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

1. Explain the concept of TEZ and how it works.

TEZ utilizes *directed acyclic graphs (DAG)* to plot the most efficient route to do the work assigned to it. By graphing it all out, we can eliminate unnecessary steps and dependencies, and can even run things at the same time. This means that when we use TEZ, we are optimizing the flow of data, and it can act as a resource manager as well.

1. What is Pig’s version of the for loop?
   1. Group
   2. Filter
   3. Join
   4. For each / generate
2. How is HBase structured? Describe or draw – whatever makes the most sense to you.

*HBase* is a non-relational, transactional database built on top of HDFS. There is no query language that goes with HDFS, but there is an API that allows you to perform CRUD operations using HBase.

HBase Architecture

HBase has two components to it: *region* servers, which contain groups of keys that can automatically grow and re-partition as needed, and *master servers*, which keep track of everything that's going on with the region servers.

Data Structure

Like other non-relational databases, in HBase, every row is a reference to a unique key. However, HBase has the ability to store data in *column families* instead of just key-value pairs. Each column family can contain multiple related individual columns. For instance, we might have a column family named address, and stored inside are columns for street number, street name, city name, state name, and zip code. HBase excels at dealing with missing data, since it does not take up space within a column if no data populates that column.

HBase also has something called a *cell*, which is the intersection of a row and a column. we can store multiple versions of each cell to get an idea of the history if data is being updated. History can include a timestamp as well, so we know exactly when a change occurred.