**Shrija Chavan**

**A20381511**

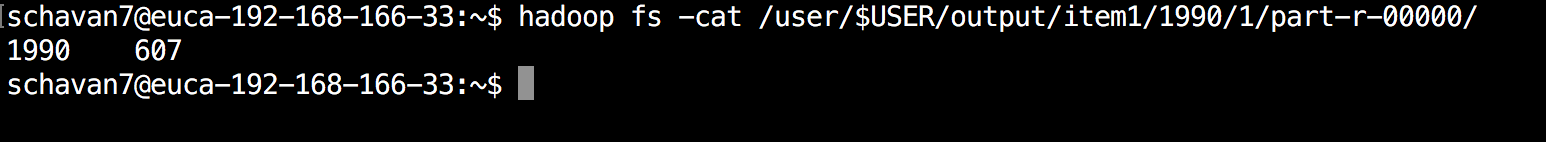
**ITMD 521**

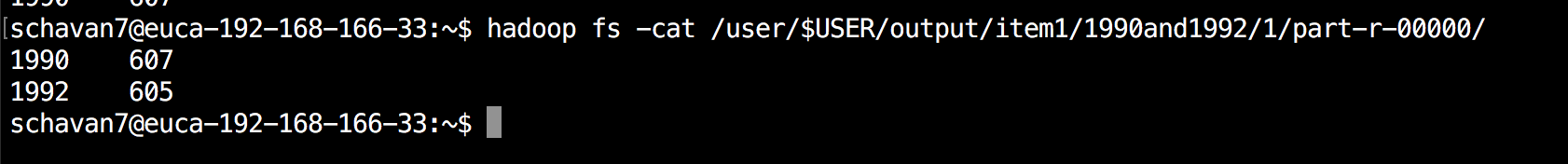
**Week 07**

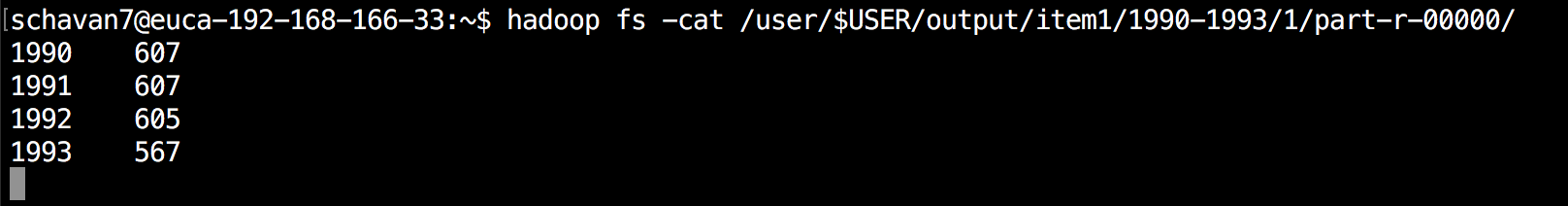
**ITEM1: Analysis**

1.Explain why the run times are different - note the amount of resources currently in use (8088) and explain in detail how MR takes advantage of the distributed cluster (Cite diagrams from the various chapters in the book explaining what is going on)

1. The run times increases as the datasets increases. The least amount of the time taken is for 1990 dataset and the maximum amount is taken by 1990-93 dataset.
2. One more reason might be the number of resources available to carry out the MR task. If there are few nodes the time taken will be more, since the resource manager waits until the node finishes the first task before giving it the next one.
3. We can refer to figure 2.2 and 7.1 which explains how the resource manager assigns the task to the Node Manager and carries out the task.
4. Data locality might also be one of the factor, if the data local (same node or same rack) the time taken would be comparatively lesser than If the data is not local (Different rack).
5. The number of resources in use to run Item one was ……….

****

****

****

**ITEM2: Analysis**

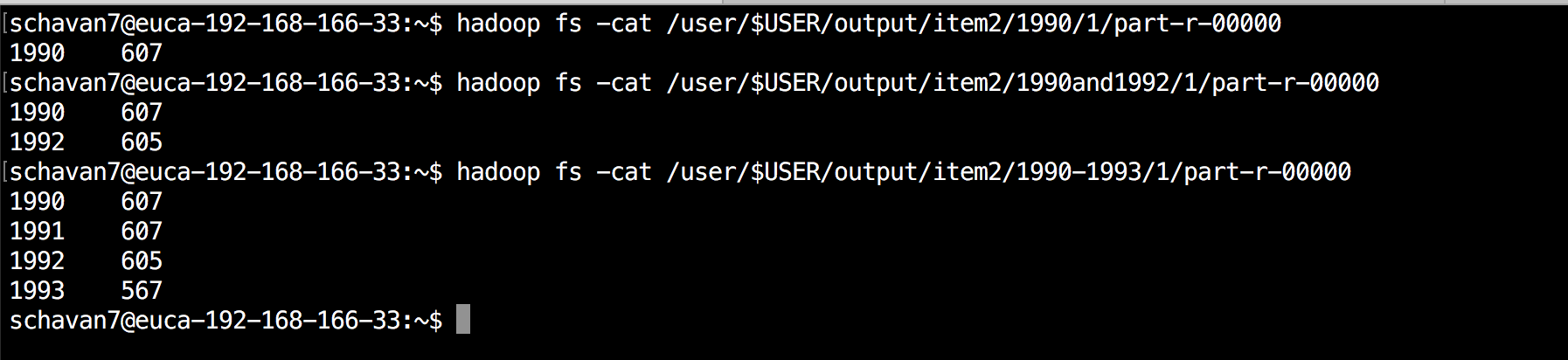
Run the above jobs again this time using the MaxTemperature with combiner example code

Repeat your graphing and analysis and explain why and how using Figure 7-1 the execution time changed - note the amount of resources in use at the time of execution (this will vary depending when you are working and results will not be the same for everyone)

1.In item two the time taken could be comparatively lower than item one, since we are running the combiner class, which acts as the local reducer in the Map phase to reduce the amount of data that is transferred to the reduce phase in the MR job.

2. By following fig 7.1 and fig 2.4(page 34, the combiner function)from the book, we can see that the Map tasks are carried out first and will be reported to the Application master upon completion, the reduce phase cannot start until the map phase is complete as the reduce phase will need all the outputs of the map phase. One all the outputs of the map phase are reported, the Application master will assign the reduce tasks to the node. Here the amount of data transfer is very less as the combiner class will sort the data which is the output of the map task so that the data transfer is less over the network which may cause delay in carrying out the MR job.

3. The number of resources in use to run Item two was ……….

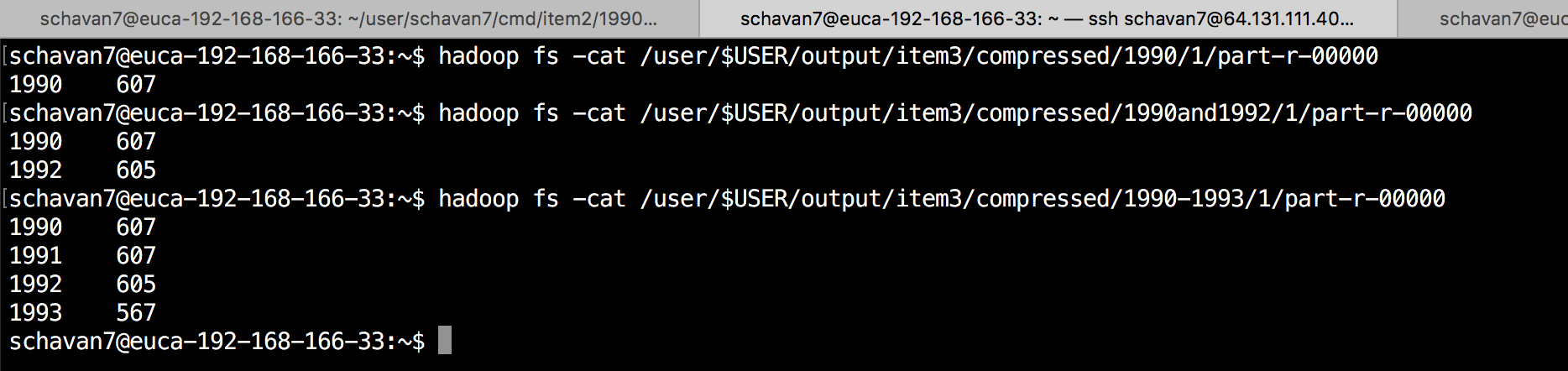
****

**ITEM3 : Analysis**

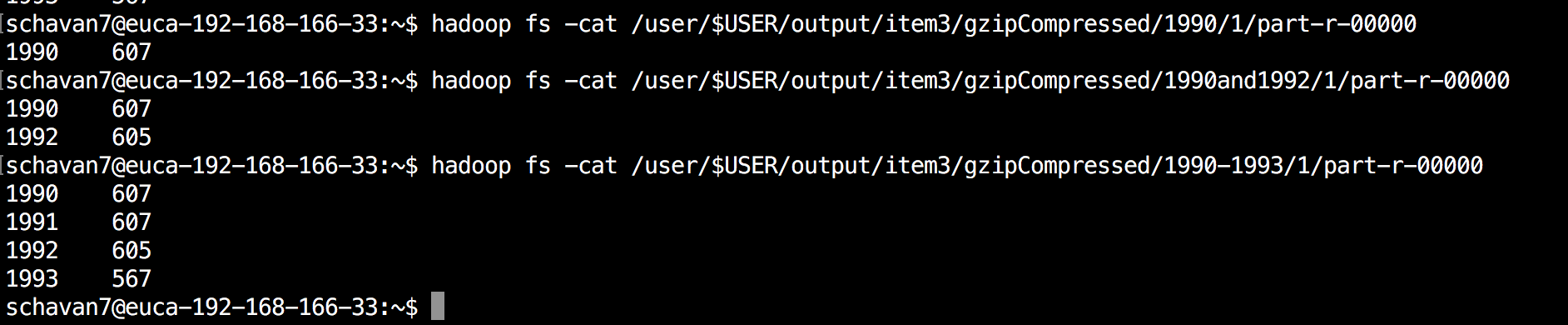
Based on P 118 of the Text Book (ePub) You are to rerun the above MaxTemperature and MaxTemperature with combiners in matrix of variables

Run the jobs with the input uncompressed, gzip compressed, and bzip2 compressed (over each dataset) Note the execution times, the resources in use, and explain the results for the jobs with a combiner and without a combiner and explain from the book why the results occurred.

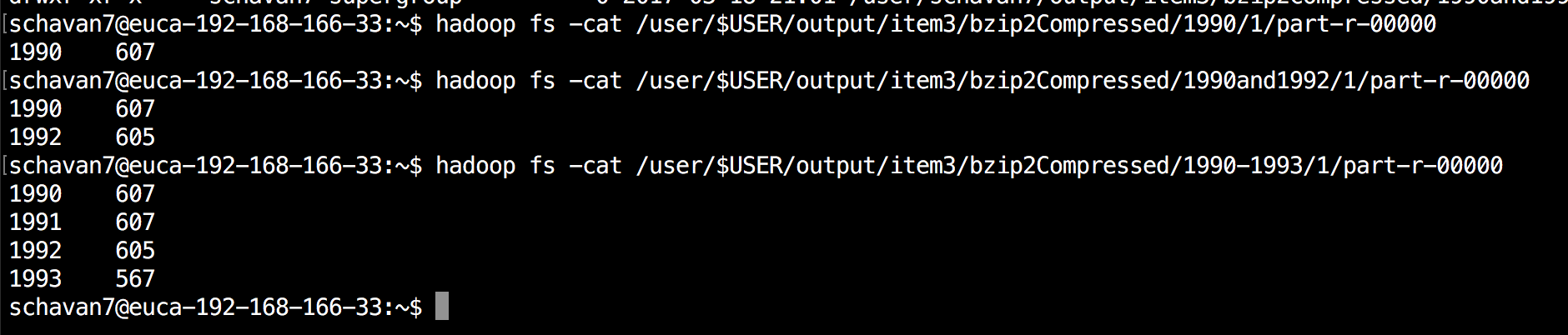
**Compresed:**

****

**GzipCompresed:**

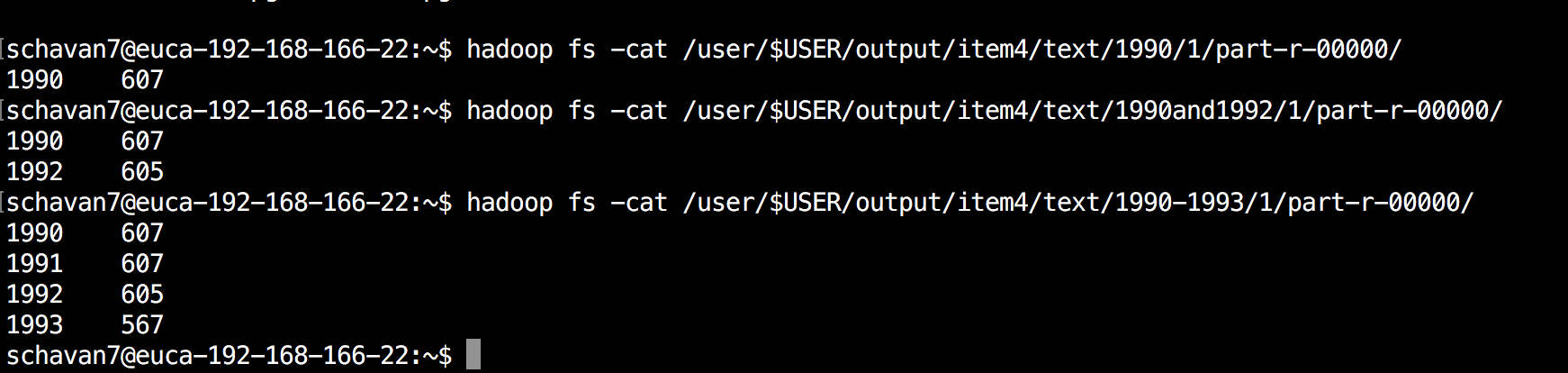
****

**bzipCompresed:**

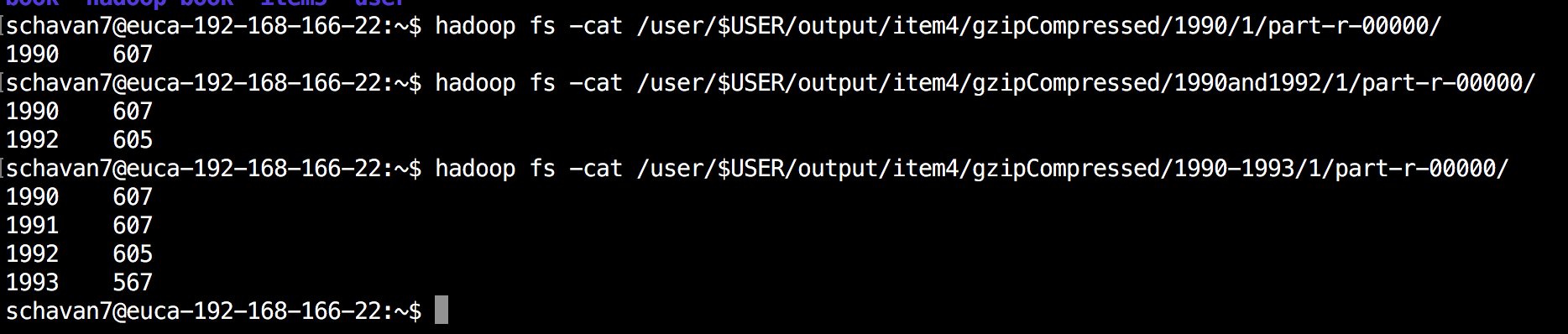
****

**Item 4:**

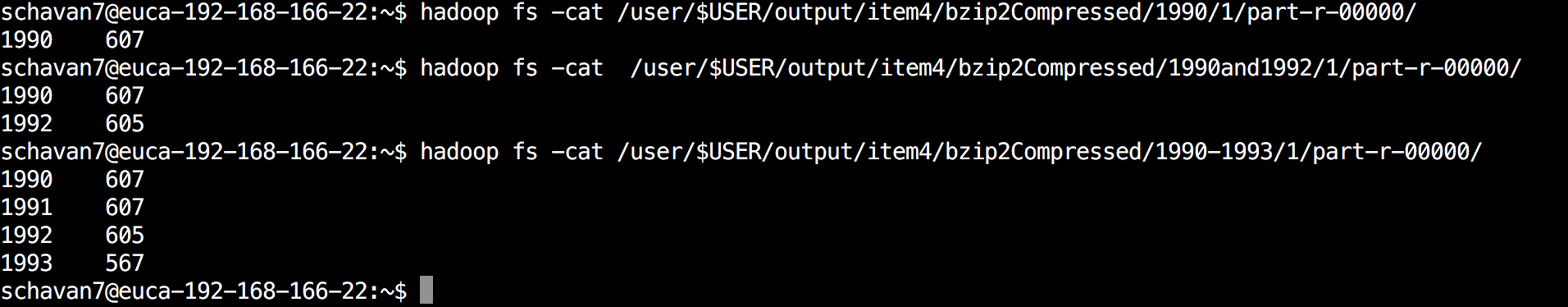
**Text:**

****

**Gzip:**

****

**Bzip:**

****

**Item5:**

**Text**

**Gzip**

**Bzip**