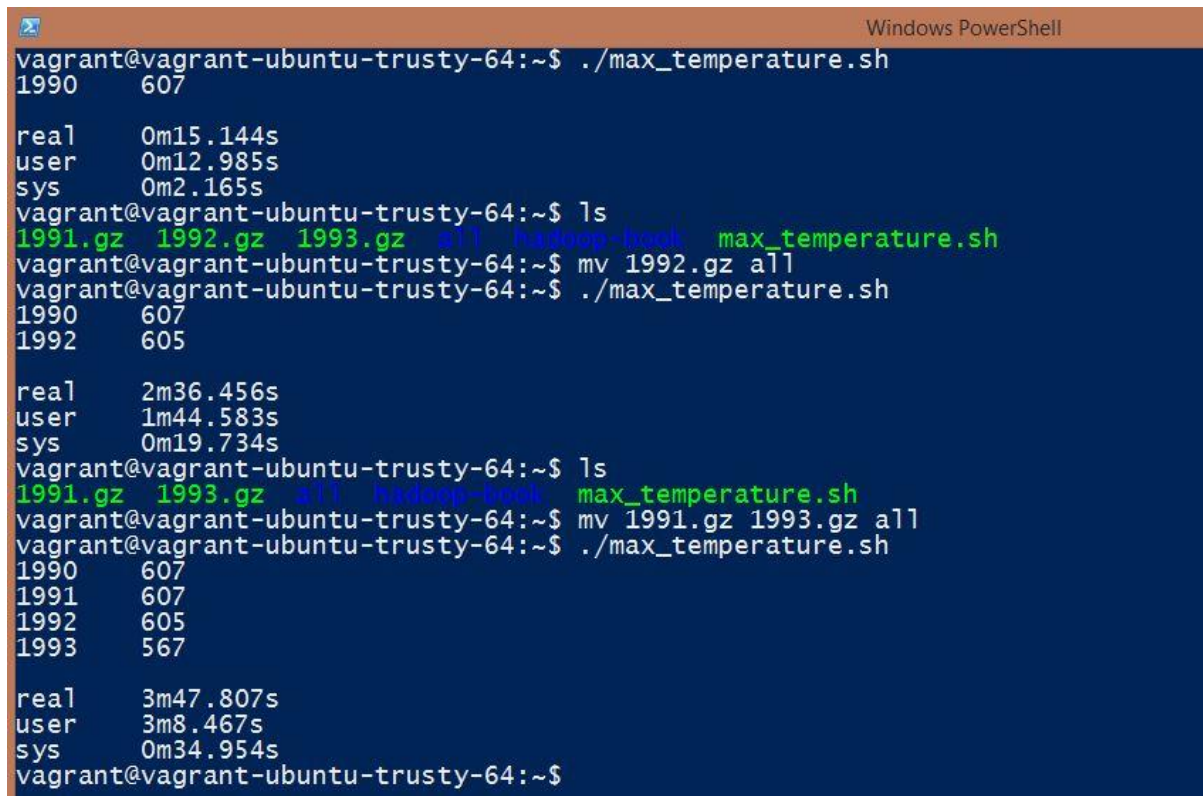


Data Analysis – Comparative Assignment

Part 1:

In this part the gzip files are placed inside the data folder under trusty64 vagrant box and “vagrant reload –provision” command will be run to push all the gzip files into “vagrant_data” which is in the virtual machine. Later all these files will be moved to “all” folder which is under home directory. In the next step, “max_temperature.sh” awk script will be run across all the four gzip files, to calculate the maximum temperature based on the year. Below is the graphical report which demonstrates the time consumed to display the maximum temperature across year 1990, 1991, 1992 and 1993.



```
vagrant@vagrant-ubuntu-trusty-64:~$ ./max_temperature.sh
1990      607

real      0m15.144s
user      0m12.985s
sys       0m2.165s
vagrant@vagrant-ubuntu-trusty-64:~$ ls
1991.gz  1992.gz  1993.gz  all  hadoop-book  max_temperature.sh
vagrant@vagrant-ubuntu-trusty-64:~$ mv 1992.gz all
vagrant@vagrant-ubuntu-trusty-64:~$ ./max_temperature.sh
1990      607
1992      605

real      2m36.456s
user      1m44.583s
sys       0m19.734s
vagrant@vagrant-ubuntu-trusty-64:~$ ls
1991.gz  1993.gz  all  hadoop-book  max_temperature.sh
vagrant@vagrant-ubuntu-trusty-64:~$ mv 1991.gz 1993.gz all
vagrant@vagrant-ubuntu-trusty-64:~$ ./max_temperature.sh
1990      607
1991      607
1992      605
1993      567

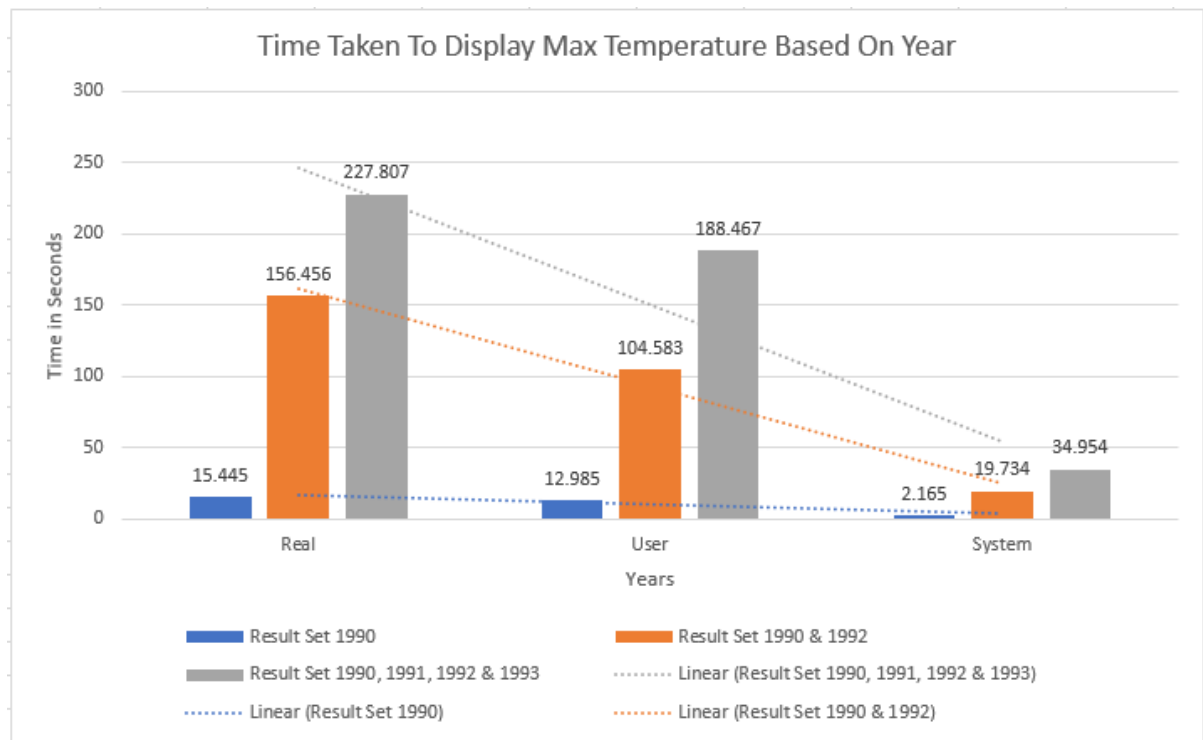
real      3m47.807s
user      3m8.467s
sys       0m34.954s
vagrant@vagrant-ubuntu-trusty-64:~$
```

The data displayed in the above image is plotted graphically to show the timing trends. The calculation is done by considering all the possible combinations and it is spread across three components, real, user and system.

	Result Set		
	1990	1990 & 1992	1990, 1991, 1992 & 1993
Real	15.445	156.456	227.807
User	12.985	104.583	188.467
System	2.165	19.734	34.954

CPU Details:

	Before Execution	During Execution
CPU	1.3 GHz	2.5 GHz
RAM	2 GB	4 GB



From the above graph, it is evident that the time taken by the system, user and real entities are gradually increasing with respect to data size. Hence, we can conclude that time taken to fetch a result is directly proportional to size of the data.