Caleb Post
Hadoop Lab 2

Question 1: Explain the data flow of MapReduce Job in Yarn. See pages 195 – 202 in the book. Please don't skim on detail and the explanation should be in your own words.

Whenever a new MapReduce job is requested in YARN there are many steps that the process must go through before it is complete. The steps are below.

1. MapReduce program is run on the client node inside a new JVM.
2. Once the new program is running a new application is gotten from the resource manager. This is the id for the new program.
3. Job resources are copied over to HDFS. This is the calculation of the input splits and moving over the job resources such as the JAR and configuration to HDFS.
4. Job and application is submitted. At this point the resources to run the job are all copied over to HDFS and nodes need to be allocated in order to run the process.
5. Container is started on node. The scheduler allocates a new container and the resource manager launches the application master's process there. The application master is under the node manager's management.
6. Input splits are retrieved from the file system as previously calculated by the client. A map task object is created for each split and the number of reduce tasks are also created as determined by the property.
7. Request is made from the Resource manager for each of the reduce tasks. These requests are made in the form of a container. The container is a request for a specified amount of memory.
8. Container is started on node and task is launched in its own JVM.
9. Job resources for the job are retrieved from HDFS.
10. Map Task is run.

Progress is given by the App Master. The app master aggregates the progress of all the nodes and tasks for the map reduce program.