Mapping of Financial Services datasets using Human-in-the-Loop

Shubhi Asthana
sasthan@us.ibm.com
IBM Research
San Jose, California, USA

Ruchi Mahindru
rmahindr@us.ibm.com
IBM Research
Yorktown Heights, New York, USA

ABSTRACT

Increasing access to financial services data helps accelerate the monitoring and management of datasets and facilitates better business decision-making. However, financial services datasets are typically vast, ranging in terabytes of data, containing both structured and unstructured. It is a laborious task to comb through all the data and map them reasonably. Mapping the data is important to perform comprehensive analysis and take informed business decisions. Based on client engagements, we have observed that there is a lack of industry standards for definitions of key terms and a lack of governance for maintaining business processes. This typically leads to disconnected siloed datasets generated from disintegrated systems.

To address these challenges, we developed a novel methodology DaME (Data Mapping Engine) that performs data mapping by training a data mapping engine and utilizing human-in-the-loop techniques. The results from the industrial application and evaluation of DaME on a financial services dataset are encouraging that it can help reduce manual effort by automating data mapping and reusing the learning. The accuracy from our dataset in the application is much higher at 69% compared to the existing state-of-the-art with an accuracy of 34%. It has also helped improve the productivity of the industry practitioners, by saving them 14,000 hours of time spent manually mapping vast data stores over a period of ten months.

CCS CONCEPTS

• **Information systems** → *Business intelligence*; *Data extraction and integration*; *Data extraction and integration*; **Entity relationship models**; **Semi-structured data**; **Data model extensions**; **Data analytics**.

KEYWORDS

Data Mapping, Financial Services, Data Analytics, Human-in-the-Loop, Clustering

ACM Reference Format:
Shubhi Asthana and Ruchi Mahindru. 2022. Mapping of Financial Services datasets using Human-in-the-Loop. In *3rd ACM International Conference on AI in Finance (ICAIF '22), November 2–4, 2022, New York, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3533271.3561705

ICAIF '22, November 2–4, 2022, New York, NY, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9376-8/22/10…$15.00
https://doi.org/10.1145/3533271.3561705

1 INTRODUCTION

As data engineers seek to access modern financial services to manage their datasets, one of the greatest barriers to doing so is the lack of integration between different systems handling the datasets. Datasets across systems, which are otherwise available in silos, need to be brought together as integrated knowledge, manipulated, and exposed as ready-to-use knowledge for faster and more effective analytics and inferencing. To realize the full potential of the data sources and provide a comprehensive analysis of how they relate to each other, it is necessary to utilize a data mapping engine. For example, a quick look at the comprehensive analysis of a financial services dataset can provide insights into potential risks that can occur. Providing a comprehensive analysis helps businesses use data mapping for effective data analytics by relating the columns in the datasets to each other. While many tools are available to process datasets by clustering and supervised machine learning, relatively few tools try to relate the columns inside the dataset to each other, in the context of pre-existing knowledge around the data sources. Relating the columns helps users get a clearer picture of the data flowing between datasets, and can be used to build data analytics and insights around it. They lack a generic mapping mechanism to relate the columns. Interpreting the columns in the data and mapping them can be a manually laborious task. This complex problem is central to data mapping and merging. [2, 14].

In the current state of the art, there have been a few solutions targeted toward a cognitive approach for mapping datasets. Different sub-tasks of data mapping including schema mapping, columns values mapping, and ontology construction have been captured in the prior art. These solutions include building a graph-based data mapping model using pattern recognition [12]. Ontology mapping, schema mapping [1, 4, 6, 13] are other data mapping methods used for performing this data mapping. Commonly used data mappings are available in form of entity-relationship diagrams in toolsets and are widely used in industry too. Another solution involves using the Unified Modeling Language (UML) extension, an ontology-based approach for mapping datasets [8, 11, 17]. Yamada et. al. in [16] provides an interactive service system that self-organizes data schema, facilitating exploratory data visualization. While this graph structure-based work is novel, it does not serve the purpose of providing a mapping between all the columns in the datasets. We have applied an ensemble of techniques like ID-based, semantic as well as relational mapping along with analysing extracted keywords covering both structured and unstructured data to provide mapping for maximum number of columns.

Additionally, the prior art focuses on entities extracted from the data versus focusing on keywords in order to broaden the mapping, which is our intention. For the purpose of this paper, we define *entity* as encompassing column labels and their values, ID definitions, and keywords extracted from unstructured data.

2 CHALLENGES

As part of the problem understanding and requirement gathering phase of our project, we interacted with several customers in the financial services. Below we share a compiled list of challenges that must be taken into consideration to realize the importance and full potential that data mapping capability can bring to our customers:

(1) ***Lack of Standardization:*** It has been our observation that there is a lack of standardization at multiple levels, both within and across industries. The different organizations within the same industry may utilize different entity definitions for the terms that are synonymous with similar semantic meanings. Similarly, within an organization, different departments may utilize one entity versus another department may utilize different entity definitions for the same term. For example, in a financial organization "transactions","contracts", "deal" are used interchangeably by the industry. Another example includes "Purchase order","PO","sales order" etc.

(2) ***Manual Mapping and Maintenance:*** As a result, the data engineers need to translate the definitions between data warehouses, if they want to be able to map them to create an augmented dataset for more comprehensive data analysis. This can take hours to weeks depending on the number of columns and the vastness of datasets. Based on our discussion with clients, such translation is typically performed and managed manually. Such laborious tasks are prone to errors and wastage of time spent on such mundane tasks, causing an adverse impact on productivity.

(3) ***Lack of Governance:*** Enterprise service providers layout processes to be followed for maintaining and monitoring contract data. However, over time different departments may digress from the business process, and there is difficulty in governing the data across departments and organizations. Additionally, departments may choose to maintain data manually leading to subjectivity, errors, and inaccuracies that may be hard to pinpoint.

(4) ***Lack of Integration:*** The challenge with large organizations handling data warehouses is that there are disintegrated systems resulting from loosely integrated departments. For example, a department may be dedicated to order management, another department may be responsible for contracts management, and yet another department may be responsible for managing invoices billing, and customer data. Each such department may have its own processes, leading to separate data maintenance and monitoring, hence, a disconnect in the end-to-end process. In an enterprise setting, different departments manage different parts of the data, further complicating the communication around orders and invoices. Hence, there is a need for a trustful data mapping engine that can connect and link the different data tables. In turn, this can enable service providers to monitor data consistently and hence provide alerts for potential issues based on dynamic analysis of data.

With the understanding of the challenges shared above, we designed the approach for data mapping as discussed in the next Section 3.

3 APPROACH

In order to address the previously mentioned limitations and inefficiencies, this paper contributes a semi-supervised approach for data mapping called Data Mapping Engine (DaME) shown in Figure 1, described as below.

The proposed system comprises of three major steps namely *Data Extraction and Processing*, *Data Mapping Engine*, and *Human-in-the-Loop*. Each of the steps are detailed below in their respective Sub-Sections.The pseudocode of DaME algorithm is shown below in Algorithm 1

Algorithm 1 Pseudocode for DaME engine

```
for ALL STRUCTURAL DATA do
    Case 1: ← $LabelMapping$
    Case 2: ← $ID - basedMapping$
    Case 3: ← $SemanticMapping$
    Case 4: ← $RelationMapping$
    return columnMapping
end for
for ALL UNSTRUCTURAL DATA do
    Case 1: ← $KeywordClustering$
    Case 2: ← $ThemeClustering$
    return columnMapping
end for
for ALL COLUMN MAPPING DATA do
    Case 1: ← $ComputeColumnWeight$
    Case 2: ← $ConflictResolution \leftarrow max(columnWeight)$
end for
```

3.1 Data Extraction and Processing

Input to the proposed system is a *Data Warehouse* comprising multiple datasets. Such datasets may contain structured or unstructured data, where each type of data requires different extraction and processing steps, which are explained next. For example, a structured dataset from financial services domain contains tables for *Orders*, *Contracts*, and *Invoices*, while unstructured data may be free form natural language comments, email exchanges, etc., see examples shown in Table 1.

Structured Data: The column labels and their values are extracted from structured data available in tabular form. Next, both the column labels and their corresponding values are validated. For example, if a column label or the value in that column is *null*, then it is not considered as input to the model training.

Unstructured Data: Unstructured data like communication text with customers, email correspondences for renewals or disputes, etc. are also leveraged. The unstructured data is parsed using System-T [7] to extract the keywords. Examples of unstructured data like comments and the respective keywords extracted are shown in Table 1.

The validated column labels, values, and the extracted keywords are input to *Data Mapping Engine*, described in the next Sub-Section 3.2.

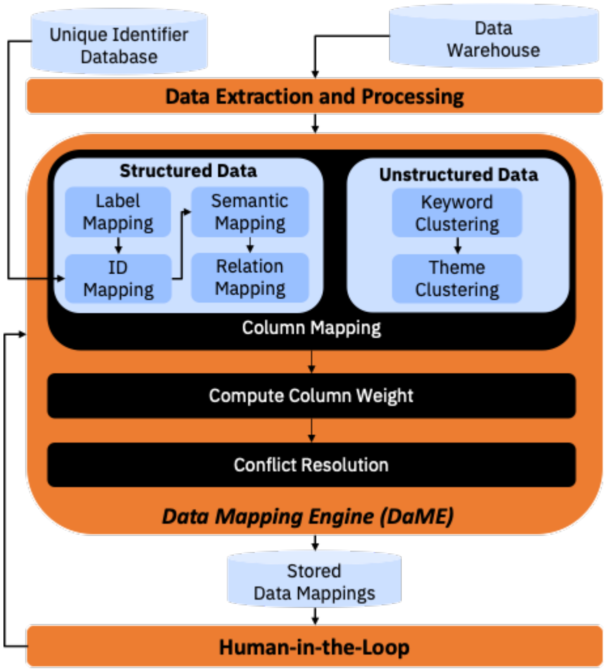


Figure 1: System Overview of DaME: Data Mapping Engine

Table 1: Examples of Customer Comments from Unstructured Data and Keywords Extracted

Comments	Keywords
Order is already cancelled and invoice settled	order, cancelled, invoice, settled
this order was cancelled by customer so no further invoices will be issued.	order, cancelled
Order terminated by ticket to representative	order, terminated, ticket, representative
Notification sent to customer@abc.org for settlement of invoices	invoices, settlement, notification, customer
This order will only autorenew on 09/30/2022, so there is no reason to contact the customer.	order, autorenew
Customer disputed the cloud services product	customer, disputed, cloud services product
Please cancel this contract and bill remaining invoices for server	contract, cancel, invoices, bill, server
Customer contact about order out of funds	customer, contact, order, out of funds
This order has already autorenewed and customer paid for the new invoice full amount	order, autorenewed, invoices, full amount, paid

3.2 Data Mapping Engine

Data Mapping Engine consists of two pipelines, one to handle *structured* data and another for *unstructured* data. Each pipeline is a novel and experience-driven ensemble of several techniques employed for identifying data mappings with high accuracy.

3.2.1 Structured Data. For structured data, four different techniques have been defined along with their confidence level in mapping identification, as shown in Table 2. The techniques are applied sequentially based on the confidence associated from *High* to *Low*. For example, *Label Mapping* technique is the most accurate, as it is an exact match on the column headers across multiple datasets, hence the confidence level associated with it is *High*. Therefore, it is applied first on the structured data input. Similarly, *ID Mapping* also has a *High* confidence as it is a direct match on the unique identifier and then the additional columns with more descriptive content are used for data mapping. Given that *Semantic* and *Relation Mapping* are fuzzy matches involving mathematical modeling of column values, hence the associated confidence level is *Medium*. Such confidence level is used to resolve issues with multiple data mappings as described in Section 3.2.4.

(1) ***Label Mapping:*** This is the most simple and accurate technique as it is syntax-driven, where the column labels that are identical across multiple datasets are mapped together. For example, CustomerName in Order Table and customer_name in Invoice table are mapped together as they are identically similar.
(2) ***ID-based Mapping:*** Next, for the remaining unmapped column labels, the ones that are dedicated to storing unique identifier information are considered for *ID Mapping*, which is also a syntax-based technique. *Unique Identifier Definition Database* is the mapping of file identifiers taken from datasets description is utilized for matching with the respective definitions of column labels to bring more context to the data. Such Unique Identifier Definitions may be readily available as part of the data source definitions. Additionally, they can be learned over time using *Human-in-the-Loop* from step 3 and reused across datasets. The objective is that description of the identifier can be used instead of using the identifier which may be an encoded value and not useful for human comprehension and data analysis. For example, a column in Order Table has country ID's for multiple orders called

Table 2: Data Mapping Techniques by Data Type along with their corresponding Confidence-Level and Sample Mappings

Data Type	Mapping Technique	Example	Confidence Level
Structured	Label Mapping	'customer_name' <–>Customer_Name	High
	ID-based Mapping	Country id '7' <–>'USA'	High
	Semantic Mapping	'Servers' <–>'linux, windows'	Medium
	Relation mapping	'Region' <–>'country'	Medium
Unstructured	Keyword Clustering	{invoices billed, settled, invoice dispute}	Low
	Theme Clustering	Orders <–>autorenewed, terminated	Low

Country_ID_X. In order to provide meaning to the country ID's in this table, we can obtain the value of the country ID's from the Country Table. The Country table is a dataset of 100 rows corresponding to 100 countries and their ID's. By mapping the Country_ID_X, we can map the Country_ID_X to the name of the country in the Orders table. For example Country id '7' mapping to 'USA'; Country id '9' mapping to 'Canada').

(3) ***Semantic Mapping:*** After *Label* and *ID Mapping* is performed, the remaining unmapped labels are evaluated using semantic similarity between column values. This similarity is useful when matching product components between *Orders* and *Invoices.* Such product components include sub-product items that were requested in the order and are billed in the invoice items. Semantic similarity between sub-product components help map the columns between the datasets like *Orders* and *Invoice* tables. The similar sub-product components are identified using the vector distance between the values in the column using cosine similarity. The angle determines the degree of similarity between the components. The output of this step is the mapping of similar column values that can be mapped to each other. For example, in Figure 3, the product components like Server, Security, etc. are semantically related, which aids in the data mapping of the columns.

(4) ***Relation Mapping:*** Using the aforementioned mapping approaches, four columns were mapped as show in Figure 3 using Label, ID-based and Semantic Mapping. For the remainder of the unmapped columns, the objective is to find the most correlated columns, and analyze whether they share a linear or non-linear relationship with each other. For each of the remainder column pairs the Pearson coefficient correlation [3] was computed and column *Contract* from *Orders* and *Contract ID* from *Invoice* table was determined to be linearly related.

The Pearson coefficient correlation analysis is conducted between column labels and values to find the most correlated columns in the datasets. The columns are considered highly correlated if their correlation coefficient score is above a predefined threshold. The mapping relationship between columns is based on these highly correlated columns.

If the relationship between columns x and y is linear, it is denoted as follows:

$$x- > p_o + p_1 y \quad (1)$$

where the linear regression [10] is defined as x changing wrt y. If the relationship is non-linear, a generalized additive models (GAMs) [15] is used to fit the columns and find the relationship.

The output of this step is a relationship between columns, which is stored in the *Stored Data Mappings* database.

3.2.2 *Unstructured Data.*

Next, we describe the pipeline for data mapping within unstructured data. We utilize the unstructured data to show a relation with structured data columns. This mapping is performed for the keywords extracted from unstructured data in the previous step, described in 3.1

(1) ***Keyword Clustering:***
We use the keywords (see example comments and extracted keywords in Table 1) and their definition obtained using a dictionary source like Wordnet [9] as input to the k-means clustering [5] model. Table 4 shows an example of keywords clustered using this approach. For example, the *order* was clustered together with *cancelled*, *terminated*, *autorenewed*, *autorenew*, *out of funds* because they show state variations of *orders*; *invoice* was clustered together with *settled*, *settlement*, bill, *full amount*, *paid* because such keywords indicate different states of invoices billed; the *customer* was clustered with *disputed*, *contact*, *notification*, *cancel* because these keywords indicate interactions with the customer.

(2) ***Theme Clustering:*** The intuition behind this step is to further abstract the clustered keywords into higher-level concepts, there by, adding more context around them.
Table 1 shows examples of Themes like order, invoice, communication, invoices, renewal, product, contract, where {order, cancelled}, {order, terminated}, {order, autorenewed}, {order,renewal} are several clusters at the theme level 'order'. These clusters group keywords with similar semantic meanings together.
Themes are mapped to structured data columns using the semantic mapping to map to relevant column headers in the Trained Mapping engine. This step is important as the unstructured data should be showing some mapping to structured data columns, in order to show their relationship.

3.2.3 *Compute Column Weight.*

A weighting function is used to identify the confidence score of the mapping between columns. This weighting function depends on the confidence score associated with the type of technique used in order to establish the mapping. It also factors in the number of different techniques that recommended the same mapping between columns to boost the confidence score. Weighting function can be defined in equation 2 as

$$W\_c = \sum f(cf, n\_t) \quad (2)$$

where W_c is the weight of the column c, cf is the confidence score and n_t is the number of mapping techniques that recommended the same mapping.

The output of this step is to provide weights to columns mapped, and stored in the database.

3.2.4 *Conflict Resolution.*

The objective of the data mapping engine is to achieve one-on-one mapping between data columns. In some cases, it is likely that the data mapping engine provides multiple mappings between columns, due to mapping techniques that may intersect with each other. Therefore, the weights from the previous step are utilized to identify the most relevant mapping between columns. The column mapping with the maximum confidence score is selected as the final mapping.

3.3 Human-in-the-Loop

Finally, in this step human-in-the-loop, Subject Matter Expert (SME) who has domain-specific knowledge and expertise around the datasets, helps analyze the identified mappings between the datasets. For example, in the financial services industry, such an SME may be a member of financial teams that maintain contracts and invoices data. SME can analyze the multiple mappings and validate if the mappings are incorrect due to reasons not factored in by the data engine. Examples of discrepancies can be erroneous spelling mistakes leading to incorrect mappings, mixing up definitions of the same terms like 'terminated' and 'closed' etc. The SME can also iterate on the mapping rules and point out if discrepancies in data are leading to inaccurate mapping.

The final output of the system is the mapping between columns in structured and unstructured datasets, which is stored in the *Store Data Mappings* database. An exemplar flow of this method is shown in Figure 3.

4 INDUSTRIAL APPLICATION

In this section, the industrial implementation and validation of DaME in a production environment are described. We start with explaining our real-world application of DaME based on the financial services industry in Sub-section 4.1, then proceed to evaluation and results in the following sections.

4.1 Dispute Elimination Through Analytics (DELTA) application

We built the Dispute Elimination Through Analytics (DELTA) tool for our financial services team. A user interface snapshot is shown in Figure 2. DELTA has three major functions 1) *Monitor:* it monitors real data from large data warehouses comprising contracts, orders, invoices, and customer data, 2) *Integrate:* uses DaME to integrate data across previously mentioned datasets, and 3) *Analyze:* provides risk analysis on the integrated dataset created using the data mappings from the previous step. For purpose of this paper, we focus on the evaluation of the DaME (i.e. Integrate), Monitor and Analyze are out-of-scope, hence, no further elaborations are provided.

We next present the training of DaME in Section 4.2 and share the results of the empirical evaluation to show the efficacy of our approach along with *Human-in-the-Loop* feedback in Section 4.3.

4.2 Evaluation setup and training data mapping engine

4.2.1 Dataset. We obtained various datasets spanning multiple years from 2019 to 2022 comprising of purchase *Orders* (8462 orders), *Invoices* billed and settled dataset (1.2 million invoices), *Customer* profile (3457 customers), *Countries* (238). There are multiple invoices for orders as billing is conducted at weekly, monthly, quarterly or annual frequencies. The *Invoices* dataset was the largest consisting of 1.2 million records of invoices billed for the purchase *Orders.* For evaluation of **DaME**, we selected *Orders* dataset which contains 32 columns and used it as the seed dataset, with the goal that the columns *Orders* should map to columns of other datasets like *Customer*, *Invoices*, *Country* etc. From the *Orders* dataset multiple smaller datasets based on country (USA, UK, and AU) namely, DSet1, DSet2, and DSet3 were created, as shown in Table 3. In the actual implementation, the datasets typically contain rows of data with columns from multiple tables. For example, as shown in Figure 3 if there are 2 tables - *Orders* table with 5 data sample rows and *Invoice* table with 4 data sample rows, we get a training set of 9 data sample rows to be mapped to each other.

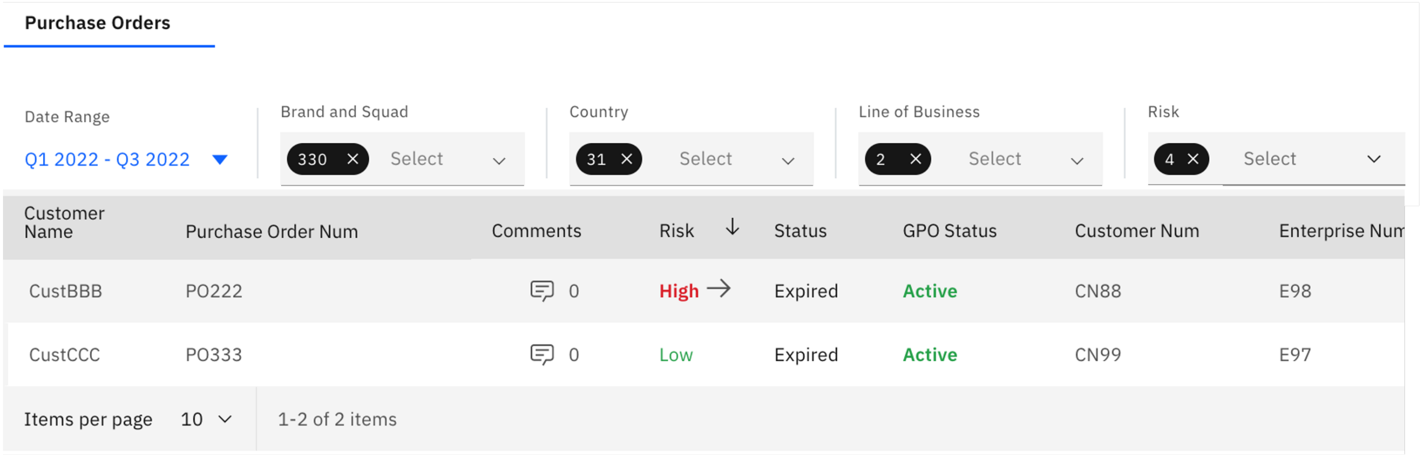
In an ideal scenario, the higher volume of the dataset would lead to a higher accuracy of the mapping results. Therefore, for our evaluation in the next sections, we utilized the USA dataset (with 3000 training and 514 test samples), given that is the largest dataset among the three datasets available. However as seen in Table 3, we observed that the UK dataset had higher accuracy of 95% in data mapping in comparison to USA (69%) and AU (45%). This is because one of the business units in the USA had bigger *Data Gaps* and *Unavailable Data*, which affected its accuracy. Let us elaborate on the intermediate results of the implementation in the following sections.

4.2.2 Structured Data Evaluation. For each order, we extracted the structured and unstructured meta data for our model features. We map the column labels between datasets using *Label Mapping*. Using the Unique Identifier Database, *ID Mapping* technique is applied between datasets to map the ID based columns to columns with definitions or more descriptive values. Table 1 contains some examples of mappings and the technique used, for example, "Country ID X" maps to "Country Name", "Customer ID" maps to "Customer Name", and "Market ID" maps to "Industry".

Using *Label* and *ID Mapping*, 10% of columns from the DSet-1 testing datasets were accurately mapped. Another, 12% of the columns were mapped using *Semantic Mapping* technique. Finally, the remaining 78% of the columns were analyzed using the *Relation Mapping* technique to find correlations between columns. As shown in DSet-1 (Table 4), these remaining 78% columns showed non-linear relationships between the data columns indirectly. For example, "PO start date" is indirectly related to "invoice start date" wherein "invoice start date" is a date that always occurs after "PO start

Table 3: Accuracy comparison of Data Mapping using the proposed system DaME with existing Label Mapping technique from Prior Art

Dataset ID	Type of dataset	Mapping techniques	Accuracy	#of training samples	# of test samples
DSet-1	USA orders	DaME	0.69	3000	514
DSet-1	USA orders	Label mapping	0.34	3000	514
DSet-2	UK orders	DaME	0.95	1000	233
DSet-2	UK orders	Label mapping	0.6	1000	233
DSet-3	AU orders	DaME	0.45	700	157
DSet-3	AU orders	Label mapping	0.3	700	157



Purchase Orders

Date Range: Q1 2022 - Q3 2022 | Brand and Squad: 330 ✕ Select | Country: 31 ✕ Select | Line of Business: 2 ✕ Select | Risk: 4 ✕ Select

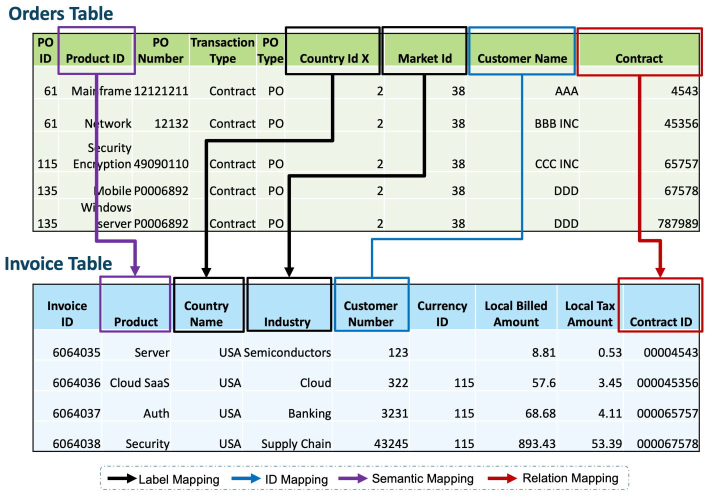
Customer Name	Purchase Order Num	Comments	Risk ↓	Status	GPO Status	Customer Num	Enterprise Num
CustBBB	PO222	0	High →	Expired	Active	CN88	E98
CustCCC	PO333	0	Low	Expired	Active	CN99	E97

Items per page 10 | 1-2 of 2 items

Figure 2: User Interface of Dispute Elimination Through Analytics - DELTA tool

Table 4: Examples of Keyword and Theme Clustering

Keywords	Theme
{order, cancelled},{invoice, settled}	{order, invoice}
{order, cancelled}	{order, invoice}
{order, terminated},{ticket, representative}	{order, communication}
{invoices, settlement},{notification, customer}	{invoices, communication}
{order, autorenew}	{order, renewal}
{customer, disputed},{cloud services product}	{communication, product}
{contract, cancel},{invoices, bill},{server}	{contract, invoices, product}
{customer, contact},{order, out of funds}	{communication, order}
{order, autorenewed},{invoices, full amount, paid}	{order, invoices, renewal}



Orders Table

PO ID	Product ID	PO Number	Transaction Type	PO Type	Country Id X	Market Id	Customer Name	Contract
61	Mainframe	12121211	Contract	PO	2	38	AAA	4543
61	Network	12132	Contract	PO	2	38	BBB INC	45356
115	Security Encryption	49090110	Contract	PO	2	38	CCC INC	65757
135	Mobile	P0006892	Contract	PO	2	38	DDD	67578
135	Windows server	P0006892	Contract	PO	2	38	DDD	787989

Invoice Table

Invoice ID	Product	Country Name	Industry	Customer Number	Currency ID	Local Billed Amount	Local Tax Amount	Contract ID
6064035	Server	USA	Semiconductors	123		8.81	0.53	00004543
6064036	Cloud SaaS	USA	Cloud	322	115	57.6	3.45	000045356
6064037	Auth	USA	Banking	3231	115	68.68	4.11	000065757
6064038	Security	USA	Supply Chain	43245	115	893.43	53.39	000067578

Figure 3: Example of Structured mappings

date". This relation mapping showed insight in the mapping, thus providing risk analysis in the aforementioned DELTA tool.

4.2.3 Unstructured Data. Roughly 120 comments of customer support interactions and a summary of conversations were used to extract keywords like order, canceled, settled invoice, notification, etc. Using *Keyword Clustering* algorithm [5], keywords were clustered into smaller sets. Next, *Theme Clustering* was performed, where keywords are clustered together into themes. Table 1 shows the examples of comments, extracted keywords, and theme clustering performed. The results were analyzed qualitatively by the same set of SMEs from three departments in the financial services industry in order to validate the accuracy of the keyword to the theme cluster.

4.2.4 Testing the Mapping. The mappings between columns across multiple datasets were identified and stored in the database, using the DaME system, as shown in Figure 1. Using the stored data mappings and a test set, we computed its f2 scores and confidence scores. The accuracy of the results for different datasets organized by country is shown in Table 3. We also came across cases where there were multiple mappings between data columns. For conflict resolution, we used a weighting function based on the confidence scores of the mapping. Data Mapping with maximum confidence score was selected.

4.3 Human-in-the-Loop

For the qualitative evaluation of the proposed DaME system, we randomly selected 200 examples of structured and unstructured data and the corresponding mappings predicted by DaME from the above-mentioned dataset of mapping 8462 orders. We asked 8 SMEs, who are experts in the Quote-to-Cash process (covering sales, account management, order fulfillment, billing, and accounts receivable) in financial services, to manually inspect and validate the quality of mappings (produced by DaME).

Table 5 provides a high-level overview of the categories of the data mapping feedback provided by SMEs. The qualitative analysis test set included responses to the above 200 examples from the

Table 5: Distribution of Qualitative Analysis by SME Feedback Type

SME Feedback Type	Description	% of Feedback Type
Correct Mapping	Data mapping provided by the system is correct	60%
Incorrect Mapping	Data mapping provided by the system is incorrect	15%
Unavailable Mapping	Out of scope mapping	10%
Ambiguous Mapping	Multiple mapping between column	7%
Data Gap	Missing data information for analysis	5%
Data Outdated	Data columns are not updated	3%

SMEs. Approximately 60% responses showed that our mapping was done correctly. We further investigated 40% responses which were not entirely correct mapping, based on the response from the SMEs. The table gives a breakdown of feedback received on which we took action.

Table 6 shows sample subset of *"Data Mapping"* generated by DaME for both *structured* and *unstructured* data, which were then evaluated by the SMEs and their corresponding feedback is shown in column *"SME Feedback Type"*.

We mapped the feedback provided by SMEs into six categories. Below we share our observations from SME Feedback, recommendations, and, remediation actions taken to improve the DaME system.

(1) *"Correct Mapping"* indicating that the columns mapped to each other across datasets are correct. This correct mapping was due to completeness in data columns, and labels without any corrupt or erroneous values. As per SME evaluation, shown in Table 5 60% of the data mappings are correctly mapped.

(2) *"Incorrect Mapping"* that is mapping identified by the system not correct, which happened 15% of the times as per the SME evaluation. This leads to our first observation, which is, **Incorrect Mapping typically happens in case of unstructured data**. This indicates the challenging nature of dealing with free-form text to extract the "right" keywords and then perform clustering. Since such techniques are statistical in nature, there is a need to build a bigger dictionary of keywords cluster that can be used for training data, to provide better results. Hence, we recommend using a larger dataset for evaluation to get a true picture of the underlying algorithm. Currently, the keywords training dataset is small, and out-of-the-box language models are generic in nature, and may not provide accurate clustering results.

(3) *"Unavailable Mapping"* that is mapping unavailable due to incomplete or corrupted, or null data values in the table columns, incorrect or null column labels. As per Table 5, 10% of the feedback received from SMEs is due to unavailable mappings. Since such a mapping issue is related to data, there is no remediation action planned for DaME for further improvements.

(4) *"Ambiguous Mapping"* that is multiple mappings possible between column labels, which happened in 10% cases as per SME evaluation (see Table 5). Ambiguous mappings may be considered out of scope if one or more of the similar column labels are incomprehensible. For example, column label defined as *'P'1'* or *'P num'* is considered out of scope, as one cannot infer the purpose or meaning of the column label.

In unstructured data, ambiguous mappings happened due to overfitted trained clusters. For example, as seen in Table 6, 'Invoices' and 'disputes' are clustered together due to insufficient clusters semantically linked to 'disputes'. Hence, multiple columns can map to the same cluster, even if they are not semantically similar. This leads to our next observation that **Special Attention should be paid to Overfitted Clusters** to avoid ambiguous data mapping.

(5) *"Data Gap"* feedback indicates that there is missing data information required for accurate mapping. For example, in the case of structured data columns, the mapping between "PO ID" and "Invoices ID" may only occur till 2021. The data gap lies in the date constraint in the mapping. This can be resolved by fixing any constraint that restricts the mapping. Our observation is that constraints may be used for improving accuracy but **Use of Constraints should be Targeted** for special scenarios which cannot be accomplished via other data mapping techniques.

(6) *"Data Outdated"* indicating that data columns are not updated if they have not been maintained over time. This can occur either if the datasets are too vast to be monitored, or the data collection system is automated and the front end, may be out of sync with the data schema of the backend. For example, in a distributed system, if data schema or code collecting data is updated in a few components, and is not updated in some other parts of the system, there is an issue with data collected being outdated. Our observation is that it is **Necessary to Monitor Updates that can Impact Data Collection** in distributed systems to ensure data integrity.

The underlying cause of SME feedback related to *"Unavailable Mapping"*, and *"Data Outdated"* is due to an issue with data, which is currently beyond the scope of mapping techniques implemented in DaME. Given that such errors collectively accounted for 13%, we removed such mapping examples from the evaluation. We recomputed the accuracy of DaME at 69%, 17%, 8% and 5% using 4 categories namely *"Correct Mapping", "Incorrect Mapping", "Ambiguous" Mapping"*, and *"Data Gap"* respectively.

Table 6 provides some of the incorrect mapping comments we received from the SMEs from different business units. Certain actions were taken to fix the mapping like incorrect labels, outdated columns used in mapping, etc. The actions taken for both structured and unstructured mapping, showed discrepancies in labeled data, incorrect semantic understanding of keywords, incorrect contextual understanding, etc.

Summarising the results from the quantitative and qualitative evaluation feedback, the overall data mapping engine had a mean accuracy of 69%. Only 13% of mappings were out of scope for the current dataset and were not included. By identifying data mapping accurately, DaME has also helped improve the productivity of SMEs. Over a period of 10 months, DaME was actively used by 8 SMEs. Without DaME roughly 14,000 hours may have been spent manually mapping data on a weekly basis, which were reduced to a few hours per week. Those few hours were spent on validating and providing feedback on the output of DaME. This feedback was used to take

Table 6: Examples of Qualitative analyses and action taken

Data Type	Data Mapping	SME Feedback Type	Type of Remediation Action Taken
Structured data	"PO number" mapped to "PO ID"	Correct Mapping	No Action Needed
	"Order number" unmapped	Unavailable Mapping	Out of Scope
	"Order" mapped to "Invoices billed"	Data Gap	Added Semantic Mapping
	"Servers" mapped to "linux server"	Correct Mapping	No Action needed
	"Order" mapped to "invoices" from different geography	Data Gap	Added Semantic Mapping
	"Invoice" mapped to "PO" till 2021 only	Data Gap	Removed Rule-based Constraint
	"Invoice settlement date" unmapped	Unavailable Mapping	Out of Scope
	"PO number" and "PO ID" matching to "transaction number"	Ambiguous Mapping	Define data columns correctly
	"Customer" has invalid data	Data Outdated	Out of Scope
	"Country ID" mapping to "country names"	Data Gap	Updated ID-based mapping
	"Item services" mapped to "Cloud product catalog"	Data Gap	Updated Product catalog
Unstructured data	"Customer ID" mapped with "invoices billed"	Incorrect Mapping	Added Training Samples
	"Invoices" keyword clustered with "disputes" keyword	Ambiguous Mapping	Added Cluster for Disputes
	"Country name" mapped with "contract region"	Correct Mapping	No Action needed
	"Invoices date" keyword mapped to "Customers theme" cluster	Incorrect Mapping	Added Training Samples
	"hybrid cloud" mapped to "product catalog"	Incorrect Mapping	Update Product catalog

remedial actions to improve the performance of DaME. Thus further reducing the human effort to an hour of analysis per week required for mapping validation.

5 CONCLUSION AND FUTURE WORK

In this paper, we presented a novel approach for automatic identification of data mapping, utilizing a trained data mapping engine comprising of an ensemble of techniques along with human-in-the-loop. In this approach, we first extracted data and processed it, then trained the data mapping engine. We resolved the multiple mappings between columns and utilized human-in-the-loop to iterate on the results. We also showed results based on a real-world application that illustrates the accuracy and efficiency of the models, as well as the efficiency of the overall method.

We have demonstrated the efficacy of the proposed system both via quantitative and qualitative evaluation. Additionally, we have shared our observations derived from SME's feedback.

Finally, we put forward that there are multiple directions for future research in our work. First, we can leverage machine learning algorithms to automate the mapping to further improve the accuracy of the model. By doing so, we would be able to address the *Data Outdated* and *Ambiguous Mapping* issues. Second, we can expand the scope of the datasets for more geography and business types, to demonstrate whether the proposed approach is applicable to broader environments.

REFERENCES

[1] Bazeer Ahamed, Razwan M.S. Najimaldeen, and Yuvaraj Duraisamy. 2020. Enhancement Framework of Semantic Query Expansion Using Mapped Ontology. In *2020 International Conference on Computer Science and Software Engineering (CSASE)*. 56–60. https://doi.org/10.1109/CSASE48920.2020.9142093

[2] Serkan Ayvaz, John Horn, Oktie Hassanzadeh, Qian Zhu, Johann Stan, Nicholas P Tatonetti, Santiago Vilar, Mathias Brochhausen, Matthias Samwald, Majid Rastegar-Mojarad, et al. 2015. Toward a complete dataset of drug–drug interaction information from publicly available sources. *Journal of biomedical informatics* 55 (2015), 206–217.

[3] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*.

[4] Mayanka Chandrashekar, Rohithkumar Nagulapati, and Yugyung Lee. 2018. Ontology Mapping Framework with Feature Extraction and Semantic Embeddings. In *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*. 34–42. https://doi.org/10.1109/ICHI-W.2018.00012

[5] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.

[6] Angelika Kimmig, Alex Memory, Renée J. Miller, and Lise Getoor. 2019. A Collective, Probabilistic Approach to Schema Mapping Using Diverse Noisy Evidence. *IEEE Transactions on Knowledge and Data Engineering* 31, 8 (2019), 1426–1439. https://doi.org/10.1109/TKDE.2018.2865785

[7] Rajasekar Krishnamurthy, Yunyao Li, Sriram Raghavan, Frederick Reiss, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2009. SystemT: a system for declarative information extraction. *ACM SIGMOD Record* 37, 4 (2009), 7–13.

[8] Sergio Luján-Mora, Panos Vassiliadis, and Juan Trujillo. 2004. Data mapping diagrams for data warehouse design with UML. In *International Conference on Conceptual Modeling*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 191–204.

[9] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[10] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. 2021. *Introduction to linear regression analysis*. John Wiley & Sons.

[11] Artem A Nazarenko, Joao Sarraipa, Luis M Camarinha-Matos, Oscar Garcia, and Ricardo Jardim-Goncalves. 2019. Semantic data management for a virtual factory collaborative environment. *Applied Sciences* 9, 22 (2019), 4936.

[12] Robinette Renner and Guoqian Jiang. 2020. Challenges in Using a Graph Database to Represent and Analyze Mappings of Cancer Study Data Standards. *AMIA Summits on Translational Science Proceedings* 2020 (2020), 517.

[13] Tanvi Sahay, Ankita Mehta, and Shruti Jadon. 2020. Schema Matching using Machine Learning. In *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*. 359–366. https://doi.org/10.1109/SPIN48934.2020.9071272

[14] Kavitha Srinivas, Abraham Gale, and Julian Dolby. 2018. Merging datasets through deep learning. *arXiv preprint arXiv:1809.01604* (2018).

[15] Simon N Wood. 2006. *Generalized additive models: an introduction with R*. chapman and hall/CRC.

[16] Takaaki Yamada, Yuko Kato, Yuki Maekawa, and Tomoe Tomiyama. 2017. Interactive Service for Visualizing Data Association Using a Self-Organizing Structure

of Schemas. In *2017 IEEE 10th Conference on Service-Oriented Computing and Applications (SOCA)*. 230–233. https://doi.org/10.1109/SOCA.2017.39

[17] Hansi Zhang, Yi Guo, Qian Li, Thomas J George, Elizabeth Shenkman, François Modave, and Jiang Bian. 2018. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC medical informatics and decision making* 18, 2 (2018), 129–147.