HOMEWORK ASSIGNMENT: MODULE 7

Complete the following tasks:

Use the spark-shell to run the following spark shell commands

Dataset:

The column names are given below
id,actorname,genre,noofmovies,hits

1001,Johnny Lever,Commedian,210,190
1002,Amit Mahesh,New Comer,2,1
1003,Salman Khan,SuperStar,300,291
1004,Johnny Depp,SuperStar,289,270
1005,Mallika Sherawat,Actress,20,10
1006,Amitabh Bachan,SuperStar,350,300
1007,Micheal Bijlu,New Comer,4,1
1008,Rocky Angelo,New Comer,3,1
1009,Rajani Kanth,SuperStar,400,399
1010,Arnold,SuperStar,261,242

1. Please construct the dataset as given below using Seq and case class ActorDetail.

2. Construct a data frame using the above dataset by creating 2 partitions (use parallelize method).

3. Show the all the details of the actDetailsDF and also show only the first two records of the actorDetails dataset.

4. How do you print the schema of the actor details dataframe?

5. Register the DataFrame as temp table and only select the rows for whom the no of hits are more than 250 and print the output.