24/25 Graded

Aaray Shah
60004210073
TE Comps C12

Experiment - 6

Aim: To implement python code for matrix multiplication and word count using MapReduce.

Theory: MapReduce is a Java-based distributed execution framework within the Apache Hadoop Ecosystem. It takes away the complexity of distributed programming by exploring the two processing steps that developers implement a) Map and b) Reduce. In the mapping step, data is split between parallel processing tasks. Transformation logic can be applied to each chunk of data. Once completed the reduce phase takes over to handle aggregate data from the map set. In general MapReduce uses Hadoop Distributed File System (HDFS) for both input & output.

Matrix Multiplication Algorithm

Map Function

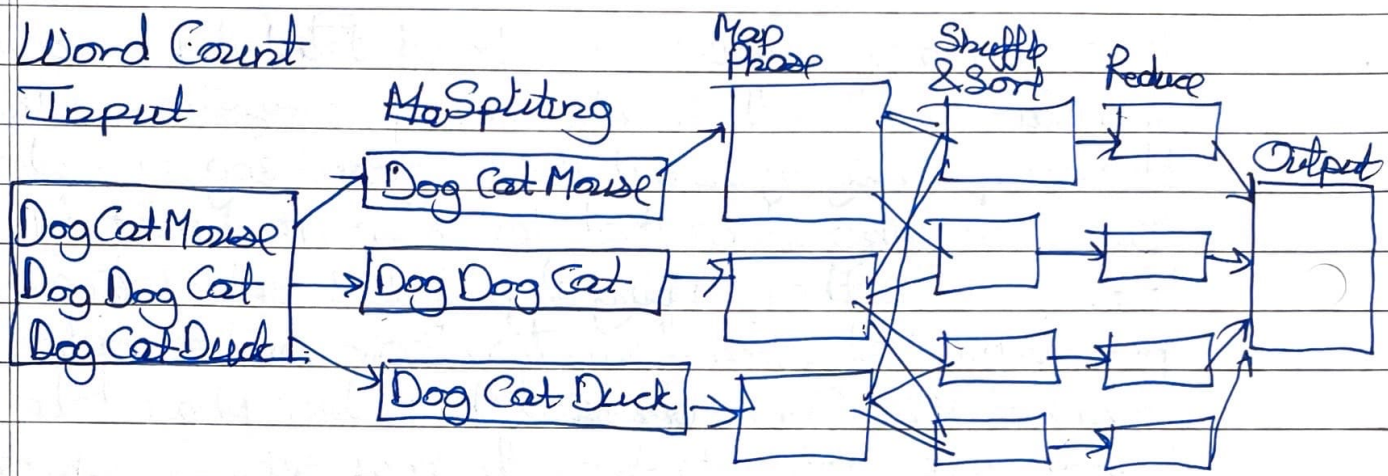
for each element m_{ij} of M do
    produce (key, value) pairs as $((i,k),(M,j,m_{ij}))$, for k = 1,2,3.. up to the number of columns of N
for each element n_{jk} of N do
    produce (key, value) pairs as $((i,k),(N,j,n_{jk}))$ for i = 1,2,3.. up to the number of rows of M
return Set of (key, value) pairs that each key (i,k) has a list with values (M,j,m_{ij}) and (N,j,n_{jk}) for all possible values of j

Reduce Function

for each key (i,k) do
    sort values begin with M by j in listM

sort values begin with M by j in list M
multiply m_{ij} and n_{jk} and j^{th} value of each list
sum up $m_{ij} * n_{jk}$
return $(i, k), \sum_{j=1} m_{ij} * n_{jk}$

Word Count



Conclusion - Thus we have implemented matrix multiplication and word count using MapReduce in python.