**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 Hive Tables 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 the identifiers 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 to a folder in HDFS.

**Assignment 2:**

Create 2 hive tables (movies\_json & artists\_json) as per the JSON structure and load the JSON files into the table. Hint: Use the JSON SerDe available from <https://github.com/rcongiu/Hive-JSON-Serde>

**Assignment 3:**

Write a query to group titles of American movies by year

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 4:**

Write a query to list MovieID, actorID, role

(movie:1,artist:15,John Ferguson)

(movie:1,artist:16,Madeleine Elster)

(movie:2,artist:5,Ripley)

(movie:3,artist:109,Rose DeWitt Bukater)

[...]

**Assignment 5:**

Normalize the schema by adding new tables with data stored in Parquet format and load data

3 new tables need to be created:

Table 1: Stores artist details in Parquet format

Table 2: Stores Movie Details in Parquet format

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

**Assignment 6:**

Create following indexes on Table 1:

Bitmap index on Genre column

Compact Index on MovieID column

**Assignment 7:**

Create a view to list MovieID, Title, Year, Genre, Country, Director and actors.