DSCI-6007: Project Report
Fall 2020

Student name – Ashwini Joshi

Geo-Location Clustering using K-means algorithm

Introduction

This project is mainly based on the concept of clustering in the machine learning domain. Clustering is an unsupervised machine learning technique. Unsupervised machine learning means in which user of the model do not need to supervise the model. But it allows the model to work on its own to discover patterns and information that was previously not detected. Also, unsupervised model extracts data features without knowing about output of the model. In the process of clustering data points are divided into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In the Geo-location clustering, we are making groups of similar locations and assigning them to into clusters. It mainly used when we need to develop some application which uses geographical location to locate place or thigs associated with geo location. For example, if hospital wants to expand its emergency room facility at some location, then the hospital will build an emergency room as per requirement of this facility in some area. This requirement can be analysed using geo location of people who are searching for the same. Here we can use Geo-location clustering.

Implementation of Geo-Location clustering is based on ML algorithm called K-means clustering. In the k-means clustering, it defines clusters with similar properties and then stores k centroids that it uses to define clusters. If point is closer to one cluster' centroid as compare to other cluster then it belongs to that cluster. This distance is calculated using Euclidian distance method.

System Configuration

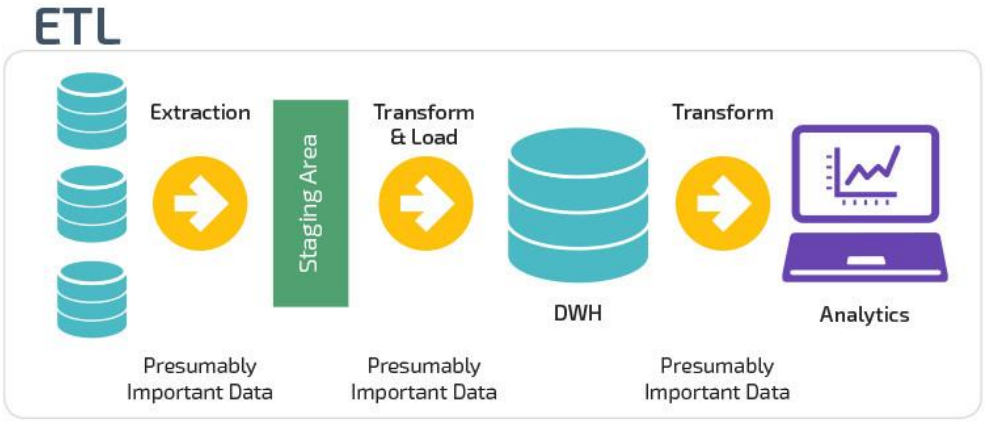
Before starting further processing we need to do some system configurations.

1. First we need to create an .pem file(keypair) required for EMR cluster.
2. Then We need to Create cluster on AWS :
   Login to AWS→Create Cluster→Enter Cluster Name , Type as Spark and Size as m4.xlarge→Add Key pair created in the first step→Wait for cluster to go in the ready state.
3. Add port 8888 to security group for master node.
4. SSH to EMR cluster using terminal.
5. Execute command for installing pyyaml ipython jupyter ipyparallel pandas boto3

6. Open file. .bashrc and add configurations in the file.
7. Execute command – source .bashrc
8. Execute command pyspark and generate token.
9. Open Jupyter Notebook on EMR instance using token generated.
10. Create Python files for required code.

Data Pre-processing

Data contains unnecessary information or attributes that is not required while processing the data. So, to improve efficiency of algorithm we need to remove or transformed irrelevant data. When data is in standardized format then we can use it for later processing. This process is commonly known as ETL(Extract-Transform-Load) process.



In this project we have provided 3 datasets:

- devicestatus.txt
  This input data contains information collected from mobile devices on Loud acre's network, including device ID, current status, location data.
- sample_geo.txt
- lat_longs.txt

As a part of pre-processing, we first did pre-processing for all data files data which is as follows:

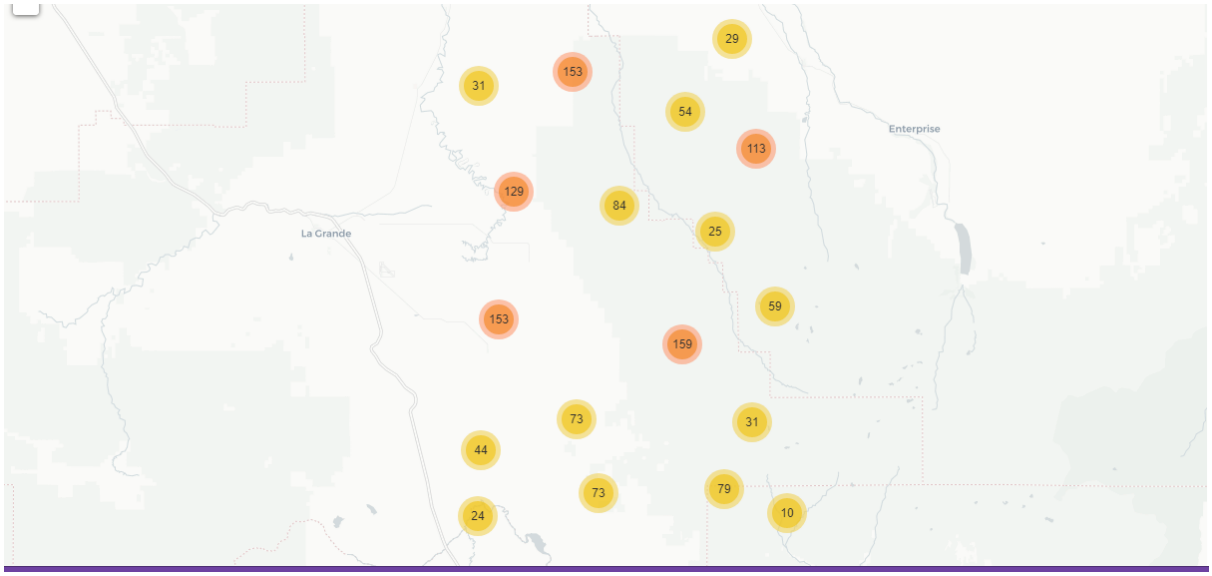
1. Upload required data files to S3 bucket for data storing purpose on Amazon cloud (AWS)
2. Load data from S3 bucket and check if data displayed properly.
3. As mentioned in problem statement, this data contains different field delimiters.

   We need to load data by providing delimiter value. As data have multiple delimiters, we need to remove such rows having other than mentioned delimiter value.

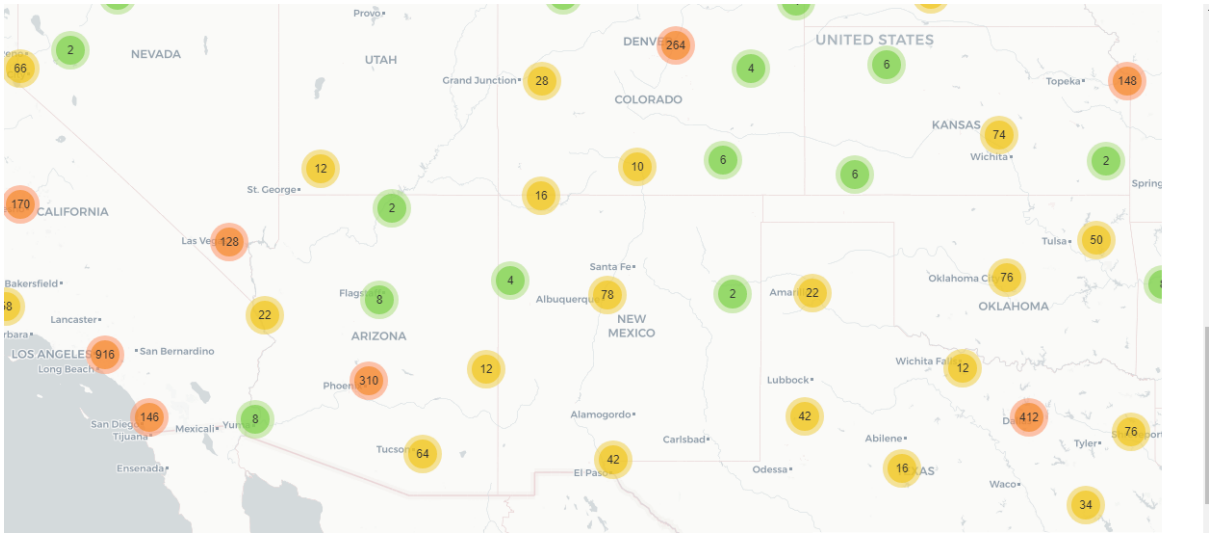
4. Column names are mentioned in the data, so we need to extract those columns date, model, device ID, latitude, and longitude.
5. Filter locations for which latitude and longitude is 0.
6. Model contains device manufacturer and model name, we need to split this field by spaces.
7. Save this data again to S3 bucket by making another folder.
8. Check if output files saved properly to S3 bucket.
9. Display and visualize data using MarkerCluste plugin of Folium.
10. This is a third part library for data visualization.

Visualization:

Device Status visualization



Synthetic Location visualization



DBPedia Location

Approach:

K-means algorithm:

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. We need to define number of centroids required in the dataset that is the target number k. A centroid is the imaginary or real location representing the centre of the cluster. K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The distance between points and centroids is calculated using Euclidean distance. K-means algorithm starts with a first group of randomly selected centroids, and then performs the same calculations repetitively to optimize the positions of the centroids. K-means stops optimizing clusters if

there is no change in their values because clustering is successfully done. Or K-means iterates for the given number of iterations.