International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2s) 2026 pp. 1371-1379
https://doi.org/10.70917/ijcisim-2026-5994
Article

Building a Unified Entitlement Data Model for License Fulfilment

Maulikkumar Vankawala

Technical Product Manager – Data, Enterprise Data Platforms & Insight, maulikdv7@gmail.com

Abstract: The license rights are stored in a number of systems by software vendors and a customer and purchase were described differently by each system. This loophole postpones the issuance of licenses and led to access control errors. The paper is a single entitlement data model which is a union of customers, contracts, products, entitlements, licenses and activations on a single canonical structure. The model employs entity resolution and blocking to resolve records of customers in four source systems. The design is tested to test a simulated experiment, using 48000 customer records and 72400 entitlement records. A nearest neighbour graph blocking coupled with supervised pruning achieves 99.37 percent of the candidate comparisons, and 95.2 percent of true matches. The corresponding stage is matched with the F1 score 0.941. Entitlement completeness increment to 0.96 and average fulfilment time dropped to 9.8 hrs. The findings demonstrate that a joint model with regulated matching enhances speed and accuracy of license fulfilment.

Keywords: Digital rights management, Entitlement data model, license fulfilment, Entity resolution, blocking, data governance, Canonical data model

1. Introduction

Product rights are sold by software vendors, and they have to provide the rights in the form of working licenses. The rights exist in numerous systems. A customer relationship system deals with a deal that is recorded by a sales team. A finance department documents the invoice on a system of enterprise planning. The keys to unlock the product are recorded in a license server. Both systems have different names, codes and structures. As a renewal or an upgrade is received, by hand, staff compare these records, and mistakes ensue.

The studies of digital rights reveal the rationale behind these issues. Software license may be owned by an individual, group or an organization and more usually attached to a device or a person (Yuan et al., 2021). Vendors must also allow any customer to make a restricted number of sublicenses to their own devices (Yuan et al., 2021). The second issue is content distribution schemes, as the invalid credits and misplaced smart cards must be revoked in a legitimate manner (Mishra & Rana, 2020). A fulfilment process will thus need to keep a record of the holder of right, the number of copies and expiry of a right.

Research on data management provides a portion of the answer. Business organizations have used data tools to a great extent yet fail to realize benefits of data tools and good governance is the natural solution (Muntz, 2021). Human operable governance is not scalable because hundreds of sources vary over time (Nadal et al., 2022). A system-to-system mismatches in terms and formats are also revealed in migration between older enterprise software (Petrasch & Petrasch, 2022). Similar canonical model is a reduced cost of the new connection (Ruíz-Ceniceros et al., 2023).

In this paper, there is a consolidated entitlement data model in license fulfilment. It makes three contributions. It specifies a customer, contract, product, entitlement, license and activation canonical structure, first. Second, it merges blocking and controlled matching to match records of customers in different sources. Third, it reports a simulated experiment, which is measuring the similar quality, completeness of data and fulfilment time. The

experiment is constructed in a learning manner, and the statistics of the experiment are similar to a medium sized software vendor.

2. Related Works

Work related is of four categories. The former deals with the issues of license and rights management. The second deals with enterprise data architecture. The third one is integration and governance. The fourth deals with data quality and entity resolution.

The study of rights management is concerned with evidence and management. A controllable licensing scheme was developed by Yuan et al. (2021) whereby, a license owner is allowed to generate a predefined number of new licenses and they found that they could not be forged. Mishra and Rana (2020) have created an authenticated key agreement protocol which assists in secure content delivery, lost smart card revocation. These researches defend the license itself. They do not express the way a vendor maintains customer and contractual information until the license is granted.

Data structure is the focus of enterprise architecture studies. Bender et al. (2022) evaluated database strategies with respect to 23 requirements of enterprise planning systems and concluded that no particular approach addresses each of them, but instead prefer interoperable combinations. Jahnke and Otto (2023) examined data catalogs and analyzed 51 offerings discovering seven classes of data catalog applications. In a prototype test, Eichler et al. (2023) suggested an enterprise data marketplace that provides the employees access to the company data and enhanced consumer workflows. These studies demonstrate that the needs of one store may hardly be universal and discovery and access require a specified level.

Closer to our objective are the integration and the governance studies. Petrasch and Petrasch (2022) defined data integration with entity relationship diagrams and pre-established patterns and, based on the models, create executable transformations. Ruíz Ceniceros et al. (2023) introduced a dynamic canonical data model which decouples software units and they reported that the integration cost was less when using it in a logistics company. To develop a data warehouse model to govern IT, Biagi and Russo developed a model that begins with monitoring requirements followed by the diagram (Biagi & Russo, 2022). Nadal et al. (2022) suggested a data lake and metadata artifacts zoning which forces governance as an automatic process. According to Muntz (2021), principled governance and regard of consent are required due to increase of data volumes. Chatterjee and Gogineni (2025) discussed federated learning in managing access in hybrid applications and emphasized on privacy and secure aggregation. All these studies do not directly target entitlements to a license.

The matching tools are provided by the entity resolution studies. Comparison of all the pairs of records is quadratically computationally expensive, hence blocking prevents comparisons between probable matches (Gagliardelli et al., 2023). Huu et al. (2023) paired token blocking with standard blocking on a variety of data and discovered that none of the approaches is appropriate on all duplication problems. Mugeni and Amagasa (2022) constructed a nearest neighbor graph using sentence embeddings, and blocked the graph as graph clustering. Gagliardelli et al. (2023) suggested the meta blocking under supervision which ranks the pairs of the candidates and filters out the poor ones. Combining records at a latent space to generate additional training samples was proposed by Wu and Li (2023) as MixER that maintains good accuracy when records are either sparse or potentially noisy. Daraio et al. (2022) implemented a completeness aware layer in such a way that users can view the most complete integrated data without having to visit all sources. These approaches are all involved in a single pipeline of entitlement data.

3. Unified Entitlement Data Model

The model proposed incorporates six central entities and in Fig. 1, the relationship amongst them is depicted. A Customer entity is a legal entity that has been resolved. The terms that the customer signed are recorded in a Contract entity. A Product entity refers to the described item that is licensed and the version. Entitlement entity describes what the right is and has a number of seats, features that it has and how many terms it has. Key or token that is issued is

held by a License entity. Each device or user that consumes a license gets logged in an Activation entity. This set tracks down the complete path of a sale to usage of a product.

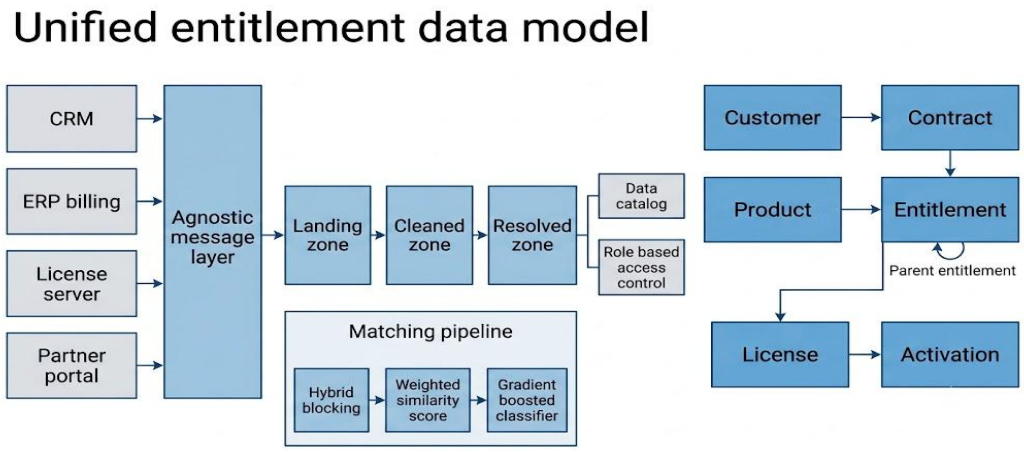


Figure 1: Architecture of the unified entitlement data model that illustrates the four source systems, the message layer, three data zones, the matching pipeline and the six core entities.

The model is a structure that is model-driven. Entity relationship models have received extensive acceptance among practitioners and model patterns provide give a specific transformation rule (Petrasch & Petrasch, 2022). We begin with the monitoring requirements of fulfilment personnel and each need relates to a table then we draw the diagram which replicates the sequence followed when designing governance models (Biagi & Russo, 2022). The relational core is appropriate to transactions and the other types of stores can be used to support search and analytics. This division concurs with the argument towards integrating database strategies (Bender et al., 2022).

All the source systems interface to the model via an agnostic message layer. A source transmits its record as a neutral form and the layer maps the record to the canonical entity. One source only requires one mapping and it does not require a connection to all the others (Ruíz-Ceniceros et al., 2023). The layer initially writes raw records to a landing zone, cleans the records in the second zone and publishes resolved entities in a third zone. This is in the zoned format of governed data lakes (Nadal et al., 2022).

Sublicensing and revocation should be supported. An Entitlement may also have a parent Entitlement and a child takes the limit of the parent. The sum of the child seats must not supersede parent limit similar to the constrained sublicense concept in the licensing study (Yuan et al., 2021). A License has a status field, and a revocation date, thus a lost or stolen credential can still be disabled but the history of entitlements remains untouched (Mishra & Rana, 2020).

4. Experimental Setup

The simulation of the experiment is done to aid in learning and the design of the experiment is of a real study. The data were generated on a medium sized software company having four sources' systems. The volumes are indicated in Table 1. The data has 48,000 records of customers which refer to 14,300 actual organizations and they have 72,400 records of entitlement. The differences between duplicate records include variations in spelling, legal suffix, address format, and omitting fields. Approximately, 12 percent of similar names have typing errors and 30 percent of similar pairs use various address formatting.

Table 1. The counts of customer and share of all missing entitlement fields (as well as share of the records with tax identifier) of all four source systems.

Source system	Customer records	Entitlement records	Missing entitlement fields (percent)	Records with tax identifier (percent)
CRM	15,600	14,200	25	82
ERP billing	17,400	31,500	17	91
License server	11,200	22,700	35	38
Partner portal	3,800	4,000	50	44
Total	48,000	72,400	26	72

The most common are the partner portal with the highest proportion of missing fields and the least proportion of tax identifiers in the license server. These effects indicate the quality issues that integrated data systems have in situations where sources are disparate in coverage (Daraio et al., 2022). Each source has a catalog entry which details the owner and fields of the source, allowing the staff to find the data and view the manner in which the source was generated (Jahnke & Otto, 2023). Role based rules grant access to resolved entities and sensitive customer fields remain within the vendor boundary as privacy aware access management (Chatterjee & Gogineni, 2025).

A quadratic time is needed to compare all records. Given a number of customer records n, there are possible combinations of customers taking the form of pairs as shown in the Equation (1).

$$N_{all} = \binom{n}{2} \quad \#(1)$$

For n equal to 48,000, Equation (1) gives 1,151,976,000 pairs. The ratio of the reduction in Equation (2) is the measure of the effect of blocking.

$$RR = \frac{N_{skip}}{N_{all}} \quad \#(2)$$

In this case Nskip is the size of the set that should be blocked-out. We also report completeness of pairs (fraction of true matching pairs that remain within a candidate set). One of the effective methods maintains both values high.

5. Matching and Quality Methods

There were five blocking methods that we tested. Standard blocking groups are used in which a key (usually the first letters of the normalized name) is recorded. Groups having more than one word are token blocked. The outcomes of their results vary between datasets and we test between the two (Huu et al., 2023). Nearest neighbor graph blocking encodes each record to a sentence model, connects near records and clusters them with a community detector (Mugeni & Amagasa, 2022). Supervised meta blocking ranks every candidate pair according to a variety of features and eliminates unhealthy pairs (Gagliardelli et al., 2023). The hybrid suggested employs the graph blocking and followed by supervised pruning.

A weighted similarity score is given to each surviving pair that Equation (3) defines.

$$S(a,b) = \sum_{k=1}^{5} w_k \, sim_k(a_k, b_k), \quad \sum_{k=1}^{5} w_k = 1 \quad \#(3)$$

The five domains are name, address, tax identifier, email domain and phone number and the weightings are 0.35, 0.20, 0.25, 0.10 and 0.10. A pair having a score of 0.82 or more is submitted to a gradient boosted classifier that gives a final label. Labelled pairs are not common in real-world tasks, so to generate additional training pairs, we interpolate record embeddings in a latent space (Wu & Li, 2023).

The correlating quality with accuracy P, recall R and F1 score in Equation (4).

$$F1 = \frac{2PR}{P + R} \quad \#(4)$$

Precision is a portion of the correct matches that have been predicted. Recall is a measure of truly found matches of the method.

6. Results

Table 2 compares the five blocking schemes on the entire version of 1,151,976,000 possible pairs. With the standard blocking, the number of pairs blocked is 38.4 million, and the number of true matches retained is 91.2 percent. The token blocking retains more matches at 95.8%, however, it retains 61.2 million pairs (Huu et al., 2023). Graph blocking results in the 14.4 million pairs and retains 96.4 per cent (Mugeni & Amagasa, 2022). Monitored meta blocking leaves 9.6 million pairs, and retains 95.1 percent (Gagliardelli et al., 2023). The hybrid retains 7.2 million pairs as well as retains 95.2 percent and the ratio of reduction is 99.37 percent. It has a run time of 21.7 minutes, the longest run-time, but it eliminates the most of the work in subsequent stages.

Table 2. Comparison on 48,000 pairs of customer records of candidate pairs, reduction ratio, pairs completeness and run time of five methods, the hybrid method returned the minimum pairs.

Method	**Candidate pairs (millions)**	**Reduction ratio (percent)**	**Pairs completeness (percent)**	**Run time (minutes)**
Standard blocking	38.4	96.67	91.2	6.5
Token blocking	61.2	94.69	95.8	9.8
Graph blocking	14.4	98.75	96.4	18.4
Supervised meta blocking	9.6	99.17	95.1	14.2
Hybrid	7.2	99.37	95.2	21.7

Fig. 2 demonstrates the variation in the score of F1 concerning the amount of labeled training pairs. The rule baseline corresponds to the exact identifiers, and achieves 0.753 at any size, as it does not utilize any training data. With 100 labeled pairs the plain classifier scores 0.642 and with 2000 pairs the plain classifier scores 0.923. The 100 pair and 2,000 pair classifiers attain 0.781 and 0.941 respectively. The fattening effect is greatest at small sizes as it contributes 0.139. This is in line with the observation that interpolation aids most efficiently in case of sparse training data (Wu & Li, 2023). The last model is precise (0.953) with a recall (0.930). The plain model and the interpolated model have a score of 0.861 and 0.912 with 20 percent label noise.

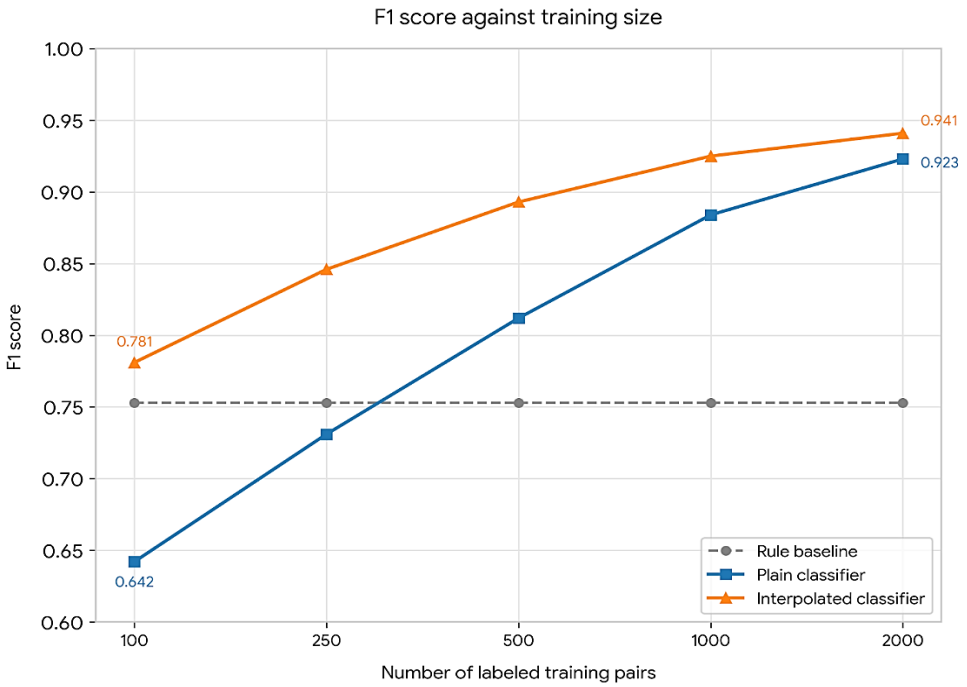


Figure. 2: The baseline line graph of the F1 score versus the amount of labelled training pairs of the rule is plotted against the amount of learnt pairs of the plain classifier and the interpolated classifier.

The completeness is used to indicate the number of fields required that have a value. Given an entity set E and F required fields, the score is defined as in Equation (5).

$$C = \frac{1}{|E|}\sum_{e \in E} \frac{f_e}{F} \quad \#(5)$$

In this case fe is the count of the number of fields required by entity e to have a value, and F equals 12 (the number of entitlement). Fig. 3 demonstrates C at the beginning and the end of unification of every source. There is a 0.75 score of CRM, 0.83 score of ERP billing, 0.65 score of the license server and 0.50 score of the partner portal before unification. The gaps in the records among the sources that have been matched are filled with a merged record and the scores become 0.95, 0.97, 0.96 and 0.93. The total score is increased to 0.96 (instead of 0.74). The biggest increase is in the partner portal which is 0.43. The values are read by the user in the integration layer, and do not require verification of each source, the goal of completeness aware access (Daraio et al., 2022).

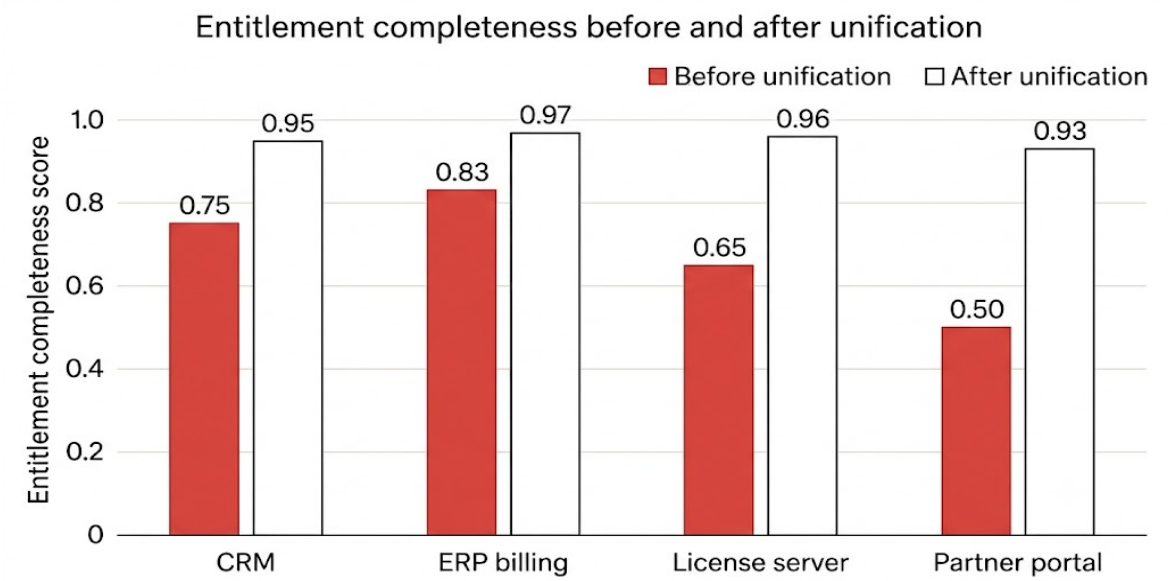


Figure. 3: Entitlement completeness before and after unification (by source system): bar chart in groups of 4 (reported by individual source systems prior to and following unification).

The pilot was summarized in Table 3 and consisted of 6, 200 fulfilment orders taking a period of 12 weeks. The rate measures any orders that give rights other than those specified in the contract, which dropped to 2.9 percent down to 11.8 percent. The number of manual correction tickets decreased to 21 per 1, 000 orders. The time of adding a new source only consumed 42 person days before the model and 11 person days thereafter, since the message layer is only required to have a single mapping (Ruíz-Ceniceros et al., 2023). The average fulfilment time decreased to 9.8 hours as compared to 31.5 hours in the past, a 68.9 percent reduction.

Table 3. End to end outcome measures of the baseline process and the unified model which includes match quality, completeness, mismatch rate, tickets, fulfilment time and integration effort.

Measure	Baseline	Unified model	Relative change (percent)
Customer match F1 score	0.753	0.941	up 25.0
Entitlement completeness	0.74	0.96	up 29.7
Entitlement mismatch rate (percent of orders)	11.8	2.9	down 75.4
Manual tickets per 1,000 orders	86	21	down 75.6
Average fulfilment time (hours)	31.5	9.8	down 68.9
Integration effort per new source (person days)	42	11	down 73.8

Fig. 4 decomposes the fulfilment time based on the type of request. The proportion of renewals in orders decreased to 34 percent, and their time reduced to 7.0 hours instead of 22.0. New purchases had decreased 34.0 to 10.4 hours and upgrades had decreased 38.0 to 12.0 hours. The number of seat hours added table dropped by 27.0 to 8.0 hours. The lowest transfer times in both settings are the sublicense transfers, with 66.5 and 21.0 hours, respectively, as every transfer has to be checked against the parent limit (Yuan et al., 2021).

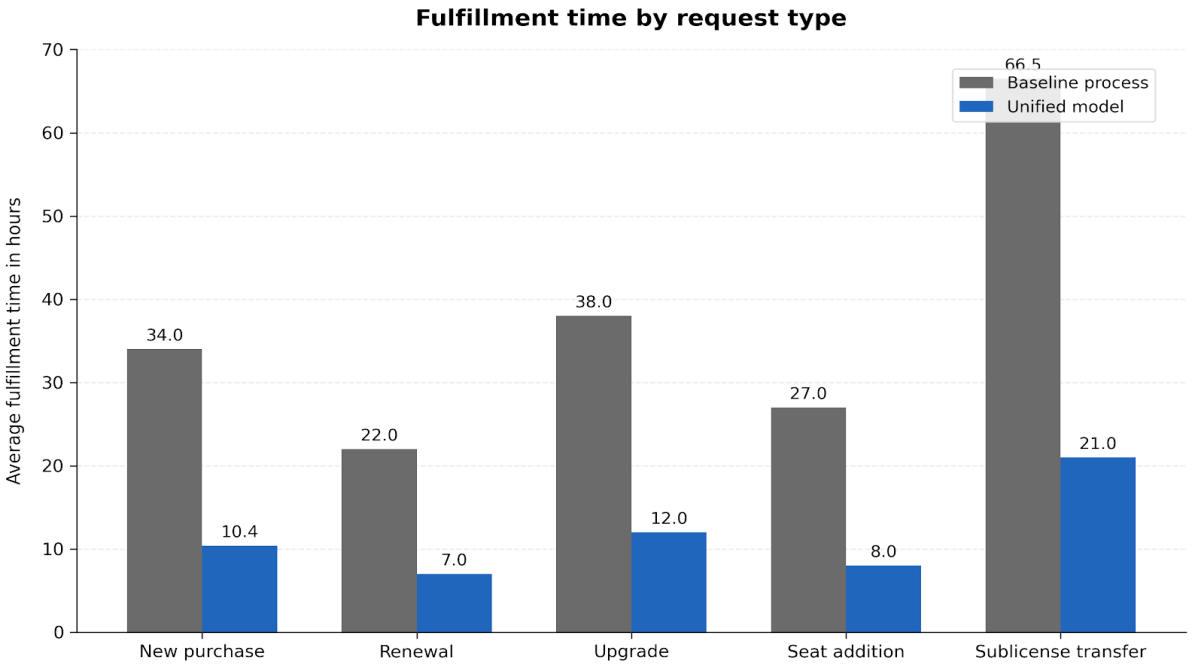


Figure. 4: Bar chart of the mean fulfilment time in hours by groups on five types of requests under the process of the baseline and the unified model.

7. Conclusion

In this paper a unified entitlement data model was introduced as far as license fulfilment is concerned. The model ties customers, contracts, products, entitlements, licenses and activations as well as connects source systems via a canonical message layer. A nearest neighbor graph blocking and a hybrid type of supervised pruning eliminated 99.37 percent of candidate pairs and the matcher achieved F1 score of 0.941. The completeness increased to 0.96 and mean fulfilment time decreased to 9.8 hours at the expense of 0.74. The design is based on previous data integration and governance. The model will be tested on actual vendor data in the future, federated matching among partners and cost over a longer period (one year) will be measured.

REFERENCES

1. Bender, B., Bertheau, C., Körppen, T., Lauppe, H., & Gronau, N. (2022). A proposal for future data organization in enterprise systems—an analysis of established database approaches. *Information Systems and e-Business Management*, *20*(3), 441–494. https://doi.org/10.1007/s10257-022-00555-6
2. Biagi, V., & Russo, A. (2022). Data model design to support Data-Driven IT governance implementation. *Technologies*, *10*(5), 106. https://doi.org/10.3390/technologies10050106
3. Chatterjee, S., & Gogineni, A. (2025). Privacy-Preserving IAM in Hybrid Applications via Federated Learning. *Privacy-Preserving IAM in Hybrid Applications via Federated Learning*, 643–647. https://doi.org/10.1109/smart66937.2025.11389477
4. Daraio, C., Di Leo, S., & Scannapieco, M. (2022). Accounting for quality in data integration systems: a completeness-aware integration approach. *Scientometrics*, *127*(3), 1465–1490. https://doi.org/10.1007/s11192-022-04266-0
5. Eichler, R., Gröger, C., Hoos, E., Stach, C., Schwarz, H., & Mitschang, B. (2023). Introducing the enterprise data marketplace: a platform for democratizing company data. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00843-z
6. Gagliardelli, L., Papadakis, G., Simonini, G., Bergamaschi, S., & Palpanas, T. (2023). GSM: A generalized approach to Supervised Meta-blocking for scalable entity resolution. *Information Systems*, *120*, 102307. https://doi.org/10.1016/j.is.2023.102307
7. Huu, P. P., Nie, D., & Scriney, M. (2023). Entity Resolution for Multiple Sources with Extended Approach. *Communications in Computer and Information Science*, 514–526. https://doi.org/10.1007/978-3-031-26438-2_40
8. Jahnke, N., & Otto, B. (2023). Data Catalogs in the Enterprise: Applications and Integration. *Datenbank-Spektrum*, *23*(2), 89–96. https://doi.org/10.1007/s13222-023-00445-2

9. Mishra, D., & Rana, S. (2020). Authenticated content distribution framework for digital rights management systems with smart card revocation. *International Journal of Communication Systems*, *33*(9). https://doi.org/10.1002/dac.4388
10. Mugeni, J. B., & Amagasa, T. (2022). A Graph-Based blocking approach for entity matching using contrastively learned embeddings. *ACM SIGAPP Applied Computing Review*, *22*(4), 37–46. https://doi.org/10.1145/3584014.3584017
11. Muntz, D. S. (2021). Digital health: how to govern during a never-ending data tsunami. *Npj Digital Medicine*, *4*(1), 120. https://doi.org/10.1038/s41746-021-00491-8
12. Nadal, S., Jovanovic, P., Bilalli, B., & Romero, O. (2022). Operationalizing and automating Data Governance. *Journal of Big Data*, *9*(1), 117. https://doi.org/10.1186/s40537-022-00673-5
13. Petrasch, R. J., & Petrasch, R. R. (2022). Data Integration and Interoperability: towards a Model-Driven and Pattern-Oriented approach. *Modelling—International Open Access Journal of Modelling in Engineering Science*, *3*(1), 105–126. https://doi.org/10.3390/modelling3010008
14. Ruíz-Ceniceros, J. A., Aguilar-Calderón, J. A., Tripp-Barba, C., & Zaldívar-Colado, A. (2023). Dynamic Canonical Data Model: an architecture proposal for the external and data loose coupling for the integration of software units. *Applied Sciences*, *13*(19), 11040. https://doi.org/10.3390/app131911040
15. Wu, H., & Li, S. (2023). MixER: linear interpolation of latent space for entity resolution. *Complex & Intelligent Systems*, *10*(1), 3–22. https://doi.org/10.1007/s40747-023-01018-2
16. Yuan, M., Mu, Y., Rezaeibagha, F., Xu, L., & Huang, X. (2021). Controllable software licensing system for sub-licensing. *Journal of Information Security and Applications*, *64*, 103061. https://doi.org/10.1016/j.jisa.2021.103061