1. Describe the setup architecture of your exercise.

I divide the whole program into three parts, including Mapper, Shuffle, Reducer respectively. Mapper is aimed to process matrix’s data and Shuffle will distribute data with same key into same iterate, and Mapper will calculate the data from Shuffle, finally return and print result.

1. Describe Mapper and Reducer function in Python.

Mapper will mark the data (aij) from matrix A with form of <key, value>, key=”i, k”, i is number of row and k is number of matrix B’s column, value = “a:j, aij”. As well as mark the data (bij) from matrix B with form of <key, value>, key=”k, j”, j is number of column and k is number of matrix A’s row, value = “b:i, bij”. Based on this, it can easily tell which matrix elements belong and specific position in matrix. And cij can be calculated with data companied with same key value.

Reducer will do work of calculation, it will multiply data with same number after type of matrix and clearly know the position of data regarding to key value.

1. Describe your experience step-by-step in your own words and provide screenshots of executed MapReduce programs.

文本

描述已自动生成

Modify permission of mapper file and reducer file.

图形用户界面, 应用程序

描述已自动生成

文本

描述已自动生成

Test the result on my own localhost.

文本

描述已自动生成

文本

描述已自动生成

Run hdfs file system and yarn resource manager. Type ips command to see what services are open now.

文本

描述已自动生成

蓝色的标志

描述已自动生成

Have to put input matrixs on hdfs in advance.

文本

描述已自动生成

Upload input, output, mapper, reducer into Hadoop’s HDFS by Streaming, and run program.

文本

描述已自动生成

Program is completed.

文本

描述已自动生成

Print the result from HDFS and its result is same with one completed on localhost.