Names: Bob Skowron, Jason Walker
Keys: rskowron, jwalker
SVN: jwalker: https://svn.seas.wustl.edu/repositories/jwalker/cse427s_fl17/

1. a.

   b.

   c.

Names: Bob Skowron, Jason Walker
Keys: rskowron, jwalker
SVN: jwalker: https://svn.seas.wustl.edu/repositories/jwalker/cse427s_fl17/

2. a.
   b.
   c.
   d.
   e.
   f.

Names: Bob Skowron, Jason Walker
Keys: rskowron, jwalker
SVN: jwalker: https://svn.seas.wustl.edu/repositories/jwalker/cse427s_fl17/

3. a. See SVN

   b. *spark-submit CountJPGs.py /loudacre/weblogs*
   Number of JPGs: 64,978
   The Driver program and processing are done locally. The result is stored locally.

   c. *spark-submit –master yarn-client CountJPGs.py /loudacre/weblogs*
   The Driver program is run locally. The processing is done on the cluster. The result is stored in HDFS.

   d. 1 stage and 311 tasks were executed

   e. *spark-submit –master yarn-cluster CountJPGs.py /loudacre/weblogs*
   The Driver program and processing are completed on the cluster. The result is stored in HDFS.

Names: Bob Skowron, Jason Walker
Keys: rskowron, jwalker
SVN: jwalker: https://svn.seas.wustl.edu/repositories/jwalker/cse427s_fl17/

4. a. To generate a sample of the data you can use the unix commands head or tail to create a file with a specified subset of data. Using PIG, you could load the data and then write out a sample file with LIMIT and STORE. Both of these would only take the first or last n records. If you wanted to generate a more representative sample, you could use PIG with SAMPLE or unix with shuf.

   It is much faster to test PIG scripts with a local subset since PIG will generate a MapReduce job based on the script. If you run the script on the cluster with the full set of data, this would be equivalent to running an entire MR job on the data which could take a long time.

   b.

   c. (diskcentral.example.com,68)
   (megawave.example.com,96)
   (megasource.example.com,100)
   (salestiger.example.com,141)

   d. (bassoonenthusiast.example.com,1246)
   (grillingtips.example.com,4800)
   (footwear.example.com,4898)
   (cofeenews.example.com,5106)

Names: Bob Skowron, Jason Walker
Keys: rskowron, jwalker
SVN: jwalker: https://svn.seas.wustl.edu/repositories/jwalker/cse427s_fl17/

5. a. Will copy in later
   b. (TABLET,3193033)
   (DUALCORE,2888747)
   (DEAL,2717098)