**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**

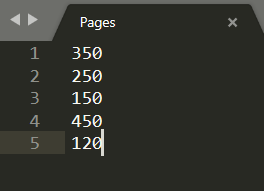
**250**

**150**

**450**

**120**

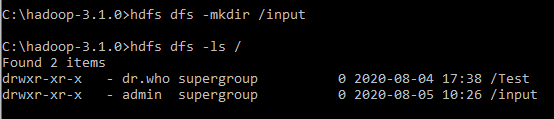
* The file ‘**Pages.txt**’ is created in local file system and the above data is stored line by line.



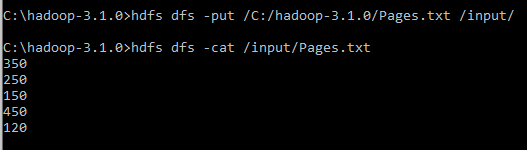
**STEP 1.2:**

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

* First, we will create ‘input’ directory using **mkdir** command.



* Using the **put** command we will put the file ‘Pages.txt’ in the ‘input’ folder to hdfs from local file system.



**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**

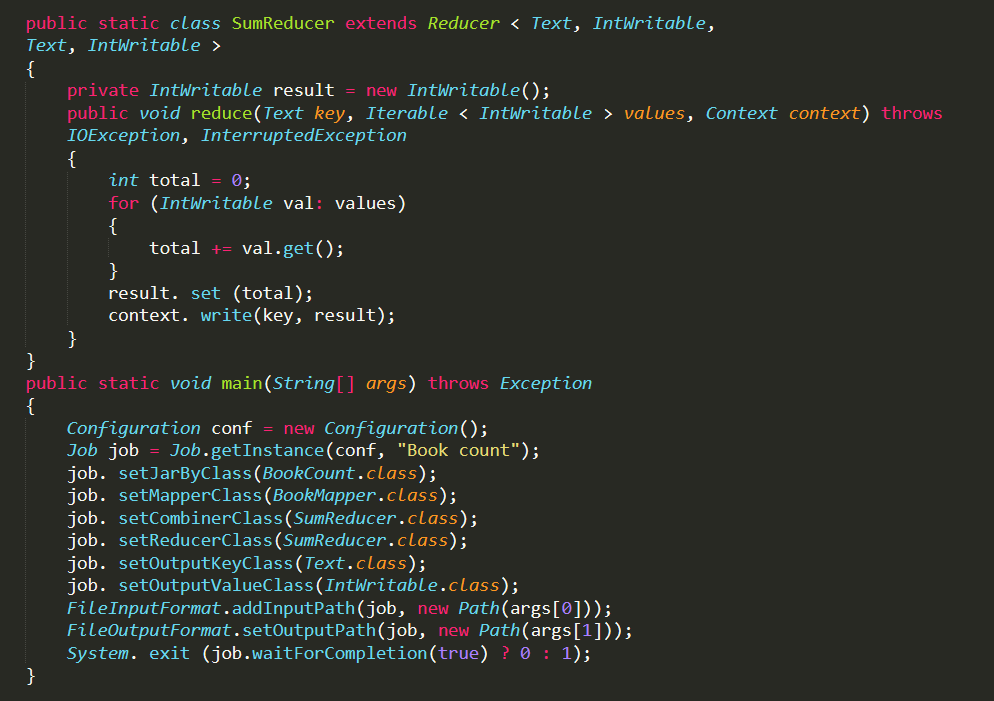
**"Big Books" 2**

**"Small Books" 3**

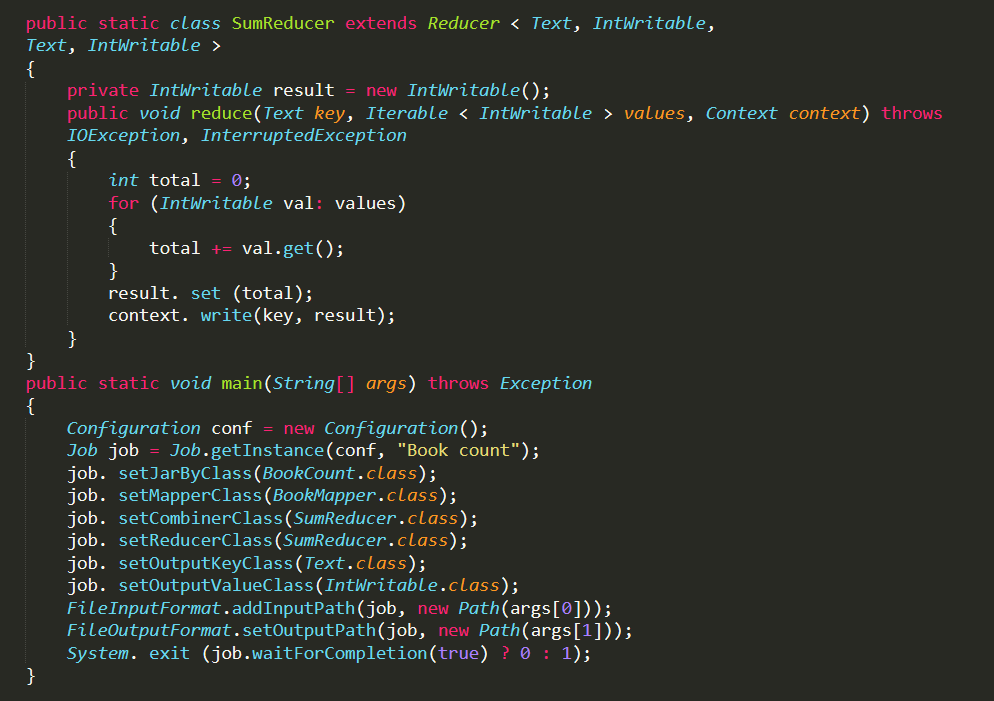
* map function

****

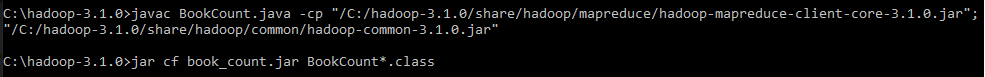
* reduce function

****

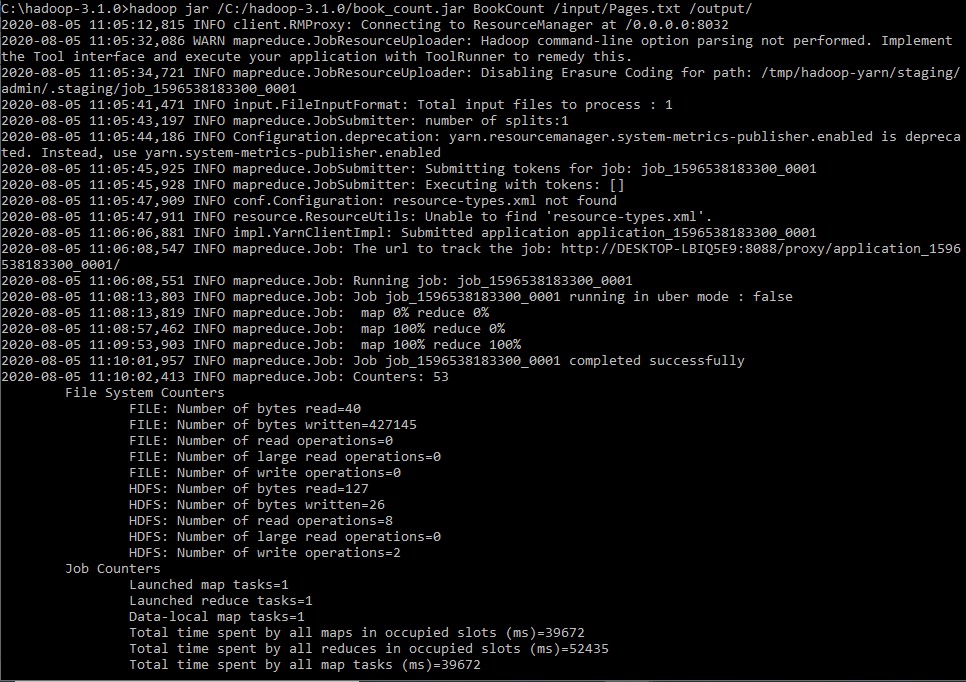
* Driver Function



* Compile the above file using command prompt as below which will generate 2 class files for map and reduce functions and a jar file ‘**book\_count.jar**’.



* Store the output in the hdfs ‘**output**’ folder using **hadoop jar** command.



* Output:

