**DAKOTA STATE UNIVERSITY INFS 774:**

**Big Data Analytics Guidelines for Week Six**

**6/24/20-7/1/20**

Description: This week’s topics include Yarn (optional) and Apache Spark (required)

Yarn, a redesigned resource manager implemented in Hadoop version 2, eliminates Hadoop's sole reliance on the MapReduce programming model. As a result, since 2014, we have been witnessing the rapid development of a new big data processing technology called Apache Spark that does not rely upon the MapReduce model. You can install and run Apache Spark on a computer cluster in a standalone mode or on top of Hadoop. Many people believe that Apache Spark is going to replace and eliminate Hadoop MapReduce (but probably not the whole Hadoop Ecosystem) (e.g., <http://aptuz.com/blog/is-apache-spark-going-to-replace-hadoop/>). The article has <https://sdtimes.com/big-data/is-spark-replacing-hadoop/> more detailed discussion on this issue. This article <https://www.intellectsoft.net/blog/spark-vs-hadoop/> has detailed comparison between Hadoop and Spark. As many companies have already heavily invested in Hadoop, and these two big data systems will co-exist in the foreseeable future, you want to know both well, in order to find a Big Data scientist job.

Relevant Learning Objectives:

* Optional: Gain an overview of Yarn, the data operating system in Hadoop.
* Gain an overview of Apache Spark

Required activities:

Watch Dr. Liu’s week 6 lecture on Apache Spark

Optional activities:

1. Watch the Hadoop2.0 and Yarn tutorial by Edureka. <https://www.youtube.com/watch?v=uncKbeS5MqA>
2. Watch the Lynda.com course “Apache Spark Essential Training” by Ben Sullins