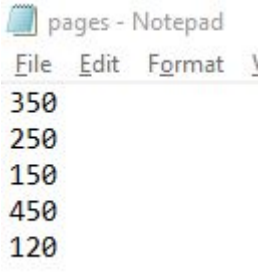
BIG DATA ANALYTICS
LAB4

AIM: Write a map-reduce program to count the frequencies of words from a distributed storage source and understand the phases involved in map-reduce programming.

EXERCISE:

Step 1.1: Create a file named 'pages.txt' in the local file system. Store line by line content as shown below. Each line data represents the number of pages of a sample book.
350 250 150 450 120



Step 1.2: Put the file from the local file system to hdfs with a folder named 'input'. Confirm the presence of above data.

```
C:\>hdfs dfs -mkdir /input

C:\>hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - jak78  supergroup          0 2020-08-05 00:00 /input
drwxr-xr-x   - dr.who supergroup          0 2020-07-24 18:32 /test
```

```
C:\>hdfs dfs -put C:\pages.txt /input/

C:\>hdfs dfs -cat /input/pages.txt
350
250
150
450
120
```

Step 1.3: Write a map and reduce functions to split the books into the following two categories: (a) Big Books (b) Small Books

Books which have more than 300 pages should be in the big book category. Books which have less than 300 pages should be in the small book category. Count the number of books in each category. Store the output as follows as result file within hdfs 'output' folder.

Book Category	Count of the books
"Big Books"	2
"Small Books"	3

Solution:

```
C:\>javac BookCount.java -cp "C:\Users\jak78\Desktop\hadoop-3.1.0\share\hadoop\mapreduce\hadoop-mapreduce-client-
core-3.1.0.jar";"C:\Users\jak78\Desktop\hadoop-3.1.0\share\hadoop\common\hadoop-common-3.1.0.jar"

C:\>jar cf bc.jar BookCount*.class
```

```
C:\>hadoop jar C:\bc.jar BookCount /input/pages.txt /output/
2020-08-02 12:38:36,054 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2020-08-02 12:38:38,087 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool inte
rface and execute your application with ToolRunner to remedy this.
2020-08-02 12:38:38,147 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/jak78/.stagin
g/job_1596350248124_0001
2020-08-02 12:38:38,717 INFO input.FileInputFormat: Total input files to process : 1
2020-08-02 12:38:39,923 INFO mapreduce.JobSubmitter: number of splits:1
2020-08-02 12:38:40,224 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead,
 use yarn.system-metrics-publisher.enabled
2020-08-02 12:38:40,595 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1596350248124_0001
2020-08-02 12:38:40,623 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-08-02 12:38:41,451 INFO conf.Configuration: resource-types.xml not found
2020-08-02 12:38:41,453 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2020-08-02 12:38:43,290 INFO impl.YarnClientImpl: Submitted application application_1596350248124_0001
2020-08-02 12:38:43,435 INFO mapreduce.Job: The url to track the job: http://DESKTOP-OOQJGQT:8088/proxy/application_1596350248124_000
1/
2020-08-02 12:38:43,436 INFO mapreduce.Job: Running job: job_1596350248124_0001
2020-08-02 12:39:21,797 INFO mapreduce.Job: Job job_1596350248124_0001 running in uber mode : false
2020-08-02 12:39:21,816 INFO mapreduce.Job:  map 0% reduce 0%
2020-08-02 12:39:44,720 INFO mapreduce.Job:  map 100% reduce 0%
2020-08-02 12:40:06,347 INFO mapreduce.Job:  map 100% reduce 100%
2020-08-02 12:40:09,457 INFO mapreduce.Job: Job job_1596350248124_0001 completed successfully
2020-08-02 12:40:10,107 INFO mapreduce.Job: Counters: 53
        File System Counters
                FILE: Number of bytes read=40
                FILE: Number of bytes written=427345
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=125
                HDFS: Number of bytes written=26
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=19324
                Total time spent by all reduces in occupied slots (ms)=18995
                Total time spent by all map tasks (ms)=19324
                Total time spent by all reduce tasks (ms)=18995
```

```
        Total megabyte-milliseconds taken by all map tasks=19787776
        Total megabyte-milliseconds taken by all reduce tasks=19450880
    Map-Reduce Framework
        Map input records=5
        Map output records=5
        Map output bytes=76
        Map output materialized bytes=40
        Input split bytes=102
        Combine input records=5
        Combine output records=2
        Reduce input groups=2
        Reduce shuffle bytes=40
        Reduce input records=2
        Reduce output records=2
        Spilled Records=4
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=203
        CPU time spent (ms)=3651
        Physical memory (bytes) snapshot=542892032
        Virtual memory (bytes) snapshot=776298496
        Total committed heap usage (bytes)=349175808
        Peak Map Physical memory (bytes)=292261888
        Peak Map Virtual memory (bytes)=382844928
        Peak Reduce Physical memory (bytes)=250630144
        Peak Reduce Virtual memory (bytes)=393453568
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=23
    File Output Format Counters
        Bytes Written=26
```

```
C:\>hadoop dfs -cat /output/part-r-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
Big Books       2
Small Books     3
```