Zhihao Xue
CM 3357
Extra Credit Week 3
CSSE490 Dr. Mohan
Sun,Sept 27, 2015

- What is the distributed cache?
  Distributed Cache is a Hadoop feature that can boost efficiency when a map or a reduce task needs access to common data. If your cluster depends on existing applications or binaries that are not installed when the cluster is created, you can use Distributed Cache to import these files. This provides a service for copying files and archives to the task nodes in time for the tasks to use them when they run. This feature lets a cluster node read the imported files from its local file system, instead of retrieving the files from other cluster nodes. To save network bandwidth, files are normally copied to any particular node once per job.

- How can you use the distributed cache to do a join efficiently when one of the datasets is small?
  We can use Pig implements the fragment-replicate join by loading the replicated input into Hadoop?s distributed cache.Pig runs a map-only MapReduce job to preprocess the file and get it ready for loading into the distributed cache. If there is a filter or foreach between the load and join, these will be done as part of this initial job so that the file to be stored in the distributed cache is as small as possible. The join itself will be done in a second map-only job.

```
big = LOAD 'big_data' AS (b1,b2,b3);
tiny = LOAD 'tiny_data' AS (t1,t2,t3);
mini = LOAD 'mini_data' AS (m1,m2,m3);
C = JOIN big BY b1, tiny BY t1, mini BY m1 USING 'replicated';
```