Matthew Ruffner

HW7 Write Up

Questions:

1) What is the purpose of the mapper?
   Inside the MapReduce framework, since data is distributed across clusters and stored on different nodes, the mapper does the same computation across all data items then stores the results in key-value pairs. In my specific code, I have two mappers, my first mapper stores the individual salaries as keys and gives them a value of one. My second mapper takes the output from the first reducer and returns a 'None' for the key and a key-value tuple as the value.

2) What is the purpose of the reducer?
   In the MapReduce framework, a reducer will take in the output of a mapper and perform some sort of reduce function based on the specific key the mapper gave and perform an aggregation on them. In my code, I have two reducers, the first reducer takes the salaries (key) that occur multiple times and groups them together. The second reducer takes the tuples (value) from the output of the second mapper and takes the all the salaries and adds them to a list. Next it finds the max current max value of the list, adds it to another list then removes that value from the previous list. It performs this operation 10 times to get a top ten list.

3) How does distributed computing fit into the MR context?
   MapReduce uses distributed computing to split up programs to be executed on a large cluster of commodity machines in a parallel manner. The data is distributed across the cluster and stored on different nodes. A distributed computing system has the benefit that it protects data by replicating data so that if one node fails then the data isn't lost for good.

Matthew Ruffner

```
Matthews-MacBook-Pro:hw7 ruffner$ python count.py bsalaries1.txt bsalaries2.txt bsa
laries3.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /var/folders/1y/m5bzsfbd60q1m82lxfyqyzm00000gn/T/count.ruff
ner.20201114.032606.595904
Running step 1 of 2...
Running step 2 of 2...
job output is in /var/folders/1y/m5bzsfbd60q1m82lxfyqyzm00000gn/T/count.ruffner.202
01114.032606.595904/output
Streaming final output from /var/folders/1y/m5bzsfbd60q1m82lxfyqyzm00000gn/T/count.
ruffner.20201114.032606.595904/output...
80431   1
66551   1
66111   1
56781   1
48201   1
47651   1
31911   1
28431   1
27831   1
16931   1
Removing temp directory /var/folders/1y/m5bzsfbd60q1m82lxfyqyzm00000gn/T/count.ruff
ner.20201114.032606.595904...
Matthews-MacBook-Pro:hw7 ruffner$
```