**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):**

Difference Representation

Naming Conventions

Spelling variations

Abbreviation

Dirty Data with Duplicate Records

Unit Difference

**CLEANING STRATEGIES**

Consistent Data/ Cleaned Data Records

5.4.1. Dataset description Next, we cleaned a hospital’s patient database containing 22122 records. This database contains 60 fields, including the NRIC number2 , sex, date of birth, date of screening, the clinic number and various fields containing codes for the results of a diabetes screen test. 5.4.2. Data cleaning process The steps taken to clean this dataset are: 1. Pre-processing: We standardized the date representations of the date of birth and date of screening fields. Validation checks were also performed on the NRIC number, date of birth, date of screening and sex fields. 2. Processing: We defined two duplicate identification rules for this database and one merge/ purge rule. The merge/purge rule removes all instances of exact duplicates and keeps only one instance. The rules are described after this. 3. Human verification and validation 5.4.3. Rules used The two duplicate identification rules used are: 1. Duplicate Id. Rule 1: certainty factor 1 \* All 60 fields of the record match. 2. Duplicate Id. Rule 2: certainty factor 0.9 \* Matching date of screening field. \* and matching NRIC number after removing all non-numeric characters. The merge/purge rule used is defined as: 1. Merge/purge rule \* For all duplicate record groups with certainty factor 1, keep one record from the group and delete the rest. 5.4.4. Performance results The alerts raised during the pre-processing stage are shown in Table 6. The pre-processing stage took 118:5 s: The results of the processing stage are shown in Table 7. Manual inspection of the database revealed that we achieved 100% recall and precision for this experiment. Although the records identified by Rule 1 of Table 7 form a subset of those identified by Rule 2, we still include this rule as it identifies exact duplicates, which we can easily use a merge/purge rule to automatically remove all, except for 1 record. The 8 records (4 pairs) that were exact duplicates were automatically processed in this manner. The rest were marked for manual processing. Generally, we can have more complex merge/ purge rules for dealing with more complicated cases. The pre-processing and processing stages (using Rules 1 and 2) took about 10 min for this dataset. 5.5. Comparative study Table 8 compares the methodologies used in IntelliClean with the closest works, namely Fig. 8. The effect of window size on recall. Tabl

**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