Assignment 2

Map Reduce in Spark

Question 1: Matrix Multiplication

In order to perform matrix multiplication I have used Two sets of mapper and reducer and one mapToPair to generate the output in given format.

Mapper 1 : FlatMapToPair

This mapper will create tuples in form of

((A,row_number),(column_number,value))

((B,column_number),(row_number,value))

Reducer 1: GroupByKey + MapValues

The goal of this combination is to generate rows and columns sorted by their respective indices.

GroupByKey will group elements in the row of A and columns in B

The MapValues will transform the value part into a list of values sorted with respect to the index. In case of matrix A column_number will act as the index and similarly in B row_number will act as the index.

Communication Cost of this task: m*n+n*p

Since, The mapper 1 is creating a tuple for each element

Mapper 2: FlatMapToPair

Now This will create copies of rows and columns in such a way the rows and columns which are supposed to be multiplied are mapped to same keys.The format of the output from this stage is:

In case of A:

((row_number,column_number),(Elements of Row))

In case of B:

((row_number,Column_number),(Elements of Column))

Communication Cost of this task: m + p

Since, the number of tuples of reducer 1 is number of rows in A+ number of columns in B

Reducer 2 : ReduceByKey+MapValues

This combination will take the two lists multiply the numbers on same index and add them which will result in the element of the index given in the key.The output format of this phase is:

((row_number,col_number),Value)

Communication Cost of this task: 2*m*p

Since, The output of Mapper 2 is a set of row and column in

Mapper 3: MapToValues + sortByKey

This stage will change the output in given output format.

Communication Cost of this task: m*p

Therefore, Total Communication cost: m*n+n*p+2*m*p+m+p

Total Computation Cost: m*n*p