = None

current\_count = 0

for line in sys.stdin:

    # remove leading and trailing whitespaces

    line = line.strip()

    # parse the input we got from mapper.py

    word, count = line.split('\t')

    count = int(coun

    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))

1. TO run word count program locally use following 4 and 5 commands
2. cat ri.txt |python3 mapper.py
3. cat ri.txt |python3 mapper.py |sort |python reducer.py python
4. TO run word count program on Hadoop framework use following command:
5. If the above 7th command runs successfully: Then in local host -browse utilities-specific folder : 2 files will get created (status: successful and Part-00000)
6. To view that : hdfs dfs -ls/bda/output
7. To display : hdfs dfs -cat /bda/output/part-00000
8. Through local
9. Hdfs dfs -get /bda/output/part-00000 /home/hdoop
10. Cat part-0000