Lab 10 Hadoop

In this lab, we explored the use of Hadoop distributed file system and MapRecuce for two basic operations. The generic word count example, and a more complicated taxi mapping problem. We learned how to list files in an hdfs directory, create new directories, and delete existing files/directories. The word count was run and outputted to result.txt shown below in Figure 1. The format is “word, number of instances”. This code works using to functions, mapper.py and reducer.py. The mapper.py function strips the text of all characters and creates a key-value pair. The key-value pair is “word, 1” with the key being “word” and the value being “1”. Indicating that “word” was found once. The reducer function then combines and aggregates all instances of “word” and adds together the values such that if “word” appears 5 times in the text the result will be “word, 5”. This is map reduce methodology is easily parallelized and can be distributed across several nodes. The individual functions of mapper and reducer can be seen in Figures 2 and 3 respectively.

The second task of this lab was to generate a heat map of taxi pickups in New York during Jan 2013. The output result is shown in Figure 4. The result was generated by making each taxi pick up location a key, with the value 1. Then reducing (ie aggregating) the values together by location. The write2Grid function then organizes the data into a grid fashion that will be easy to rasterize, the output is a “.asc” file. Finally plotTaxi.sh takes the “.asc” output and saves it as a .png file. The data was then visualized and a screen shot was taken. (fig 4)

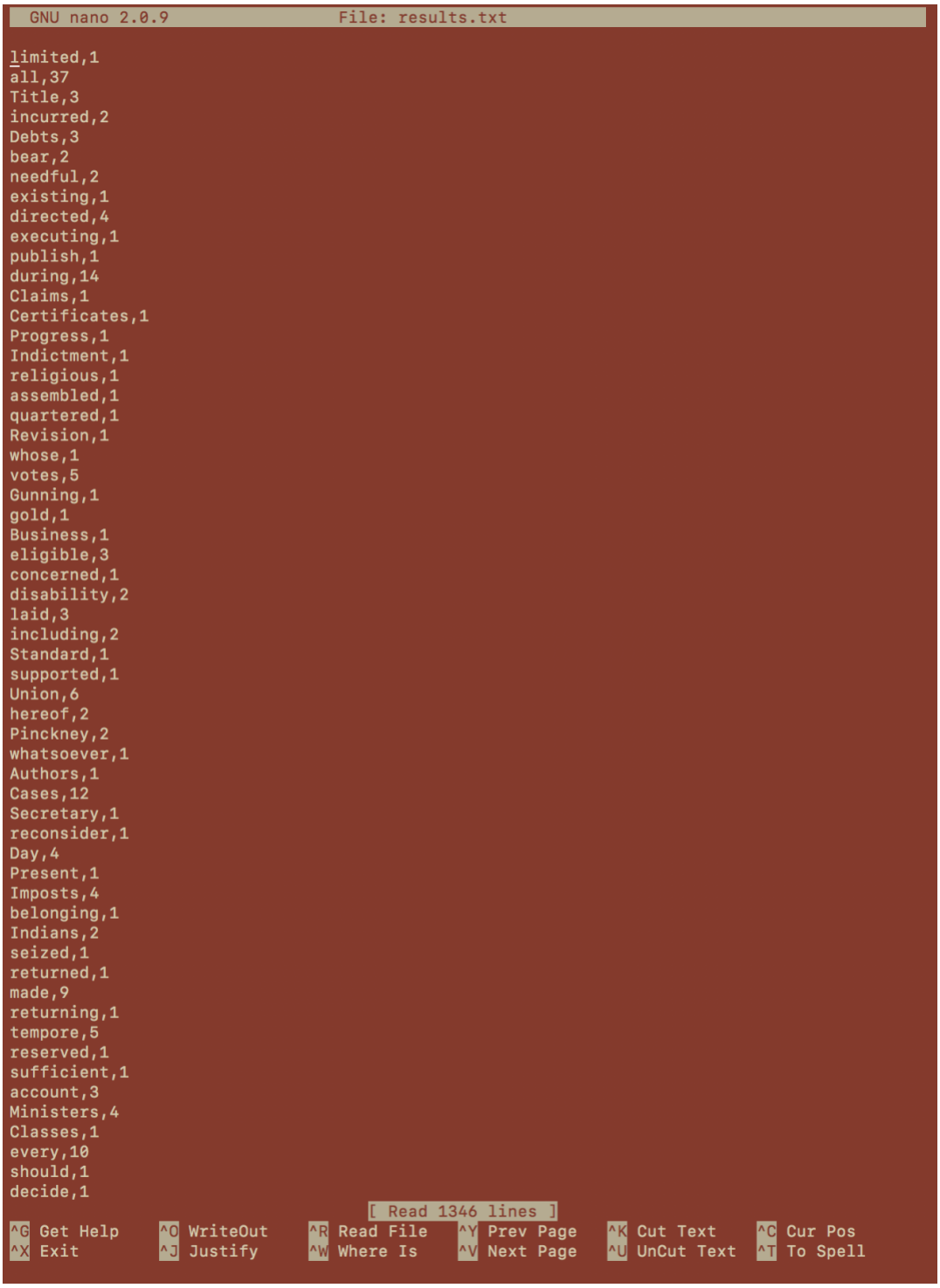

Fig 1: The output of the first word count example.

```
[karoth4@cg-hm08 word_count_hadoop]$ echo "This is a cold day. So it is grea
t that we are sitting inside a not so cold room doing coding. Great not to b
e out in the cold." | /projects/geog479/lab10/word_count_hadoop/mapper.py
This,1
is,1
a,1
cold,1
day,1
So,1
it,1
is,1
great,1
that,1
we,1
are,1
sitting,1
inside,1
a,1
not,1
so,1
cold,1
room,1
doing,1
coding,1
Great,1
not,1
to,1
be,1
out,1
in,1
the,1
cold,1
```

Fig 2: Output of the mapper.

```
[karoth4@cg-hm08 word_count_hadoop]$ echo "This is a cold day. So it is grea
t that we are sitting inside a not so cold room doing coding. Great not to b
e out in the cold." | /projects/geog479/lab10/word_count_hadoop/mapper.py |
/projects/geog479/lab10/word_count_hadoop/reducer.py
doing,1
is,2
coding,1
it,1
are,1
in,1
cold,3
out,1
sitting,1
to,1
be,1
Great,1
that,1
This,1
So,1
not,2
day,1
a,2
great,1
room,1
we,1
inside,1
so,1
the,1
```

Fig 3: Output of the reducer, which was given the mapper output as input.

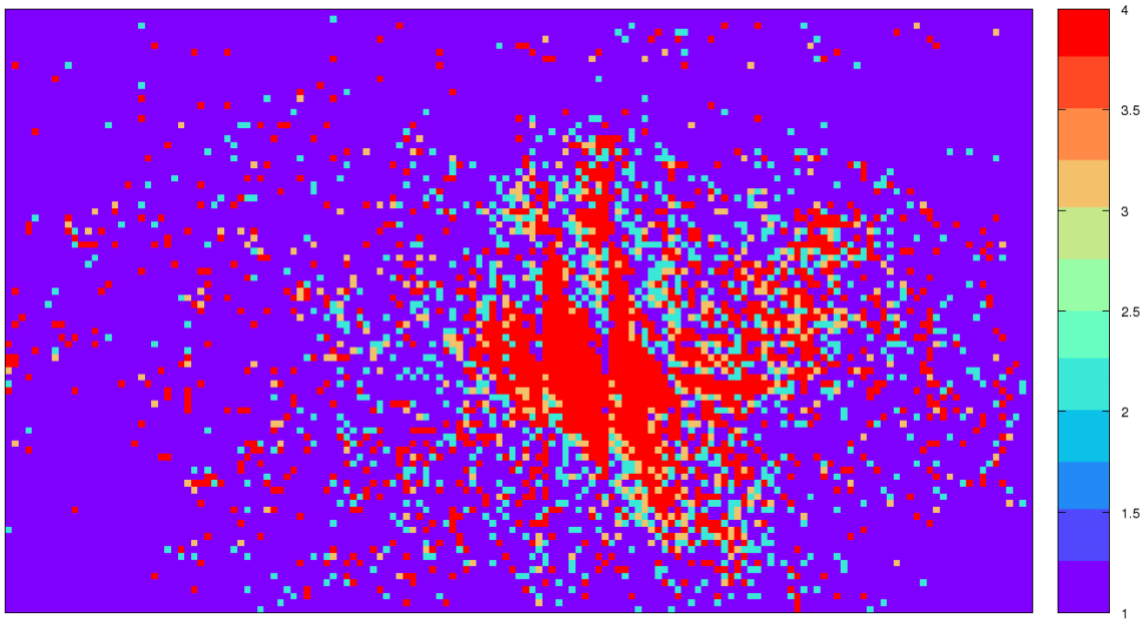

Fig 4. Plot of NY taxi data.