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

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

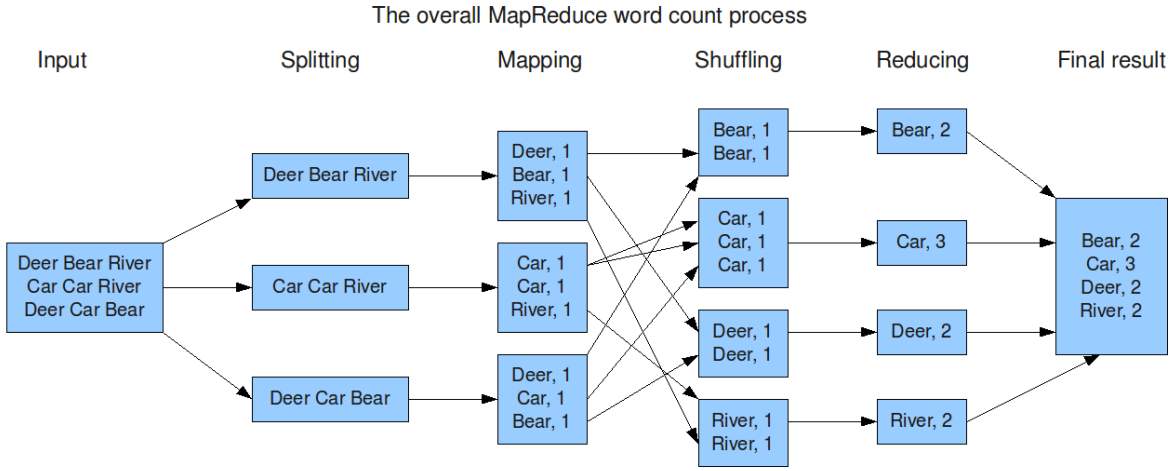
Lesson Description: In this lesson, we are going to discuss about Hadoop MapReduce and Hadoop Distributed File System (HDFS)

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. Counting the frequency of words in the given input with MapReduce algorithm

Use the given text file in canvas to count the frequency of words.

Refer the following link for step-by-step explanation of wordcount program runs on 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 given text file that starts with letter ‘b’

Refer following example:

Input text: base
base
boy
bug
time
cat

Output: base 2
boy 1
bug 1

3. Determine the prime number in input and print number only once

Input:
2
3
3
7
7
6
8

Output:
2 0
3 0
7 0
6 1
8 1

0 -> Prime
1 -> Not Prime

4.determine the odd number of input

Input:
2
11
14
32
33

Output:
2 0
11 1
14 0

32 0
33 1

0->not odd
1->odd

Marks will be distributed between logic, implementation and UI

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.