Big Data Computing

Assignment-VI

October 6, 2020

A. Word Count

Run the Wordcount program in **Apache Spark**. The source code of the wordcount program is provided which you need to show it running on a sample dataset. Run the program with the training dataset provided

Input: An unstructured text file with multiple lines

Output: Number of occurrences of each word, appearing in the text file.

Input: *input.txt*

IIT Patna is a great place to ready great lessons about Java, Big Data, Python and many more Programming languages. Big Data lessons are difficult to find but at Big Data Computing Lab, you can find some excellent pieces of lessons written on Big Data.

Output: (ready,1), (are,1), (excellent,1), (languages.,1), (find,2), (Python,1), (is,1), (you,1), (about,1), (can,1), (a,1), (on,1), (IIT,1), (Big,4), (many,1), (lessons,3), (some,1), (Java,,1), (to,2), (written,1), (at,1), (Lab,,1), (more,1), (pieces,1), (of,1), (place,1), (Data.,1), (Data,,1), (great,2), (difficult,1), (but,1), (and,1), (Data,2), (Patna,1), (Computing,1), (Programming,1)

B. Web traffic analysis

one of the most common uses of Apache Spark is analyzing and processing log files. In this question you are provided with a large text file containing the log of a web server. Each line of the file is associated with an URL request. Now write a program using apache spark to output the list of distinct IP addresses associated with the connections to a “google” page written onto a HDFS file using the new dataset provided in the link below.

Dataset: http://www.almhuette-raith.at/apache-log/access.log

Execution command: spark-submit --class WordCount --master local target/GVexample-1.0.0.jar /home/iitp/spark/wordcount/output