**Approach Note for RE CLUSTEREING**

**Methodology of RE Clustering:-**

* RE Clustering will be initiated on the bases of RE Type Id i.e. on the basis of category like Banking, Mutual Fund etc.
  + In 1.0, RECATGEORY is represented as a String of 5 Characters. A few example are,
    - BADCB – District Cooperative Banks
    - BAFOR – Foreign Banks
    - INBRO – Share Brokers
    - INTRU – Trustees to Trust Deeds
  + In 2.0, RE Type is a String of 2 Characters. A few example are,
    - BA – Banks
    - BF – Brokerage Firms
    - FI – FI/NBFC/Others
* Each segregated RE will be clustered on the basis of their filed KYC Count, Amount and Report differentiated on the basis of STR,PTR,CTR,NTR,CBWTR & CCR
* Each RE will be assigned RE\_CLUSTER\_ID on the bases of RE Type and Cluster ID

**Assumptions:**

* 'RE Cluster Creation' should run only once in a Year wherein entire Clustering exercise and RE\_CLUSTER\_ID assignment to RE's would be done. This in line with the expectation that frequent Updates to RE\_CLUSTER\_ID of existing RE's would not be done since it will have a cascading effect.
* 'RE\_CLUSTER\_ID Assignment' would be run every month and only new REs (REs which have not been assigned any RE\_CLUSTER\_ID) would be considered and a RE\_CLUSTER\_ID will be assigned.
* No of Cluster Groups for each RE\_TYPE has been FIXED to 10. The Distribution of Thresholds within the 10 Clusters would be based on ML Technique (of K Means).

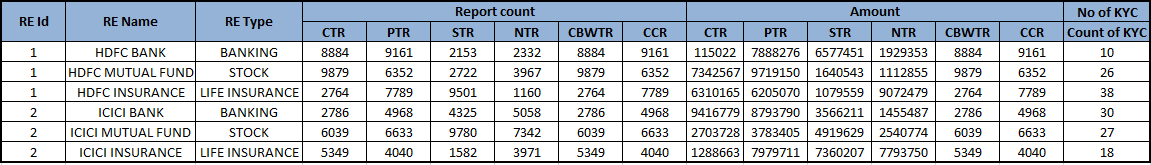
**Steps Implemented To initiate RE Clustering:-**

| STEPS | Description |
| --- | --- |
|  |  |
| Step 1 | Extracting the data of the sum of the report count, amount and KYC filled by RE during some specific time frame on the basis of the Report Type. |
| Step 2 | Clustering will be initiated on the final data where optimal number of clusters will be chosen on the basis of AI/ML and below Column Values   * RE TYPE * KYC COUNT * STR\_REPORT\_COUNT * STR\_AMOUNT * CTR\_REPORT\_COUNT * CTR\_AMOUNT * NTR\_REPORT\_COUNT * NTR\_AMOUNT * PTR\_REPORT\_COUNT * PTR\_AMOUNT * CBWTR\_REPORT\_COUNT * CBWTR\_AMOUNT * CCR\_REPORT\_COUNT * CCR\_AMOUNT |

Table Structure for the RE Clustering is embedded along with:



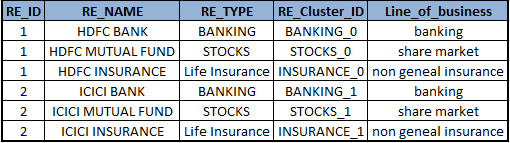
**Step1: Extracted Data**



**Step2: Clustering**

* K means (unsupervised machine learning algorithms) clustering will be used on the output of the data. K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids and then perform clustering.
* For each Retype id, there will be multiple clusters, the optimal numbers of cluster of each category will be chosen on the bases silhouette score. (Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1)

Output of sample clustering will be as follows:



**One RE belonging to multiple peer groups:-**

**Pro:-**

1. Same RE under different category will have different behavior and attributes.

For example, HDFC is one Reporting Entity but different Line of Business like Banks, Mutual Funds, Insurance will have different RE\_CLUSTER\_ID so that it justifies and the characteristics of RE and its Line of Business.

2. RE ID along with Line of Business will be the bases of Clustering.

3- One RE should not be clubbed as Single Entity as average outcome can be different as there may be extreme behavior on their different category