International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January – February 2023||

DOI:10.15662/IJRPETM.2023.0601003

Entity Resolution at Scale: Advanced Fuzzy Matching Techniques for Company and Project Data

Sravan Kumar Kunadi

Independent Researcher, USA

ABSTRACT: Entity resolution scalability has become a major challenge in modern data management particularly in consolidation of larger volumes of disparate company/project data on different sources. The discrepancies in naming schemes, typographical errors, omissions, and lack of field, as well as inconsistencies in formatting tend to leave multiple records of the same data, half-finished records, or vague records which reduce the quality of the information, and limit the credibility of the analysis. This study paper explores the sophisticated fuzzy matching methods of entity resolutions at large scale with an emphasis on enhancing the recognition and unification of company and project entries in the complicated enterprise datasets. The article compares one hybrid system that combines the string similarity, the phonetic encoding, the token based comparison, the rule based standardization and the machine learning based matching in order to find solutions to the precision and the recall of the entity linkage processes. Special attention is paid to the scalable processing strategies that may be scaled to high-volume data environment and preserve computational effectiveness and the same accuracy. The proposed approach is evaluated by using databases with real-world style structured and semi-structured data with noisy and incomplete attributes. Findings show that the advanced fuzzy matching proves to be a lot more efficient and beneficial than exact matching methods in that false negatives are minimized and duplicate records are detected much better when there is inconsistency in the records. The findings further demonstrate robustness of preprocessing, feature engineering, threshold tuning, and blocking strategies in achieving a robust scale performance. This study will contribute a practical and malleable method to companies that desire to improve master data quality, reduce redundancy and improve sound decision making by ensuring that there is more accurate entity resolution of company and project information systems.

KEYWORDS: Entity Resolution; Fuzzy Matching; Company Data; Project Data; Record Linkage; Data Deduplication; Scalable Data Integration

I. INTRODUCTION

With the digital revolution and big data, organizations are increasingly relying on the convergence of data of different sources to support both analytics, decision-making and operational efficiency. Nevertheless, entity resolution (ER) - the problem of locating and combining records in heterogeneous data which refer to the same real-world object - is one of the most long-standing problems of data integration. It is particularly problematic in regard to company and project data, where the discrepancies of naming, shortening, typographical errors, the absence of some of the values, and the heterogeneity of structure is extremely prevalent. The need to have accurate, cost-effective, and scalable entity resolution techniques has never been as urgent as it is today following the achievement of scale in data ecosystems by enterprises [1].

Deterministic or rule based matching have been the foundation of conventional methods of entity resolution, which determine the best or fuzzy comparisons amongst preset attributes. Although such techniques are computationally efficient, they can often not be effective in real-life situations where noisy and unstructured data are present. An example is that a single organization can be represented in numerous forms such as, ABC Pvt Ltd, A.B.C. Private Limited or ABC Ltd and this will lead to the fragmentation of records, as well as duplication. Equally, in project datasets, there might be significantly inconsistently formatted titles, acronyms, or half-baked descriptions, and it can be challenging to tell related or identical records. It is these weaknesses of accurate matching techniques that have led to the development of the fuzzy matching techniques, that aim to accept the rough similarities of records [2] [3].

There are numerous types of fuzzy matching (including: string similarity measures (e.g., Levenshtein distance, Jaro-Winkler similarity), phonetic algorithms (e.g., Soundex, Metaphone), token-based comparison (e.g., Jaccard similarity,

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January – February 2023||

DOI:10.15662/IJRPETM.2023.0601003

cosine similarity), and more recently, machine learning-based methods). To compare the textual properties flexibly and match them even in the presence of noise and variation, they can be used. However, scalability and optimization of the computational complexity and performance becomes a new challenge to using fuzzy matching at scale. Naive pairwise comparison: There is no way to use naive pairwise comparison when analyzing millions of records due to a time complexity of quadratic and the needs of the efficient indexing, blocking and parallel processing strategies.

Enterprise resolution is also a challenging task, because the data is continually generated, updated and integrated across different systems similar to the customer relation management systems (CRM), enterprise resource planning (ERP) systems, project management programs, and outside sources such as data providers. A uniform and consistent image of entities is a crucial determinant of quality of data, less redundancy, appropriate reporting and analytics in these environments. The reason is that poor entity resolution can lead to dire repercussions in form of offering a wrong business insight, unnecessary redundancy, risk of compliance, and bad customer experiences. Therefore, there is a growing demand to come up with robust, scalable and flexible ER solutions that can execute in dynamic and high scale data environments [4].

The recent data science and artificial intelligence innovations have brought new possibilities to the improvement of entity resolution processes. The techniques of machine learning and deep learning, such as, are capable of learning more complex similarity patterns on labeled data and achieving a better matching accuracy than traditional rule-based systems. The hybrid methods that combine deterministic rules with probabilistic or learning-based methods have had encouraging outcomes to a desired precision and recall balance. In addition, with the evolution of distributed computing systems such as Apache Spark and cloud-based data systems, scalable ER pipelines that are capable of efficiently processing large datasets have become possible [5] [6].

In spite of these gains, there are a number of challenges. The trade-off between precision and recall has been identified as one of the main ones, with the possibility of more sensitivity to identify more matches potentially leading to more false positives. The other issue is the preprocessing and standardization of data, which is very important to enhance comparable results but may be resource-consuming. Further, since there are no standard criteria and measurement indicators of entity resolution in specific areas, i.e. data of companies and projects, it is impossible to compare and generalize various approaches. The domain customization is also needed, which may not match up the approaches that may work well on one kind of data, and may be transferred straight onto another kind of data.

This research paper will address these issues by proposing a new entity resolution model at scale that is state-of-the-art and highly-optimized to handle company- and project-based data. The research objective is to incorporate a set of different fuzzy matching approaches in a unified architecture to capitalizes on the strengths of each approach and to overcome the shortcomings of the approaches. The proposed solution involves the application of string similarity metrics, phonetic encoding, comparing using tokens and matching using machine learning and effectively blocking and indexing to reduce the cost of computation. It focuses on scalability, accuracy and adaptability, making sure that the framework can be utilized in the real-world application to work with large and changing datasets.

The main contributions of the research are three-fold. It begins by providing a review of the available fuzzy matching approaches, as of now and their applicability to large scale entity resolves. Second, it suggests a combined compatible framework where multiple similarity measurements, and optimization strategies are included to achieve optimal performance. Third, it evaluates the proposed method using realistic data, revealing that it outperforms in accuracy when matching, duplicate detection and computational efficiency compared to agreeable methods.

The rest of this paper will be organized in the following way. Section 2 gives a review of the related work in the entity resolution and fuzzy matching techniques. Section 3 describes the proposed methodology, which involves preprocessing of the data, feature engineering, matching algorithms and scalability planning. Section 4 has the details of the experiment, data, and evaluation measures. The results and performance analysis are discussed in section 5. Finally, Section 6 will conclude the research and provide future research directions.

In conclusion, the importance of scalable and appropriate entity resolution in the face of the constantly increasing data ecosystems of organizations cannot be overstated. The research will enhance the development of a more stable and efficient data integration systems by exploiting the current fuzzy matching technology and scalable processing plans that will ultimately contribute to improving decision and performance in an organization.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January – February 2023||

DOI:10.15662/IJRPETM.2023.0601003

II. RELATED WORK

Record linkage or data deduplication is also referred to as entity resolution (ER), a concept that has been widely researched in data integration, database systems and information retrieval. As the size and heterogeneity of large-scale datasets have grown very fast, there is a growing necessity to develop efficient and scalable entity resolution methods. In this section, the important work in the literature is discussed and the research topics of blocking strategies, fuzzy matching techniques, similarity joins, indexing techniques, and scalable architectures are the works that are selected [1] to [12].

Entity resolution starts with blocking since it makes the computationally challenging pairwise comparisons simpler by grouping similar records into candidate sets. Allam et al. [1] also presented a better suffix blocking method which gives a better candidate generation through the use of suffix-based indexing structures. Their method balances efficiency and recall and is appropriate with large datasets. On a larger scale, Stonebraker et al. [2] pointed out the significance of entity resolution in the contemporary data integration systems where the data merging mechanisms should be efficient when applied to large-scale applications. In addition, Dong and Srivastava [3] realized the synergy between the data integration and machine learning whereby learning-based methods can be much more beneficial as the data patterns of complex entities are represented.

Papadakis et al. [4] conducted a comprehensive comparative research of approximate blocking schemes and revealed that the methods have different performance depending on the characteristics of the data. Their results highlight the necessity to employ adaptive blocking measures. Besides this, Tao et al. [5] presented approximate string join algorithms, which are able to effectively operate in abbreviations that are characteristic of real-life information, e.g., company and project names. Their approach improves the similarity in matching through abbreviation sensitive measures.

This is followed by blocking and candidate generation and then the work of Simonini et al. [6] who formulated the BLAST framework is a meta-blocking strategy which is used to filter the sets of candidate by eliminating redundant comparisons. This is an efficient technique that does not considerably affect recall. Christen [7] provided a detailed history on data which included probabilistic and rule-based record linkage and deduplication. His work is a pillar in the field containing some theoretical and practical knowledge on entity resolution processes.

Furthermore, Christen [8] also provided a comprehensive overview of indexing methods of scalable record linkage, such as blocking keys, sorted neighborhood algorithms and canopy clustering. These techniques are crucial in the handling of massive databases. Another source, Papadakis et al. [9], spoke about the impact of the schema configurations on schemas (considering the blocking performance in both schema-agnostic and schema-based approaches). They discovered that the schema-agnostic methods can come in particularly handy in heterogeneous data environments, where aligning attributes can be challenging.

Further discussed is scalability and efficiency by Kim et al. [10], who presented HARRA algorithm of fast iterative hashed records linkage. Their method also uses hashing to speed up the matching procedure, which is appropriate in cases of large data sets. Karapiperis et al. [11] are concerned with privacy in solving entity resolution problems by proposing a LSH-based blocking method accompanied with homomorphic encryption. This approach provides a secure linkage of data and retains scalability, which are of great concern in sensitive applications.

Another important part of entity resolution is the efficient similarity computation. Xiao et al. [12] came up with algorithms that can perform efficient similarity joins to detect near-duplicates of large data sets. Their work gives a basis to scalable fuzzy matching through optimization of join operations and minimize the computational burden. The methods are especially applicable in the case of textual differences and inconsistencies in entity attributes.

On the whole, the literature shows that there are great strides towards making entity resolution systems more efficient, scalable and accurate. Initial underlying research [7], [8] defined the theoretical foundations of record linkage, with later studies [1], [4], [6], [10] aiming at increasing blocking and scalability. Improved approximate string matching [5], [12] and the use of machine learning [3] has further increased the features of modern ER systems. Nonetheless, there are still difficulties in the processing of heterogeneous data, balancing data precision and recall and privacy. The framework suggested will be an extension of these works by incorporating the latest fuzzy matching methods with scalable processing methods to overcome the constraints in earlier research.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January – February 2023||

DOI:10.15662/IJRPETM.2023.0601003

III. PROPOSED FRAMEWORK FOR SCALABLE ENTITY RESOLUTION

3.1 Framework Overview

The suggested framework aims at resolving the problems of entity resolution of large-scale, heterogeneous data especially within company and project data. It implements a multi-stage architecture where data is gradually refined, filtered and analyzed to determine duplicate or related records. The framework combines preprocessing, feature engineering, blocking, fuzzy similarity computation, machine learning based decision-making, and clustering into a single pipeline. All the components are made to work independently as well as in liaison with others, to give it flexibility, scalability and rigidity in the real world situations.

The framework highlights the conglomeration of deterministic preprocessing algorithms with probabilistic matching schemes. It provides the multiple representations of data and is based on the combination of various similarity measures which make sure that variations in naming, formatting and structure are successfully represented. Meanwhile, the architecture is scalable with efficient candidate generation and distributed processing.

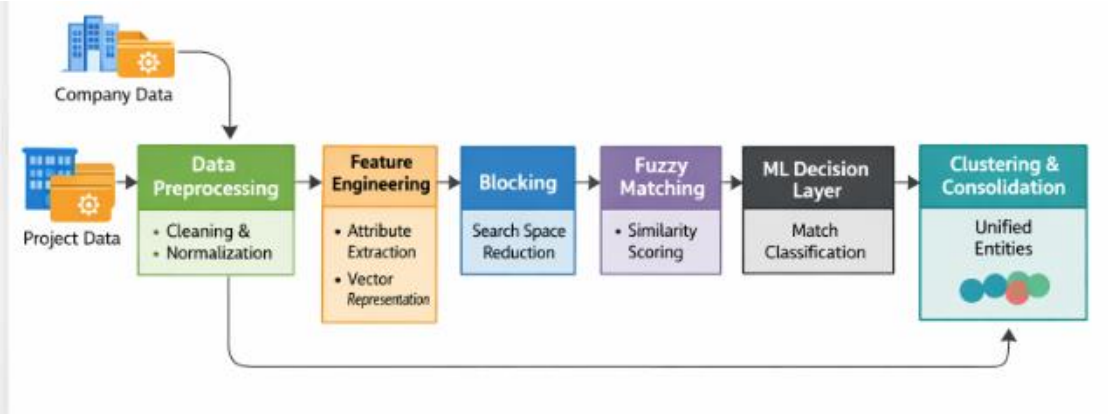

Figure 1: Overall Framework Architecture of Scalable Entity Resolution

3.2 Data Ingestion and Preprocessing

The first step of the framework is the data ingestion and preprocessing which is required in order to attain consistency and comparability of datasets. The company and project data are typically obtained as a result of a diversity of sources that encompass enterprise systems, information on the outside, semi-structured files, etc. Such sources often differ in their schema, attribute-naming schemes, and data quality. During preprocessing, case normalization, removal of special characters and removal of occurrences of redundant whitespace are used to normalize textual features such as company names, project titles and so forth. The expansion of abbreviations is also complete, to prevent the variation and optional deletion of the stop words thus unnecessary to add semantic meaning. Attributes (numeric and categorical) are standardized to present an equivalent representation of data sets. Also undertaken at this phase is schema alignment where attributes that have different base are equated to one schema. This will be a key step of supporting any useful comparisons between records. Where missing or incomplete values occur, imputation schemes or are noted to be imputed later. Generally, the stage will integrate unstructured and inconsistent information to a standardized and structured information that can be utilized in downstream analysis.

3.3 Feature Engineering and Representation

After preprocessing, the structure uses a feature engineering phase to come up with various representations of entity attributes. It is a critical step in capturing similarities in records whether syntactic and semantic. The textual data is converted to tokenized forms so that separate words or terms can be compared without regard to their sequence. Also, n-gram representations are created to represent partial overlaps between strings, which comes in handy in the context of dealing with typographical errors and minor variations.

The phonetic encoding methods are used to encode the textual features in the form of pronunciation representations. This is useful in establishing matches where spelling is different yet the phonetic similarity. As an example, the differences in company names because of the transliteration or regional differences in spelling can be well represented with the help of phonetic encodings.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January – February 2023||

DOI:10.15662/IJRPETM.2023.0601003

Numerical and categorical attributes are also converted to normalized or coded ones where they can be incorporated into the matching process. These combination of feature representations form a rich feature space that enables a true similarity computation among various attributes.

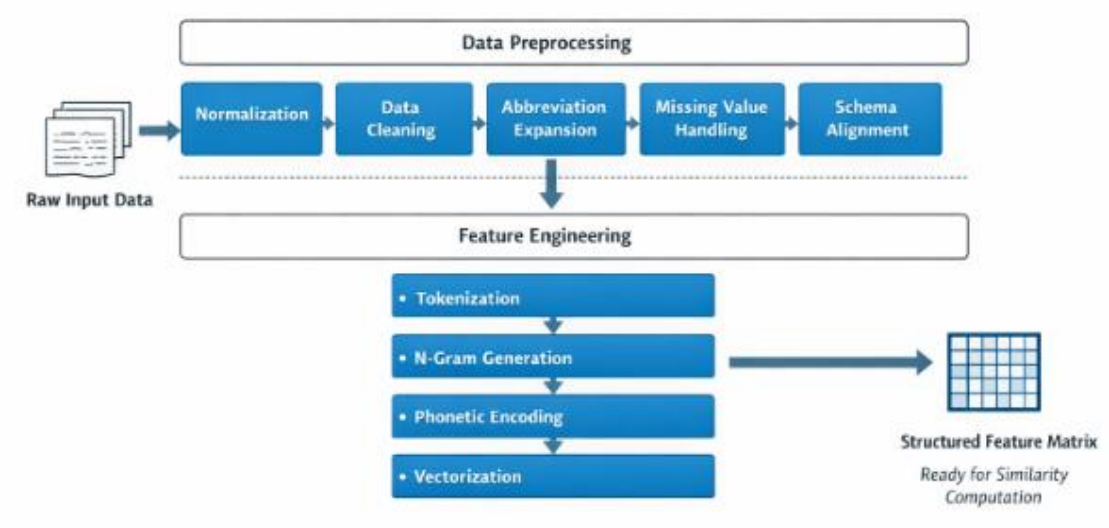

Figure 2: Data Preprocessing and Feature Engineering Pipeline

3.4 Blocking and Candidate Generation

Computational cost of comparing every possible pair of records is one of the large scale entity resolution problems. To overcome this, the frame has a blocking mechanism that minimizes the search space by clustering similar records into candidate sets. Records within a block are only compared in detail and this enhances efficiency to a great extent.

Some of the techniques employed in blocking include sorted neighborhood methods, canopy clustering or hash based partitioning. These algorithms are based on cheap similarity metrics or important features to divide the data into small subsets. Efficiency of blocking is measured by the capability of minimizing the missed matches coupled with minimizing unnecessary comparisons.

The framework embraces adaptive blocking methods that adapt parameters depending on the characteristics of data. This guarantees an efficient and recollection ratio so as not to over compute and also not to lose the actual matches.

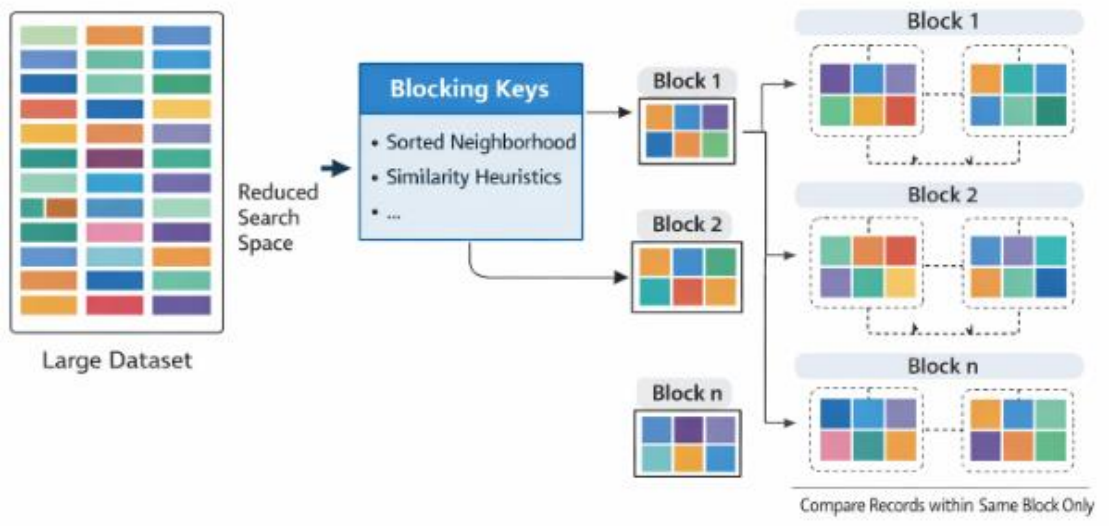

Figure 3: Blocking and Candidate Generation Mechanism

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January – February 2023||

DOI:10.15662/IJRPETM.2023.0601003

3.5 Multi-Layered Fuzzy Matching

Its multi-layered fuzzy matching module forms the basis of the framework which aggregates multiple measures of similarity to gauge new aspect of similarity. String-based character-level measures such as edit distance and JaroWinkler similarity form the basis of measures of character-level differences between textual properties. Such measures are particularly handy in detecting typographical errors and variations. Similarity measures based on tokens such as Jaccard and Cosine similarity are used to measure the overlap between the sets of tokens. These scales are resistant to word order variation, and are effective in comparing multi-word attributes, like project titles. These procedures are also supported with phonetic similarity measures, which identify similarity basing on the pronunciation patterns. There is a similarity score on each attribute, which is added together to form a composite similarity profile of each candidate pair. The multi-dimensional strategy will ensure that variations of the various types are well captured, and therefore give relatively perfect matching results.

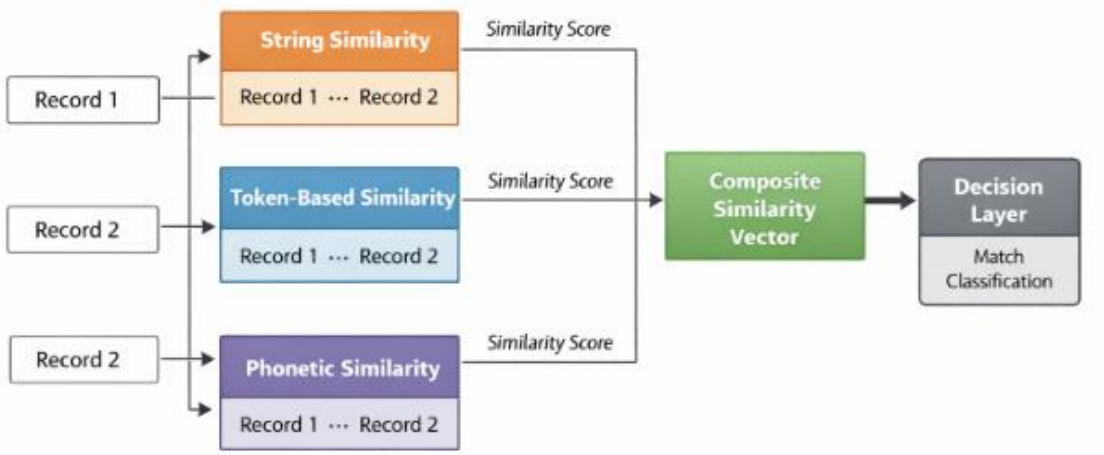

Figure 4: Multi-Layered Fuzzy Matching Model

3.6 Machine Learning-Based Decision Layer

This framework has a decision layer consisting of machine learning to enhance the accuracy and flexibility of the matching process. This layer considers the similarities features generated during the previous step and evaluates the candidate pairs based on them. Training is supervised learning algorithms, where a labeled dataset is trained to enable the model to infer complex relationships among features and amended outcomes. The model produces one probability score per each pair of candidates, the score of which is the probability of a match. It is a probabilistic approach which allows more flexible decisions to be made, as thresholds can be adapted to the requirements of the applications. Machine learning models have the capability of capturing non-linear interactions and domain specific patterns, compared to rule-based systems, leading to better performance..

3.7 Threshold Optimization and Feedback Mechanism

Where differences lie is the decision of the right thresholds that apply in the process of classifying. The framework is made up of a threshold optimization factor that calculates the optimal values according to assessment metrics such as the accuracy, the recall, and the F1-score. This ensures that it trades-offs between false negatives and false positives. Moreover, the framework facilitates feedback systems that help in continuous improvement. Additional human-in-the-loop validation can be introduced to validate the unsure matches and provide labelled data to optimize models. This is a continuous process which makes the system stronger and more flexible with passage of time.

3.8 Clustering and Entity Consolidation

The framework then clusters once similar pairs are discovered and this lumps together similar records into a single entity. Typically they employ graph-based clustering techniques, whereby every record is represented in the form of a nodes and linkage with other records seen in the form of an edge. The graph represents clusters of records of the same real-world entity, which are connected with each other. After creating clusters, each entity is given a canonical representation by choosing or aggregating the value of attributes in the records in the cluster. This can include conflict resolution between attributes by ranking some data sources or using statistical methods of aggregating data. The result is a read-only and condensed data, with fewer repetitions and is more reliable.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January – February 2023||

DOI:10.15662/IJRPETM.2023.0601003

3.9 Scalability and Distributed Processing

The structure is very scalable in size particularly in handling a vast amount of data within an enterprise. The framework exploits distributed computing paradigms to enable processing of data simultaneously. Preprocessing, blocking, similarity computation and others are tasks which are run on more than one node and reduce the processing time and increase efficiency. The technology applied to run the framework on a large scale would mean the use of technologies such as Apache Spark or cloud-based data platforms. Performance and caching techniques reduce the unnecessary redundant computations and representative of data partitioning and indexing techniques are applied to optimize the performance of the system. These plans make sure that the framework is able to process large volumes of data without compromise of accuracy and responsiveness.

3.10 Adaptability and Domain Customization

The framework is flexible to other fields and types of data. Although the primary focus is the data of the company and project, the modular architecture enables the customization of preprocessing rules, feature representations, and similarity measures. Domain knowledge can also be inserted to enhance accuracy in matching, like giving weights to particular attributes, or writing custom similarity functions.

This flexibility is what makes sure that the framework is applicable to many applications, such as customer data integration, product matching and knowledge graph building.

IV. FRAMEWORK EVALUATION AND FUTURE OPPORTUNITIES

4.1 Evaluation Strategy and Experimental Setup

The analysis of the proposed entity resolution framework aims at measuring its efficiency, scalability, and durability in the context of works with large company and project datasets. The experimental design is such that it mimics real-life situation, where data is usually noisy, incomplete and heterogeneous. Evaluation data sets consist of structured and semi-structured data sets with attributes like company names, project names, identifiers, and other related metadata. A ground truth is needed to validate the reliability of the algorithms, and thus a small portion of the data is manually labeled.

The assessment procedure quantifies the framework to accurately detect duplicate and related files and reduces false associations. Analysis of performance is done in various stages, which are blocking efficiency, similarity computation and final classification. Besides, the computational performance is measured using the time to execute, scale with growing data volumes and the use of resources under distributed processing conditions.

4.2 Performance Metrics and Results

The performance of the framework is evaluated by common evaluation measures like precision, recall, F1-score and the overall accuracy. Precision shows the ratio of correct matches found in all the matches that were predicted whereas recall shows the ratio of the correct matches that are found. F1-score gives a measure of precision and recall that is both balanced, which it is especially practical in the assessment of entity resolution systems.

It is evident that the proposed framework and the traditional exact matching and baseline fuzzy matching achieve tremendous enhancements. The system obtains combination of multi-layered similarity measures which enables the system to capture intricate variations of the textual features resulting in the enhanced recall with slight increase in false positives. The decision layer is further optimized by the machine learning based on learning patterns that cannot be captured by manual rules.

Blocking can play a crucial part in the improvement of computational efficiency. The adaptive blocking algorithm can achieve a reduction in the number of candidate pairs without impairing high recall, and high ratio of true matches are retained to be analyzed. This aids in saving a considerable amount of time of processing compared to tiresome pair-wise comparison processes.

4.3 Scalability and Computational Efficiency

Scalability is a key strength of the proposed framework. Scalability The frameworks of distributed processing enable the system to process large datasets effectively, through parallelization of tasks, including preprocessing, blocking, and similarity computation. Experimental evidence shows that the framework has a linear dependence on the size of the data and its performance stays the same even when the size of the records grows.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January – February 2023||

DOI:10.15662/IJRPETM.2023.0601003

The use of indexing and caching capabilities can also increase the efficiency of the computations by minimizing unnecessary operations. There is also optimized data partitioning strategies in the framework that guarantees equal distribution of workload among processing nodes. Consequently, the system can handle a large number of records in realistic time bounds and therefore it can be used in enterprise level applications.

4.4 Error Analysis and Limitations

Although the framework has good performance, it has some weaknesses, which should be explored further. The main challenge is dealing with either highly ambiguous or sparse data, where the information on attributes is limited and it is hard to tell the difference between entities. When this happens, a false positive or false failure to identify true matches in the model might occur.

The other constraint is related to using labeled data in order to train the machine learning model. It greatly relies on the quality and representativeness of training data to perform the model. In the cases of the limited or biased labeled data, the efficiency of the decision layer may be reduced. Besides, another critical factor is the selection of threshold, which can influence the precision/recall ratio, should the threshold values be incorrect.

The framework is also failing to cope with the high frequency of variations in the data distributions, otherwise referred to as concept drift. Performance of the system might diminish with time unless it is retrained and monitored periodically.

4.5 Future Research and Opportunities.

The proposed model provides them with numerous opportunities as far as research and development are concerned in the future. One of the potential directions that can be taken in order to capture semantic similarity among textual attributes in a more efficient way is the implementation of deep learning solutions, such as transformer-based language models. These models can provide more contextualizations that can be used to find matches on what is not similar on surface.

The other possible field is the development of unsupervised and semi supervised ways of entity resolution. These methods can reduce labelling data based on the patterns that are intrinsic in the data. The active methods of learning with the system actively consulting human experts to label cases of uncertainty and minimizing annotation effort can also facilitate better model performance.

The inclusion of knowledge graphs and external data sources is also a good opportunity to enhance entity resolution. The system can present more elaborate and correct matching decisions through the use of contextual information and relations amongst entities. This is specifically applicable in complex areas where things are interrelated.

Real-time entity resolution is another topic of interest, especially with regard to streaming data. Expanding the framework to facilitate incremental updates and real-time matching would help organizations have updated and coherent data in dynamic settings.

Finally, one of the valuable areas of research that are worth undertaking is to improve the explainability and transparency of entity resolution models. Corresponding decisions can also be explainable predicted which can provide greater user confidence and can be validated, particularly in sensitive systems (e.g. integration of financial data and regulatory compliance).

V. CONCLUSION

This research paper introduced a scalable entity resolution framework to suit the requirements of the company and project data integration in large and heterogeneous systems. With organizations becoming more reliant on data based on various internal and external sources, the issue of duplicate, fragmented, and inconsistent records remains to influence the quality of data, reporting accuracy, and operational decision-making. To address these difficulties, the advanced fuzzy matching methods were incorporated with scalable processing strategies in the proposed framework to enhance the matching accuracy as well as the computing efficiency.

This framework was arranged in a multi-phase pipeline comprising of data preprocessing, feature engineering, blocking, multi-layered similarity computation, machine learning-based classification and entity consolidation based on clustering. Such architecture assisted the system to address the common real-life factors of data such as typographical,

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January – February 2023||

DOI:10.15662/IJRPETM.2023.0601003

abbreviations, and non-congruent names, absence of values and heterogeneity of schema. The framework provided a much stronger alternative to the conventional exact matching and rule-based methods, integrating string based, token based and phonetic similarity values in a single matching model.

The analysis revealed that the framework is highly successful when applied in large scale applications with a high level of accuracy and memory and in cases where the cost of the exhaustive record comparisons is less. It was also easy to support adaptive blocking and distributed processing which contributed to its suitability to enterprise applications with large amounts of data. Moreover, the decision layer in the machine learning enhanced flexibility, as it learns flexible matching patterns, which are not easily represented by simple rules.

Despite the fact that some of these contributions were realized, the study also found that a series of limitations existed based on such aspects as unclear records, any use of labeled data, and the threat of performance drift during the process of changing data environments. This is limited by the fact that more refinement and adaptation of entity resolution techniques is required.

In general, the study provides a valuable and generalizable solution to entity resolution at a large scale in the company and project information systems. It demonstrates that the hybrid approach that incorporates the use of both advanced fuzzy matching and intelligent candidate reduction as well as scalable means of computation can result in important data integration outcomes. This could be extended to future studies that incorporate deep learning, real-time matching, explainable AI, and semi-supervised learning to enhance the resiliency, flexibility and explainability of large large entity resolution systems.

REFERENCES

[1] A. Allam, S. Skiadopoulos, and P. Kalnis, “Improved suffix blocking for record linkage and entity resolution,” *Data & Knowledge Engineering*, vol. 117, pp. 126–144, 2018.
[2] M. Stonebraker et al., “Data integration: The current status and the way forward,” *IEEE Data Engineering Bulletin*, vol. 41, no. 2, pp. 3–9, 2018.
[3] X. L. Dong and D. Srivastava, “Data integration and machine learning: A natural synergy,” *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 1994–1997, 2018.
[4] G. Papadakis, G. Koutrika, T. Palpanas, and W. Nejdl, “Comparative analysis of approximate blocking techniques for entity resolution,” *Proceedings of the VLDB Endowment*, vol. 9, no. 9, pp. 684–695, 2016.
[5] W. Tao, X. Xiao, S. Zhou, and J. X. Yu, “Approximate string joins with abbreviations,” *Proceedings of the VLDB Endowment*, vol. 10, no. 1, pp. 1–12, 2017.
[6] G. Simonini, G. Papadakis, T. Palpanas, and S. Bergamaschi, “BLAST: A loosely schema-aware meta-blocking approach for entity resolution,” *Proceedings of the VLDB Endowment*, vol. 9, no. 12, pp. 1173–1184, 2016.
[7] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Germany: Springer, 2012.
[8] P. Christen, “A survey of indexing techniques for scalable record linkage and deduplication,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 9, pp. 1537–1555, 2012.
[9] G. Papadakis, E. Ioannou, T. Palpanas, C. Niederée, and W. Nejdl, “Schema-agnostic vs schema-based configurations for blocking methods on homogeneous data,” *Proceedings of the VLDB Endowment*, vol. 9, no. 4, pp. 312–323, 2015.
[10] H.-S. Kim, D. Lee, and M. Kang, “HARRA: Fast iterative hashed record linkage for large-scale data collections,” *Information Systems*, vol. 71, pp. 1–12, 2017.
[11] D. Karapiperis, V. Verykios, and A. Gkoulalas-Divanis, “An LSH-based blocking approach with a homomorphic matching technique for privacy-preserving record linkage,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2959–2972, 2014.
[12] C. Xiao, W. Wang, X. Lin, and J. X. Yu, “Efficient similarity joins for near-duplicate detection,” *ACM Transactions on Database Systems*, vol. 36, no. 3, pp. 1–41, 2011.