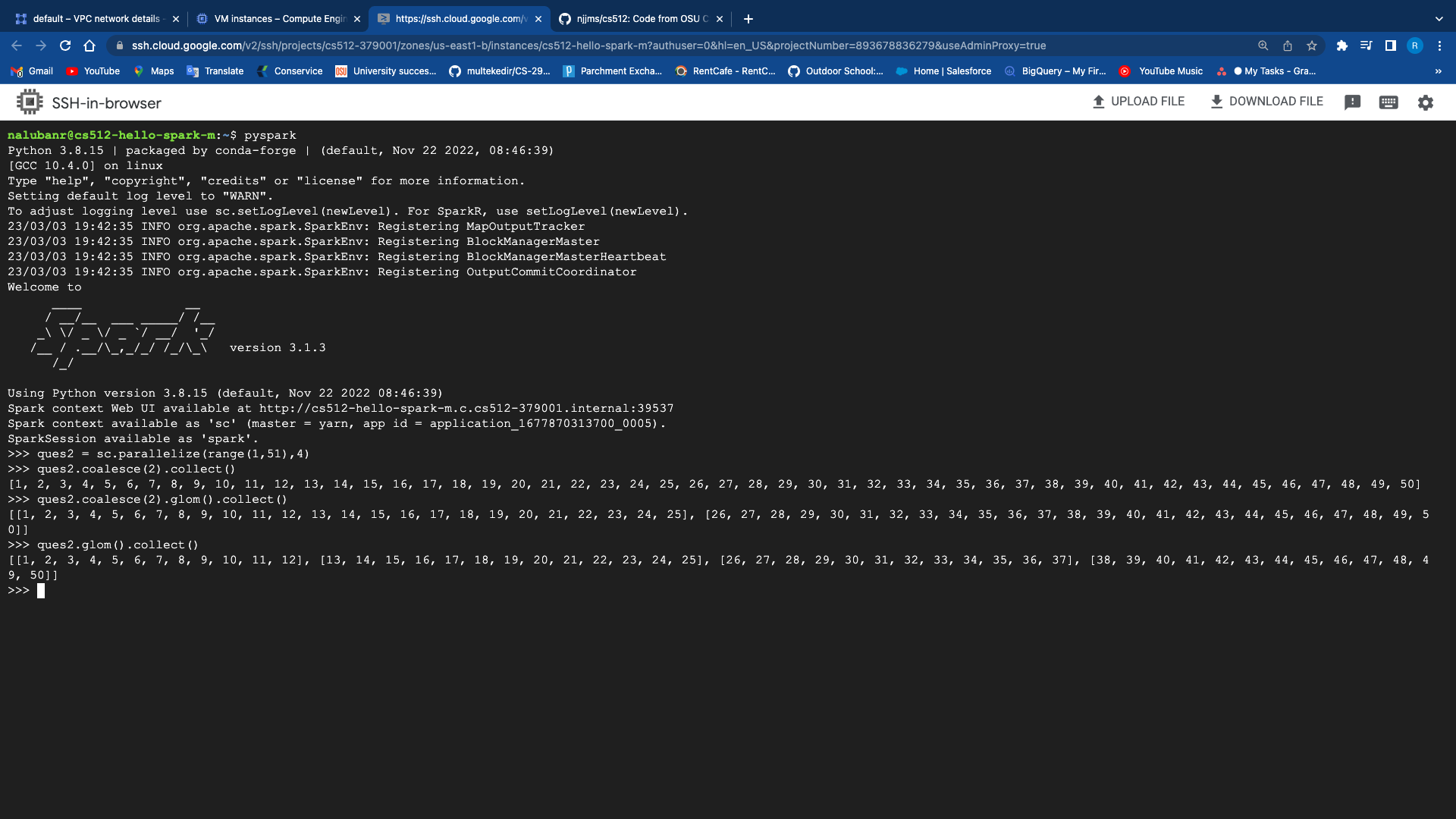
HW: Spark Activities

1) Make a list of 25 integers across 3 partitions.

Text

Description automatically generated

2) Make a list of 50 integers across 4 partitions, efficiently convert it to 2 partitions.  


3) Starting with a list of 26 integers 0 through 25 on 1 partition, end with a list of 26 integers split among two partitions, even numbers on one and odd on the other.

Text

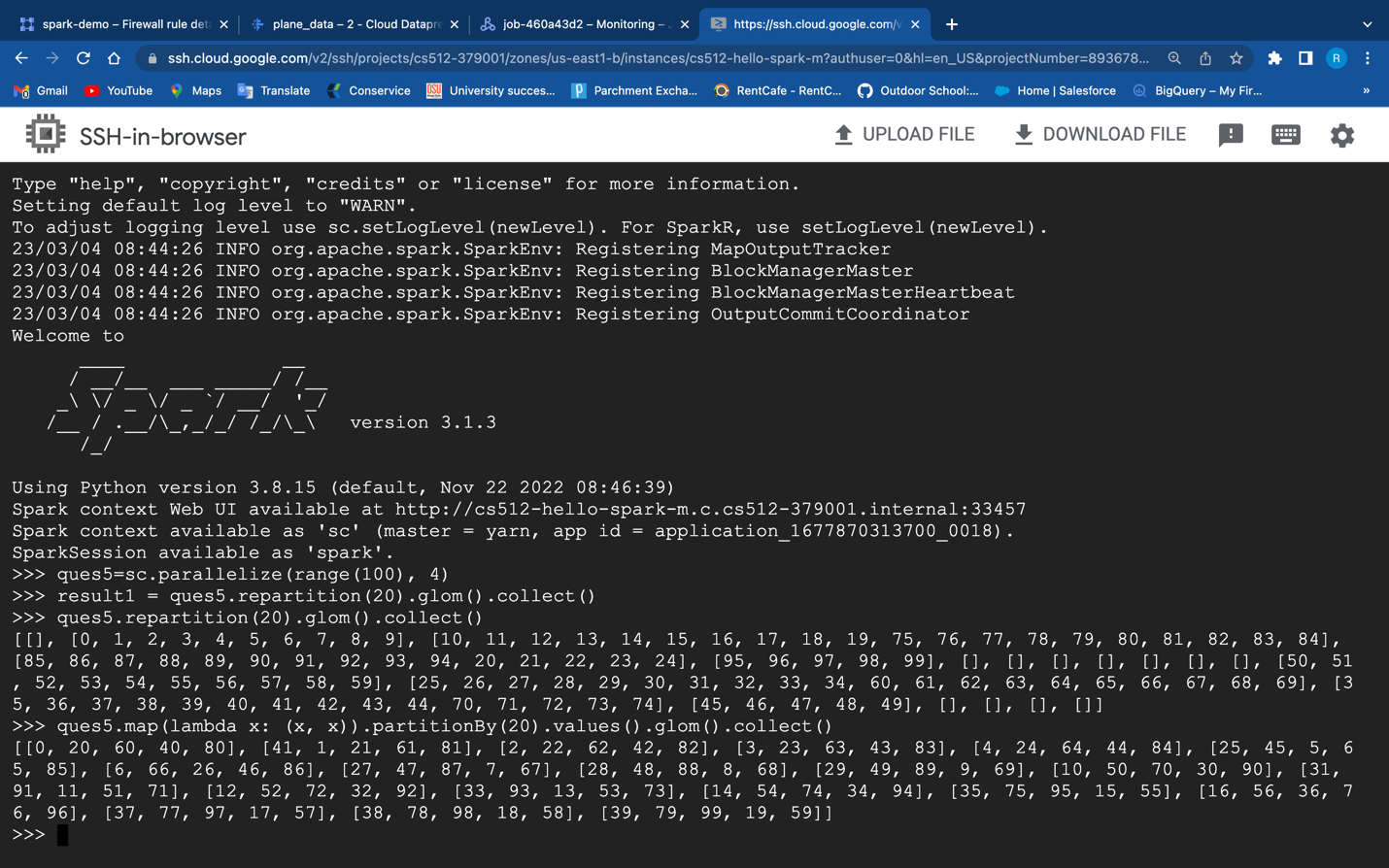
Description automatically generated

4) Starting with 20 strings split somewhat evenly across 3 partitions, end with 4 partitions will ALL the strings stored in one with the other 3 empty.

Text

Description automatically generated

5) Compare the results of using repartition(20) directly on an RDD containing the values 0 through 99 with the results of first making a key value pair using the value as the key, then using partition By(20)



Using repartition(20) directly on an RDD containing the values 0 through 99 it will result in creating 20 partitions with varying numbers of elements in each partition, depending on the distribution of the data whereas creating a key-value pair using the value as the key, and then using partitionBy(20) will create 20 partitions where each partition will contain a contiguous range of keys.