**Geo-Location Clustering using k-means algorithm**

**Contents**

1. Introduction and Motivation
2. Data Preparation
3. Visualization
4. Approach
5. Implementation
6. Results
7. Conclusion

**Introduction and Motivation**

**What is Geo-Location Clustering?**

Geolocation Clustering is done over geographically dispersed sites with computer clustering. A Cluster can be defined as a group of independent computers called nodes. Clustering has plenty of useful applications like in marketing, logistics. The clustering is done at the point of data storage or a group of points which are close to one other. Here, we are using k-means algorithm for geolocation clustering to solve the clustering problem in parallel fashion. We implemented the algorithm in Spark.

**Data Preparation**

To implement the algorithm, we need to get the pre-processed data:

* To load the dataset,
* Determine which delimiter to use
* Filter out any records more than 14 values which do not sparse correctly.
* Extract the data, extract the model, the device ID, and lat and long.
* Filtering out the locations that have a lat and long of 0.
* Split the model field by spaces to separate the manufacturer from the model.
* Save the extracted data to comma delimited text files.
* Save the data in a file correctly and confirm.

**Visualization**

**Synthetic cluster location data**

A picture containing diagram

Description automatically generated

**DBpedia Location Data**

