CSEE5590-0004/490-0004: Big Data Programming

Lesson Plan # 2

ICP Feedback and Submission Link :

https://forms.gle/xMAmr3zATrtMG5cX7

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)

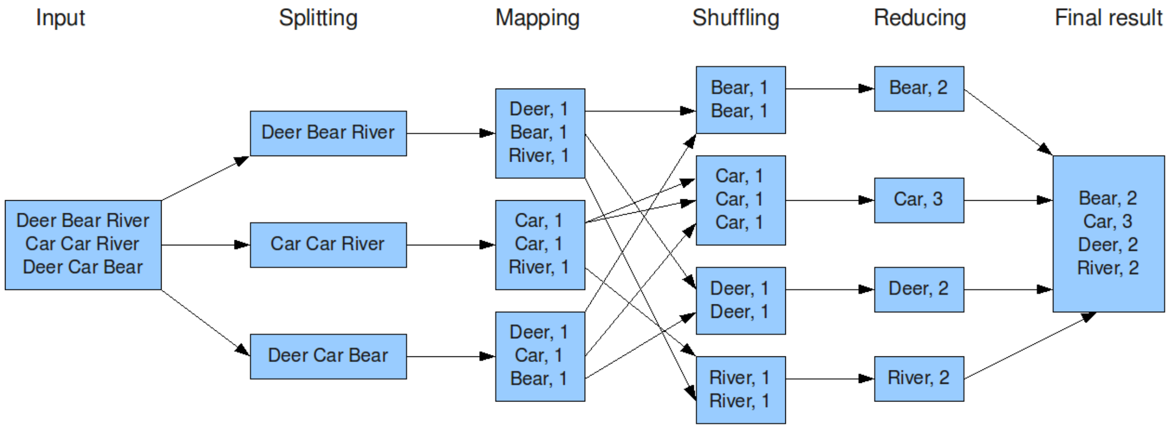
In class exercise

There are many ways to execute wordcount program:

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

Use case Description:

The overall MapReduce word count process



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

Use the following text file to count the frequency of words.

https://umkc.box.com/s/r4jtmjnoip7g0q8tzyqb2naa78u50t3c

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 ‘a’

Use the text file as input.

https://umkc.box.com/s/r4jtmjnoip7g0q8tzyqb2naa78u50t3c

Refer following example:

Input text: ashes
time
cat
add
add
amazing

Output: ashes 1
add 2
amazing 1

Bonus Question:

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

Marks will be distributed between logic, implementation and UI

Programming elements:
Hadoop MapReduce and HDFS

Prerequisites:
Ensure that Hadoop is installed, configured and is running. More details:
Single Node Setup for first-time users.
Cluster Setup for large, distributed clusters**.**
You can even execute program on any IDE like eclipse or Intellij.

After completion of your ICP fill in the form. Any available TA/instructor will come to you and evaluate ICP

ICP Submission Guidelines:

1. ICP Submission is in pairs of four students.
2. Once completed, must be presented to TA or Instructor before the completion of the class
3. Submission after the deadline is considered as late submission. (Check the late submission policy in the syllabus)
4. ICP Code with brief explanation should be pushed to GitHub.

5. Submit a demo video 2-3 min showing your assignment with a voice over explaining your work if you are unable to complete ICP within the deadline due to genuine reason.
6. Provide the video submission link through the submission form https://forms.gle/xMAmr3zATrtMG5cX7

Cheating, plagiarism, disruptive behavior and other forms of unacceptable conduct are subject to strong sanctions in accordance with university policy. See detailed description of university policy at the following URL: *https://catalog.umkc.edu/special-notices/academic-honesty/*