# Spark Evaluation Lab 6

Q1. Below is the data for the ListOfStartups from the company information bureau of India.

Sample data: (Please use the attached file: ListOfStartups.csv)

| Incubation Center | Name of the startup | Location of company | Sector |
|---|---|---|---|
| SIIC IIT KANPUR | E-Trainer Analytics Wizard Pvt ltd | New Delhi | Fit-Tech |
| SIIC IITK | Invariance Automation Private Limited | Kanpur, Uttar Pradesh | Industrial Automation |
| SIIC, IIT Kanpur | Neoperk Technologies Pvt. Ltd. | Mumbai, Maharashtra | Agri-Tech - Soil Testing |
| SIIC IIT KANPUR | WeRehab Technologies Pvt Ltd | Nagpur, Maharashtra | Health tech |
| IIT Kanpur | Arthavedika Tech Pvt Ltd | Noida, Uttar Pradesh | Fintech |

DTA is as below:

Incubation_center - Place where the startup was incubated
Name_of_Startup - Startup name
Location - Location of company
Sector - Sector of startup

Software Stack needed: Spark, HDFS (using data from HDFS is optional)

- Analyze it using Spark and answer the following questions:

1. Find which `sector` has the most startups
2. Split the `Location of company` into 2 columns, `state` and `city` . If state is not present then keep it as null
3. If `Location of company` column has a data `DIAT,Pune` then set `state` as `Maharashtra` and `city` as `DIAT Pune` .
4. If `Location of company` column has a data `Ulhasnagar` then set `state` as `Maharashtra` and `city` as `Ulhasnagar`
5. Find which State has the max number of startups
6. Find all the startups from `Maharashtra` .
7. How many startups were formed in `Healthcare` sector
8. Display all startups from `Pune` and `Nashik`
9. Sort the cities in `Maharashtra` in descending order of the count of startups
10. How many startups are in South India. That is states `Karnataka` , `Tamilnadu` , `Telangana` , `Andhra Pradesh`
11. How many startups are in `Gujarat`
12. How many startups are in North India.That is states other than `Karnataka` , `Tamilnadu` , `Telangana` , `Andhra Pradesh` and `Maharashtra`
13. What is the percentage of startup initiative from South India and Maharashtra

14. What is the percentage contribution of startup from Maharashtra
15. What is the percentage contribution of startup from Gujarat
16. Replace `state` with null values to `Unknown`
17. Store the DataFrame with following Schema into a Hive table `StartUps_Spark` with partitioning done on columns `state` and `city`

```
root
|-- Incubation_Center: string (nullable = false)
|-- Name_of_startup: string (nullable = false)
|-- Location_of_company: string (nullable = false)
|-- Sector: string (nullable = false)
|-- city: string (nullable = false)
|-- state: string (nullable = false)
```

**Hint -** Use partitionBy and saveAsTable API from pyspark

In [ ]:
```
1
```