**CSEE5590-0001/490-0003: Big Data Programming**

**Lesson Plan #1**

**ICP Feedback and Submission Link :**

<https://docs.google.com/forms/d/e/1FAIpQLScaeLhVS92maF9Z9exB-5dw2PdPAiljbAeb92zb3jAe9hqVHg/viewform?usp=sf_link>

**For Online students:** [**https://docs.google.com/forms/d/e/1FAIpQLScaeLhVS92maF9Z9exB-5dw2PdPAiljbAeb92zb3jAe9hqVHg/viewform?usp=sf\_link**](https://docs.google.com/forms/d/e/1FAIpQLScaeLhVS92maF9Z9exB-5dw2PdPAiljbAeb92zb3jAe9hqVHg/viewform?usp=sf_link)

**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:**

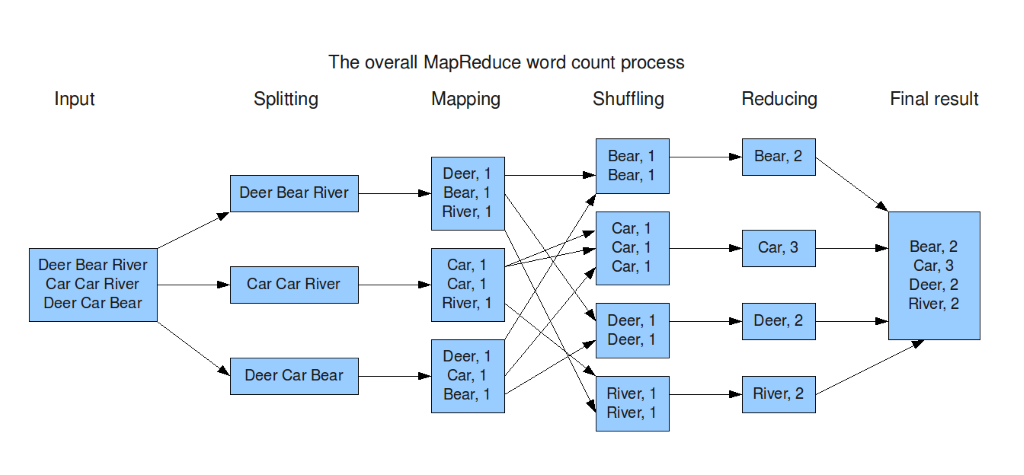
<http://allaboutscala.com/tutorials/chapter-1-getting-familiar-intellij-ide/scala-tutorial-first-hello-world-application/>

**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.**

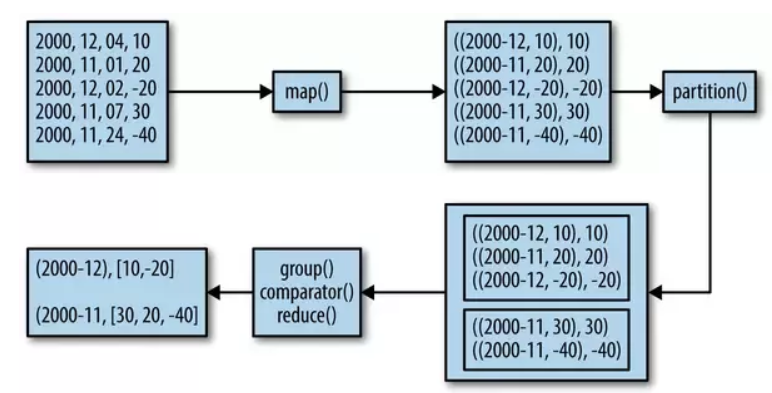
**Present your use case in map reduce paradigm as shown below (for word count).**



1. **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 Guidelines (In Class Students):**

1. ICP Submission is in pairs of two students.
2. Once completed, must be presented to TA or Instructor before the completion of the class
3. Submission after class is considered as late submission. (Check the late submission policy in the syllabus)
4. ICP Code with brief explanation in wiki should be pushed to GitHub (submit your screenshots as well. The screenshot should have both the code and the output)Submit GitHub Link through the Feedback Form [(https://docs.google.com/forms/d/e/1FAIpQLScaeLhVS92maF9Z9exB-5dw2PdPAiljbAeb92zb3jAe9hqVHg/viewform?usp=sf\_link)](https://docs.google.com/forms/d/e/1FAIpQLScaeLhVS92maF9Z9exB-5dw2PdPAiljbAeb92zb3jAe9hqVHg/viewform?usp=sf_link)

**Submission Guidelines (for online students):**

1. Submit your source code and documentation to GitHub and represent the work through wiki page properly with detailed explanation (submit your screenshots as well. The screenshot should have both the code and the output)
2. Comment your code appropriately.
3. Submit a brief demo video 2-3 min showing your assignment with a voice over explaining your work through the Submission Link.
4. Use the following Google link to submit your assignment

(ICP Submission Link#): <https://docs.google.com/forms/d/e/1FAIpQLScaeLhVS92maF9Z9exB-5dw2PdPAiljbAeb92zb3jAe9hqVHg/viewform?usp=sf_link>

***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/*](https://catalog.umkc.edu/special-notices/academic-honesty/)