RAW Big Data-Set

New Dataset with Featured Filtered Records

Finally feature extraction method PCA is applied to further refine the selected feature space

Number of group is determined using intrinsic dimensionality estimation

MLE Estimation on dataset to determine the ultimate dimension

Clustered of Dataset is created according to the redundant Using Spark

To filter irrelevant and noise features multi-strategy combination is used (FiS and IG)