Technical Interview Question on Big Data Task – 1

Exploratory Data Analysis:

Dataset : *Provided database with name 'movielens-small'*

- Write a SQL query to create a dataframe with including *userid, movieid, genre* and *rating*
- Count ratings for each movie, and list top 5 movies with the highest value
- Find and list top 5 most rated genres
- Find and list top 5 most rated tags
- By using timestamp from ratings table, provide top 5 most frequent users within a week
- Calculate average ratings for each genre, and plot average ratings of top 10 genres with descending order

Task – 2 Recommender Design:

Dataset : Provided database with name 'movielens-small'

- Provide an implicit feature by using any of the data from the given database
- Train two individual recommender models, one by using rating (from ratings table) and the other one by using your designed implicit feedback
- Present comparison between two models, by using essential metrics

Task – 3 Text Analysis:

Dataset : *http://ai.stanford.edu/~amaas/data/sentiment/*

- Create a dataframe with following schema:

```
root
 |-- content: string (nullable = true)
 |-- label: string (nullable = true)
 |-- sentiment: string (nullable = true)
```

- Design a tokenizer for content column and remove stop words, and give descriptive information about obtained content column

Notes:

- You should implement your solutions by using Python and PySpark API (via a Notebook or IDE)
- Please do not use pandas library for dataframe applications