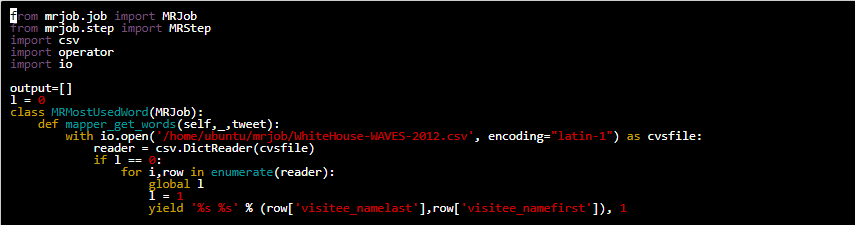
1. **Description of your workflow and program-running environment (try to be as detailed as possible), use screenshots if possible.**

**Running the cluster on local machine:**

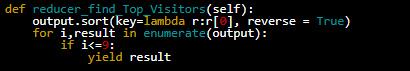
**--------------------------------------------------**

****

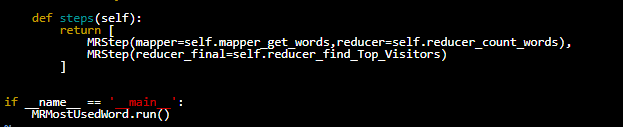
The very first job is the mapper job. The functions of mapper job is assigning the value 1 to each iterative output. Here first we are opening the csv file with encoding format as ‘latin-1’ and then reading using the Dictreader in the form of dictionary. As mapper is iterative and we are using the enumerator operator in for loop to read the column content, it will read the data multiple number of times. So to avoid multiple reads we used simple count logic(l==0) and finally we are assigning 1 to result of each row and returning the result to next reducer job.

Part2.PNG

Next job is the reducer function. Here it will do the aggregation functionality(summation part). It will get the input from the mapper function, and it will do the addition for same unique output. We are appending the the result into output list.



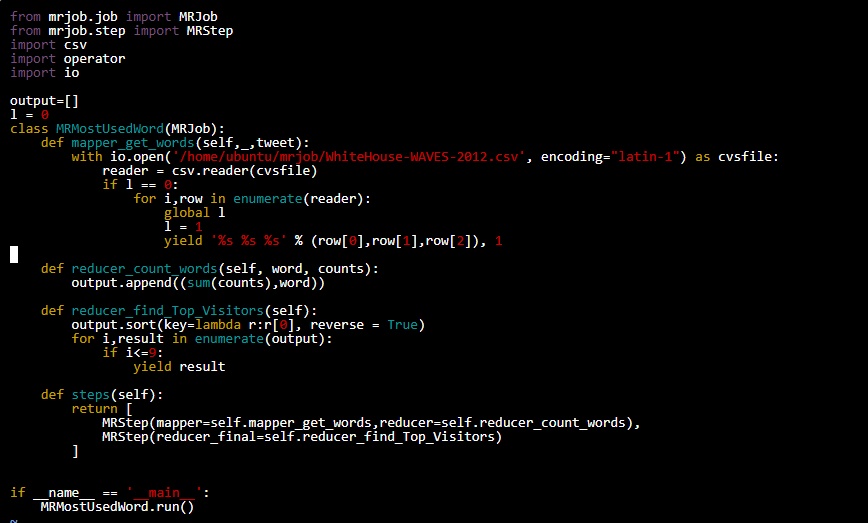
Final job is the reducer\_final function. Here we are sorting the output list in descending order. As here we are trying to get the top ten results, we are filtering the top ten results using for loop and yielding the final result.



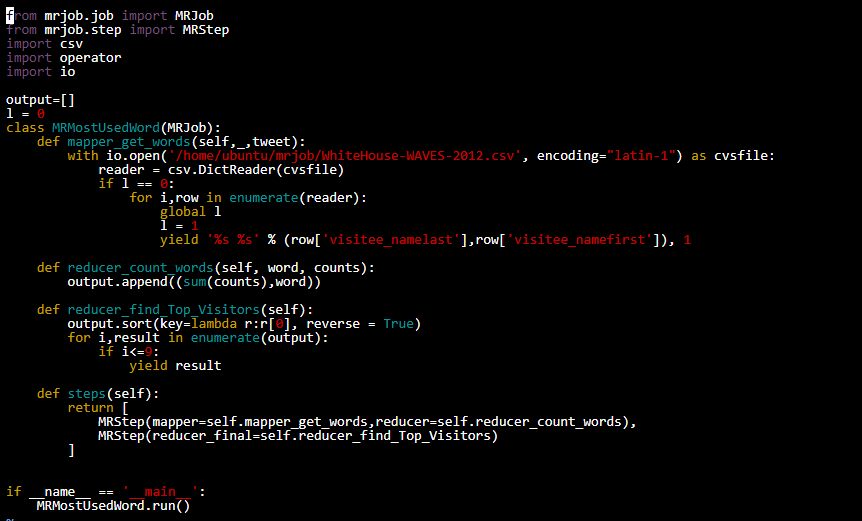
These are the steps how the output of each job will flow from one job to another job. Here mapper and reducer function will run concurrently and last reducer\_final job will print the result.

**2) Your source code**

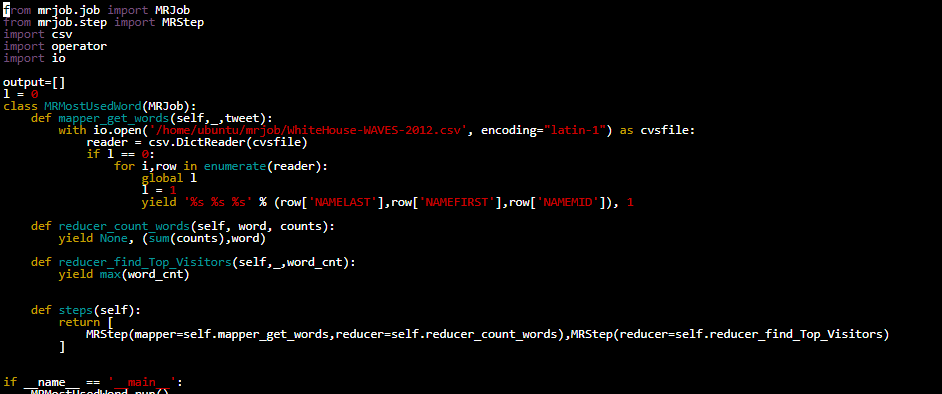
1. Who are the top 10 most frequent visitors (NAMELAST, NAMEFIRST, NAMEMID) to the White House in the year 2012



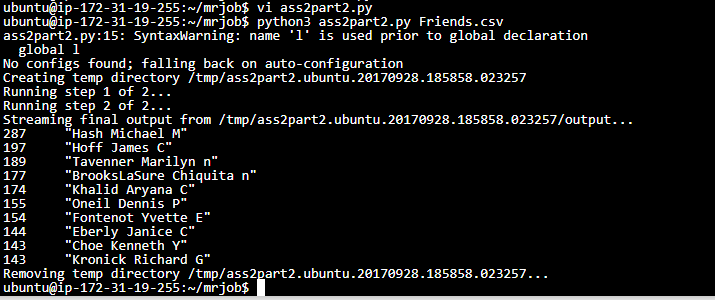
2) Who are the top 10 most frequently visited people (visitee\_namelast, visitee\_namefirst) in the White House in the year of 2012.



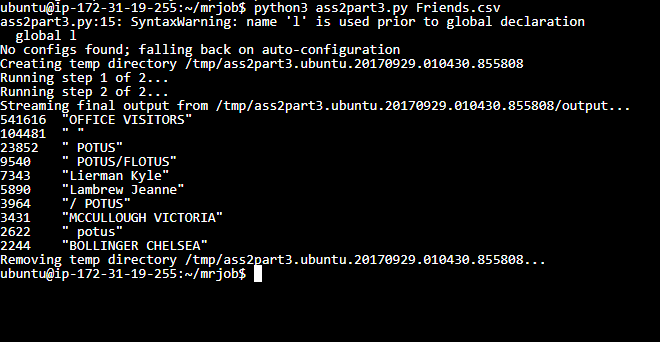
3)Who is the top visited frequent vistor(NAMELAST, NAMEFIRST, NAMEMID) to the White House in the year 2012

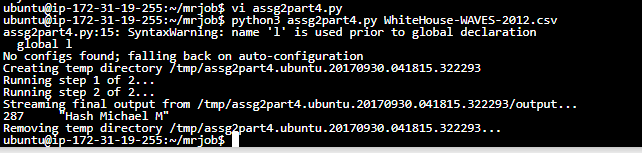


**3) Your analysis results.**

1)

2)



3)

4**) Summary of job running on AWS.**

* An Instance is created under Ubuntu AMI. Now the required files for the mrjob is created in the home directory. The files are said to stored in the same directory.
* Now the job will run with Elastic Map Reduce along with the Instance that we ran before.
* During run time, temporary files and log files are created in the S3 bucket under the tmp directory.
* **Running State, Pending State** are the job states.
* The job status as well as job queues are stored in Amazon NewService.
* A cluster with unique ID is created during run time. The EMR clusters under the analytics tab contains complete detailed information of the **Name Nodes & Data Nodes.**
* The Mapper is used to count all the frequencies and yields to Reducer.
* The reducer then aggregates the results and loads into the logs
* The output results are fetched from the logs