Wyatt Blair

EMT-678-WS: Big Data Technologies

10/3/2024

**Class 3—Spark ACM Communications 2016**

Spark is a Programming Model which extends the MapReduce paradigm. Its main improvement is using RDDs (Resilient Distributed Datasets) which provides significant computational improvements. For a sense of scale, the authors of the paper point to a simple logistic regression implementation in Spark compared to an equivalent implementation in MapReduce. On a 100GB job MapReduce takes 110 seconds, whereas Spark takes only *one second* per iteration after the initial load time.

Beyond significant runtime improvements, Spark also allows for the use of additional libraries such as Spark Streaming, GraphX, and MLlib. These libraries leverage Spark and allow users to access the tremendous speed ups for their specific problem without having to code everything from scratch.

Spark has been used for several applications such as interactive queries, stream processing, and others. But the most interesting use case described in the paper (in my opinion) are the scientific applications. Spark has been used to stream neuroscience data *in real-time*; scaling up to 1 TB / hour of whole brain-imaging data from small organisms. Such uses indicate to me that Spark not only is a leap forward for the industry surrounding Big Data, but also for the realm of possibilities in science as a whole.