Q1. Execute all the mapreduce jars as shown in the sessions

Q2. Execute the spark program to find the frequency of each word, same like shown in the sessions

Q3. Create logic for linkedIn profile views in Apache Spark.

-Create a new directory structure data/input in your home directory of the gateway node.

-Now create a new file with the name linkedin_views.csv under the above created directory using vi editor.

-Insert the below records

1,Manasa,Sumit
2,Deepa,Sumit
3,Sumit,Manasa
4,Manasa,Deepa
5,Deepa,Manasa
6,Shilpy,Manasa

save the file and quit

so basically below is the structure of your file
id,from_member,to_member

If we have to interpret the row
1,Manasa,Sumit

its like id is 1, and Manasa visited Sumit's linkedIn profile

Now copy this file linkedin_views.csv from your gateway node to HDFS under the directory /data/input in your hdfs home.

We need to find how many times each person's profile is viewed.

The following tasks has to be done in Pyspark3 notebook..

step 1 - Create a spark session using the boiler plate code

step 2 - Load the file from hdfs and create a base rdd

step 3 - perform the required transformation to find how many times each person's profile is viewed.

step 4 - Finally save the results to hdfs in a subdirectory data/output under your hdfs home directory.

Make sure you shut down the kernels after using the lab!