greatlearning

# Project

## Business Requirement:

You are working as a data analyst with the company Yelp. Yelp is one of the first companies in the world who introduced the idea of local business listing of restaurants. Today we have many similar businesses like Zomato , food panda etc. As a business analyst with Yelp, you have collected the data of restaurants and other businesses that has tied up with you. The data is in JSON format and you would like to analyse this data to get more insights about the business.

## Data Set Description

The dataset consists of many columns and is in nested JSON format. You may use the built in json reader of Spark Dataframe API to read and analyse the data. No need to provide any schema, since the reader can infer the schema automatically.

- Business_id: Unique ID of restaurant
- Full-address: complete address
- Hours: nested column containing opening and closing hours of each day
- Categories: The business category
- City
- Review_count: How many reviews the business has got
- Name
- Latitude and Longitude
- State
- Starts: The rating on a scale of 1 to 5 star
- Attributes: Extra information that the business wants to list such as whether they accept credit cards or allow pets etc.

## Outcome of the Project

After successfully completing the project, the participants will be able to

- Use Spark SQL as a SQL tool for analysing Big Data
- Get understanding about writing queries using Spark
- Approach a business problem and model the solution

## Approach to Solution

Participants can use spark shell to explore the problem and find the solution.

1. Create a dataframe and load the JSON file. After that register the dataframe as a temporary table. (10 Marks)

2. How many businesses have registered with Yelp?(10 Marks)

3. Show the state wise count of businesses registered.(10 Marks)

4. Display the name, starts, review count, city and state of businesse which have got 5 stars(10 Marks)

5. Display name, starts and review count of all businesses from Las Vegas city with descending order of starts and review count(10 Marks)