**Problem Description**

You have two files named movies\_en.json and artists\_en.json containing a small movie database in the JSON format. You need to load them into Spark Data frames and perform analysis.

Sample record: movies\_en.json - one record per line, newline characters have been added for readability:

{

"id": "movie:14",

"title": "Se7en",

"year": 1995,

"genre": "Crime",

"summary": " Two detectives, a rookie and a veteran, hunt a serial killer who uses the seven

deadly sins as his modus operandi.",

"country": "USA",

"director": {

"id": "artist:31",

"last\_name": "Fincher",

"first\_name": "David",

"year\_of\_birth": "1962"

},

"actors": [

{"id": "artist:18",

"role": "Doe"

},

{"id": "artist:22",

"role": "Somerset"

},

{"id": "artist:32",

"role": "Mills"

}]}

And here is an example record from artists\_en.json:

{

"id": "artist:18",

"last\_name": "Spacey",

"first\_name": "Kevin",

"year\_of\_birth": "1959"

}

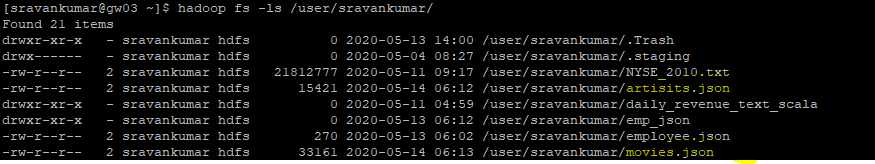
movies\_en.json contains the full names and years of birth of movie directors, but only theidentifiers of actors. The full names and years of birth of all artists, as well as their identifier, are listed in artists\_en.json.

**Assignment 1:**

Connect to the Hadoop cluster and copy the two files movies\_en.json and artists\_en.json toa folder in HDFS.



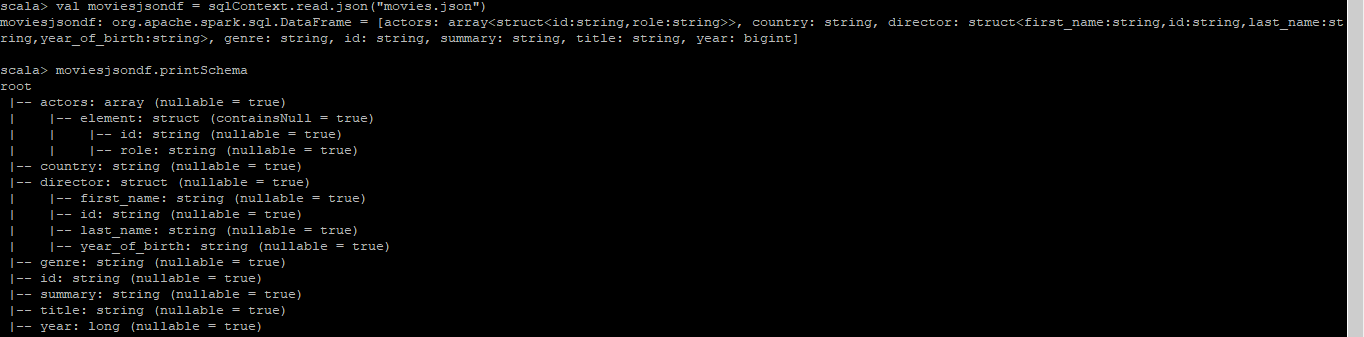




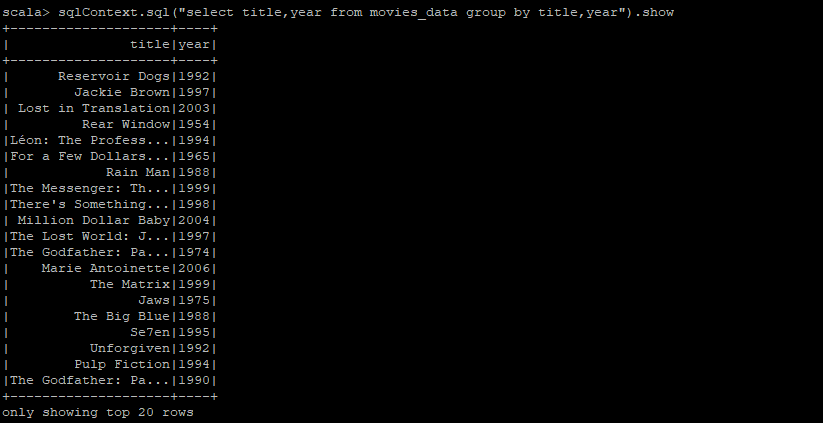
**Assignment 2:**

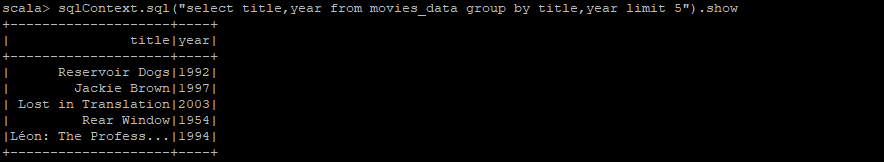
Write spark SQL code to perform the following:

1. Read the files into data frames
2. Write a query to group titles of American movies by year
3. Display the first 5 records in command line



****

****

****

Sample Output:

[...]

(1988,{(Rain Man),(Die Hard)})

(1990,{(The Godfather: Part III),(Die Hard 2),(The Silence of the Lambs),(King of New York)})

(1992,{(Unforgiven),(Bad Lieutenant),(Reservoir Dogs)})

(1994,{(Pulp Fiction)})

[...]

**Assignment 3:**

Write spark code to normalize the data frames created above and store the output as Parquet files

Output will be stored in 3 folders

Folder 1: Stores artist details

Folder 2: Stores Movie Details

Folder 3: Stores link between Movie, Artist and Role played

**Assignment 4:**

Write spark code to normalize the data frames created above and store the output as 3 Hive Table in Parquet files

**Assignment 5:**

Execute a Spark SQL query on hive tables to list MovieID, Title, Year, Genre, Country, Director and actors by joining on the tables created in assignment 4 and display the first 5 records in command prompt.