CSEE5590-0001/490-0003: Big Data Programming

Lesson Plan #8

ICP Feedback and Submission Link :

https://forms.gle/xMAmr3zATrtMG5cX7

Lesson Title: *Apache Spark*

Lesson Description: *Apache Spark Introduction*

Lesson Overview:

Apache Spark is a unified analytics engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing.

Installation:

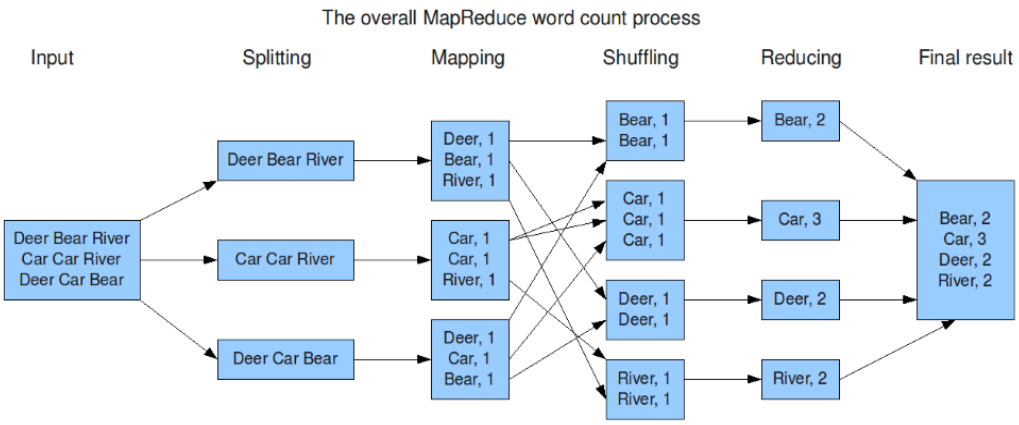
http://allaboutscala.com/tutorials/chapter-1-getting-familiar-intellij-ide/scala-tutorial-first-hello-world-application/

In class exercise:

1. Spark Programming:

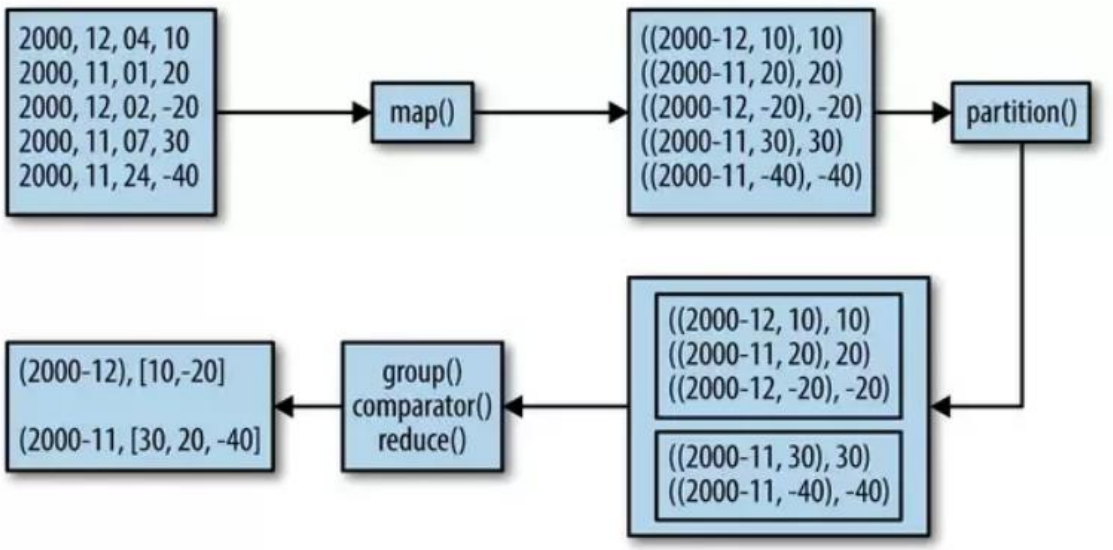
Write a spark program with an interesting use case using text data as the input and program should have at least Two Spark Transformations and Two Spark Actions.

Present your use case in map reduce paradigm as shown below (for word count).



2. Secondary Sorting in Map Reduce

Secondary sorting is used to sort the values in the reducer phase.
Take any input of your interest and perform secondary sorting on it.



Real Time Applications:

Useful with time series data

Reference Links:

https://stdatalabs.com/2017/02/mapreduce-vs-spark-secondary-sor/
https://www.quora.com/What-is-secondary-sort-in-Hadoop-and-how-does-it-work
https://www.oreilly.com/library/view/data-algorithms/9781491906170/ch01.html

Partitions:
https://www.ibm.com/support/knowledgecenter/en/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/rangepartitioner.html

ICP Submission Guidelines:

1. ICP Submission is individual however, it can be completed as a Team during session.
2. If completed, should 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 your screenshots as well to GitHub and documentation. The screenshot should have both the code and the output.
6. 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.

7. Provide the video submission link through the GitHub and 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/*