Report:

Implementation of HDFS and Spark over AWS EC2 Instances

Author: Dominick DeCanio

As seen in the images contained in the folders for tasks 1-3 the application time to completion was shortest for task 2, longer for task 3, and longest for task 1. We can see from the DAG diagrams that task 1 had many more stages than tasks 2 and 3. It is also evident that the number of task per stage were dramatically higher for task 1 than for the other tasks.

From these execution statistics we can see that the added partitioning of steps 2 and 3 speed up the execution time of the application because the application is able to leverage the parallel execution structure available in spark. Because the application is executed in parallel across multiple partitions, each partition needs to complete fewer tasks to complete the application. This helps explain the speed of execution of steps 2 and 3. In task 3, despite the killing of the worker, the completion time is relatively fast due to the automatic rescheduling of the failed tasks and the increased execution speed at the beginning of the run.