Assignment 1

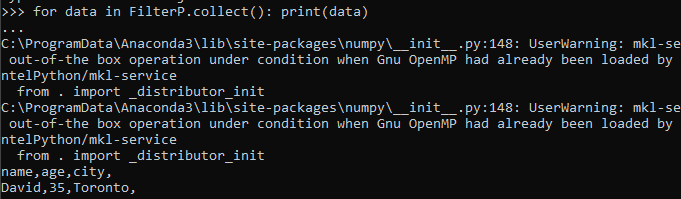
Q.1 RDD Creation: How could you create an RDD from the sample dataset provided above?



Q.2 Transformations: Which action would you apply to filter out individuals who are over 30 years?



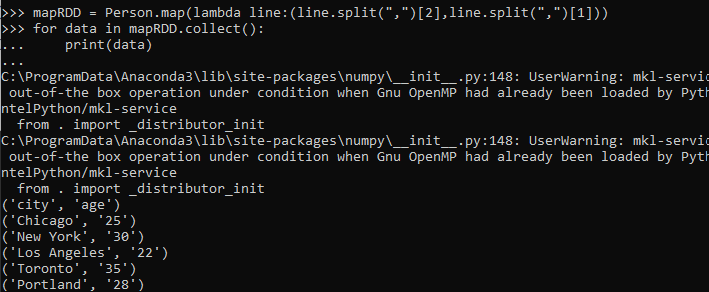
Q.3 Actions: Which action would you use to collect the results of the RDD after applying the transformation in question 2?



Q.4 Persistence: If you wanted to reuse the filtered RDD in multiple actions, what method would you call to optimize the performance?



Q.5 Key-value Pairs: How can you transform the dataset into key-value pairs where the key is the city and the value is the age?



Q.6 Shuffling: What happens during the shuffling phase, and why it is important to minimize it ?

Ans. During the shuffling phase in spark, data is moved across the network between executors to ensure that all data belonging to a particular key in a key-value pair RDD ends up on the same machine for further processing like reducing, joining etc. Its important to minimize shuffling because it involves network input output and can be performance bottleneck.

Q.7 Partitions: How does the no of partitions affect parallelism in spark?

Ans. The number of partitions in a Spark RDD directly impacts the level of parallelism achievable during processing. Each partition represents a unit of work that can be processed independently. Spark aims to distribute these partitions across multiple executors (nodes) in the cluster. The more partitions, the greater the potential for parallel execution.

Impact on Performance:

* Too few partitions: Limits parallelism, as fewer tasks can be executed concurrently. Can lead to inefficient resource utilization.
* Too many partitions: Increases overhead due to managing a large number of small tasks. Can introduce network communication overhead if data needs to be shuffled between partitions.

Generally, aiming for a number of partitions similar to the number of cores in your cluster is a good starting point. However, this might need adjustment based on your specific workload and dataset characteristics. Insummary**,** the number of partitions in Spark directly influences the degree of parallelism and can significantly impact the performance of your applications. Finding the right balance is crucial for optimal results.