CLAIMS

What is claimed is:

1. A computer-implemented method for data management of documents in a plurality of data repositories in a computer network or cloud infrastructure, the method comprising:

a computer receiving user-defined criteria and/or user-defined conditions;

the computer sampling only metadata extracted from the documents in the plurality of data repositories, wherein the metadata extracted from the documents are sampled in accordance with the user-defined criteria and/or the user-defined conditions;

after sampling only the metadata of the documents, the computer formulating representative subsets of the sampled documents by automatic weighted clustering of the sampled documents utilizing metadata features in the metadata extracted from the documents, wherein the representative subsets of the sampled documents are formulated based on the user-defined criteria and/or the user-defined conditions;

the computer then generating sampled data sets of the sampled documents to automatically reduce a cost of computation resources for processing of the sampled documents; and

the computer balancing the sampled data sets for further processing of the sampled documents,

wherein the formulating the representative subsets is performed by the computer to identify some of the representative subsets for initial processing in response to the user-defined criteria and/or the user-defined conditions.

2. The method in accordance with Claim 1 wherein sampling only the metadata of the documents in the one or more data repositories comprises the computer filtering the documents in the one or more data repositories in accordance with one or more features of the documents, the one or more features of the documents indicated by the metadata of the documents.

3. The method in accordance with Claim 2 wherein the one or more features of the documents indicated by the metadata of the documents comprise folder, department or location in the one or more data repositories, document date of creation, document date of modification, document size, document depth, document language, document extension and/or number of personal identifying information (PII) in a document.

4. The method in accordance with Claim 2 or Claim 3 wherein the one or more features of the documents indicated by the metadata of the documents comprise one or more of the user-defined criteria.

5. The method in accordance with Claim 2 wherein the filtering the documents in the one or more data repositories further comprises filtering the documents in accordance with distributors, and wherein the filtering the documents in accordance with the distributors comprises sampling the documents in a balanced manner in accordance with the user-defined criteria and/or the user-defined conditions.

6. The method in accordance with Claim 5 wherein the filtering the documents in accordance with the distributors includes filtering the documents in accordance with a

path clustering distributor, the path clustering distributor clustering ones of the documents having similar document folder paths together.

7. The method in accordance with Claim 1 wherein the generating the sampled datasets of the sampled documents comprises smart sampling the sampled datasets in either or both of a random sampling mode and a proportional sampling mode, wherein smart sampling comprises one or more of (i) the computer improving the quality of document classification within the sampled datasets, (ii) the computer creating balanced datasets which are subsets of one of the sampled datasets while following a distribution of the one of the sampled datasets, (iii) the computer providing supervised classification of the sampled datasets, and (iv) the computer reducing the cost of computation resources for generating the sampled datasets of the sampled documents.

8. The method in accordance with Claim 7 wherein the generating the sampled datasets of the sampled documents comprises smart sampling the sampled datasets by one of weighted clustering, combined distributors or hierarchical sampling prior to smart sampling the sampled datasets in either or both of the random sampling mode and the proportional sampling mode.

9. The method in accordance with Claim 1 wherein the weighted clustering of the sampled documents comprises weighted k-means clustering of the sampled documents.

10. The method in accordance with Claim 1 wherein the weighted clustering of the sampled documents comprises weighted clustering of the sampled documents in a grid search space.

11. The method in accordance with Claim 10 further comprising the computer defining the grid search space in response to a computed silhouette score.

12. The method in accordance with Claim 11 wherein the defining the grid search space comprises the computer defining the grid search space in response to the computed silhouette score and a best clustering weight and a best sampling weight.