This exam serves as the assessment for those students who cannot utilize the Hadoop system and/or Ambari GUI.

1. Describe the process of MapReduce.

Since MapReduce is a two-step process, the work will be split into two parts.

The Map Results will equal one input for every one output. The map step will process the "Arrest" column in the crimes-samples.csv file and produce a list as shown below, where "Yes" means an arrest occurred and "No" means there was not an arrest. Then we can reduce or aggregate the entries into a simpler representation which means each line is processed and the values for each key are combined.

2. What is the difference between schema on read and schema on write?
    a. **On read applies structure when the data is being processed, on write doesn't A**
    b. On read applies structure all the time, on write doesn't

"Hive uses a schema on read format that takes unstructured data and only applies structure to it when it is being read and processed." Page 3

3. True or False? "You can use SQL in Hive."
    a. **True A**
    b. False

4. Why do you think sqoop would be a useful tool?

Sqoop is a program that allows you to use Hadoop to run queries on databases that are stored in MySQL, allowing you to integrate the processing power of Hadoop with your already created database systems. The purpose is to interface existing database connections with Hadoop. It is used to import data from RDBMS to Hadoop and export data from Hadoop to RDBMS. This way we don't have to reinput all of the data in a different way.