**Solution-design**

**Title: Data Duplication Removal/ DE duplication**

**Abstract (Minimum 100 Words):**

To create a utility that will DE duplicate the records and remove duplicate entries based on rules including threshold values or other certainty factors. This utility will be a web based Client interface which will help in uploading the data import, set the rules and other certainty factors to be implied for removal of duplicate data. It will be a library backed by a web interfaces for fuzzy matching, duplicate record removal and entity-resolution. Hence this will help in creating a cleaner set of data

**Solution Approach (Minimum 200 Words):**

The data cleansing process takes two inputs

1. Data required to be cleaned

2. Rules of cleansing from rules configuration database

This is the area where actual data cleansing processing done based on rules from rules configuration repository and output of this process provides error-free and consistent data that is ready to load into data warehouse. This output data is standardized, uniform, accurate and complete with accordance to business. The cleaned data not only provides data quality but expedite the processing speed and performance of overall ETL process. Hence, will come up with the best rules for datasets to quickly and automatically find similar records, even with very large databases.

We will be clustering the large data sets into a set of smaller chunks and then will perform n-gram and distance edit like algorithms to come up with the threshold values of the possible matches with in all the chunks to figure out and discard the data those fall within certain rule range.

**Design Architecture (Well Explained):**

Data cleaning process will be conducted in following steps briefly.

The steps taken to clean this dataset are:

1. **Pre-processing**: We standardized the date representations like of the date of birth. Validation checks are also performed on the number, date of birth like fields.
2. **Processing:** We defining identification rules based on Record Linkage. Record linkage is highly sensitive to the quality of the data being linked, so all data sets under consideration (particularly their key identifier fields) should ideally undergo a [data quality assessment](https://en.wikipedia.org/w/index.php?title=Data_quality_assessment&action=edit&redlink=1) prior to record linkage. Many key identifiers for the same entity can be presented quite differently between (and even within) data sets, which can greatly complicate record linkage unless understood ahead of time. For example, key identifiers for a man named William J. Smith might appear in three different data sets as so:

|  |  |  |  |
| --- | --- | --- | --- |
| **Data set** | **Name** | **Date of birth** | **City of residence** |
| Data set 1 | William J. Smith | 1/2/73 | Berkeley, California |
| Data set 2 | Smith, W. J. | 1973.1.2 | Berkeley, CA |
| Data set 3 | Bill Smith | Jan 2, 1973 | Berkeley, Calif. |

In this example, the different formatting styles lead to records that look different but in fact all refer to the same entity with the same logical identifier values. Most, if not all, record linkage strategies would result in more accurate linkage if these values were first normalized or standardized into a consistent format (e.g., all names are "Surname, Given name", and all dates are "YYYY/MM/DD"). Standardization can be accomplished through simple rule-based [data transformations](https://en.wikipedia.org/wiki/Data_transformation) or more complex procedures such as lexicon-based [tokenization](https://en.wikipedia.org/wiki/Tokenization_(lexical_analysis)) and probabilistic hidden Markov models. Several of the packages listed in the Software Implementations section provide some of these features to simplify the process of data standardization.

1. **Human verification and validation:** This will have to be done at the user end which will consume the cleaned data set

Difference Representation

Naming Conventions

Spelling variations

Abbreviation

Dirty Data with Duplicate Records

Unit Difference

**CLEANING STRATEGIES/ Record Linkage**

Consistent Data/ Cleaned Data Records

**Business Impact (Minimum 200 words):**

It will help the developers:

• Remove duplicate entries from a data sources of names and addresses

• Link a list with customer information to another with order history, even without unique customer id's

• Take a database of campaign contributions and figure out which ones were made by the same person, even if the names were entered slightly differently for each record

Record linkage plays a key role in [data warehousing](https://en.wikipedia.org/wiki/Data_warehousing) and [business intelligence](https://en.wikipedia.org/wiki/Business_intelligence). Data warehouses serve to combine data from many different operational source systems into one [logical data model](https://en.wikipedia.org/wiki/Logical_data_model), which can then be subsequently fed into a business intelligence system for reporting and analytics. Each operational source system may have its own method of identifying the same entities used in the logical data model, so record linkage between the different sources becomes necessary to ensure that the information about a particular entity in one source system can be seamlessly compared with information about the same entity from another source system. Data standardization and subsequent record linkage often occur in the "transform" portion of the [extract, transform, load](https://en.wikipedia.org/wiki/Extract,_transform,_load) (ETL) process.