CSCE 5300 Introduction to Big data and Data Science
ICE-2

Lesson Title: Hadoop MapReduce and Hadoop Distributed File System (HDFS)

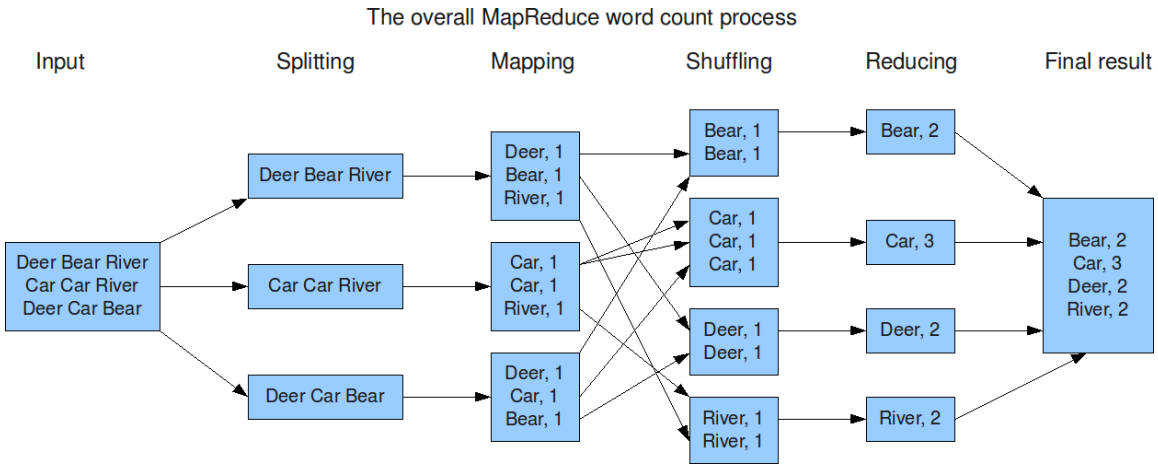
Lesson Description: In this lesson, we are going to discuss about Hadoop MapReduce and Hadoop Distributed File System (HDFS)

In class exercise

There are many ways to execute wordcount program:

1. **Using any IDE like Intellij or Eclipse**
2. **Run on hadoop clusters**

Use case Description:



1. Using the MapReduce method, count the number of words in the input. Output the words whose frequencies are a multiple of 3.

   Refer the following link for step-by-step explanation of wordcount program runs on a single node cluster.

   https://github.com/chenmiao/Big_Data_Analytics_Web_Text/wiki/Hadoop-with-Cloudera-VM-(the-Word-Count-Example)

2. Counting the frequency of words in the given text file that end with the letter ‘s’.

Refer following example:

Input text:
ashes
eventual
employers
icicle
onto
users
ashes

Output :
ashes 2
employers 1
users 1

3. Using the wordcount, make changes where necessary and try to implement a program where you can print the words whose frequency is a prime number. A prime number is any number that is divisible by itself and 1.

Input text:
Input:

Ashes
Eventual
Employers
Icicle
employers
Onto
Icicle
Ashes
Ashes

Output:
Ashes 3
Eventual 1
Onto 1

Marks will be distributed between logic, implementation and output

Programming elements:
Hadoop MapReduce and HDFS

Source Code: Given in canvas.

ICE Submission Guidelines

1. ICE Submission is individual.
2. ICE code has to be properly commented.
3. The documentation should include the screenshots of your code/results.
4. Provide the explanation of the exercise as per your understanding.
5. The similarity score for your document should be less than 15%.
6. Submit the source code (properly commented) and documentation (.pdf/.doc) with explanation and screenshot of source code having input logic and output results.
7. Submission after the deadline is considered as late submission.