Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Low-Latency Survivorship Scoring over Cloud Data Warehouses: A Workflow-Orchestrated Approach for Enterprise Record Linkage

VENKATESH ALAMURI

venkat.alamuri@gmail.com

Fusion Global Technologies and Solutions Inc https://orcid.org/0009-0000-7460-047X

Research Article

Keywords: Enterprise Record Linkage, Cloud Data Warehouse, Apache Airflow, Google BigQuery, Survivorship Scoring, Workflow Orchestration, Data Governance, Parallel Processing, Adaptive Matching

Posted Date: July 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10387632/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

Low-Latency Survivorship Scoring over Cloud Data Warehouses: A Workflow-Orchestrated Approach for Enterprise Record Linkage

Venkatesh Alamuri [1*]

[1] Fusion Global Technologies and Solutions Inc

venkat.alamuri@gmail.com

Corresponding author: venkat.alamuri@gmail.com

Keywords:
Enterprise Record Linkage;
Cloud Data Warehouse;
Apache Airflow;
Google BigQuery;
Survivorship Scoring;
Workflow Orchestration;
Data Governance;
Parallel Processing;
Adaptive Matching.

Abstract: Enterprise CRM systems use distributed cloud storage to store redundant records and improve the reliability of analytics. Repetitive deduplication can cause costly latency on cloud warehouses, or conversely, poor reliability. The paper notes the significance of using a dataset of 660K+ records and implementing Workflow-Orchestrated Adaptive Survivorship Scoring (WOASS) technique, which involves using Google BigQuery parallel processing and Apache Airflow orchestration, to achieve near-instant resolution to entities. Learned the importance of weights, a survivorship scoring scheme that's based on recency and source authority. Frequency heuristics and a composite attribute similarity function are both utilized in this context. A governance layer that can be monitored from an angle without any breaches in control. The analysis compared 97 CRM records with a database made up of 662,763 CRM records by 20 respondents from BigQuery. So, the correct number is 96. The recall revealed that 8% of the drivers hadn't yet achieved an F1 score. This puts the total count at 970. The total range was between 4% and 18. The increase is mainly due to overrule and ML, resulting in a gap of 8 percentage points between the baselines. The span of 3 seconds (4.). The index of data quality is 94% (with a 7-point speed bump), and the audit coverage is 97%. The statistical range is computed at p 0. Just one thing.

1. Introduction

The use of cloud-based data warehouses like Google BigQuery and Amazon Redshift is prevalent. The use of Snowflake by present-day companies is due to their insistence on preserving customer data in different CRM instances, marketing platforms, and transactional databases. This has added to the already prevalent inconsistency - multi-channel customer visits. Degraded master records, affecting the effectiveness of customer analytics, personalization engines, and regulatory compliance mechanisms. It's said that businesses spend roughly $12 per day on bad data quality, according to sources within the industry. The high number of records, which exceed 9 million a year, can be attributed to doubts. Entity resolution (ER) techniques now feature a mix of exact matching, rule-based systems, machine learning classifiers, and deep learning embeddings [4], [5]. The shift of enterprise data lakes to cloud data warehouses poses additional challenges, including the need for low-latency processing at petabyte scale and the flexibility to operate with serverless compute models. The robustness of integration with managed workflow orchestration services [6,], [7]. The de facto standard for managing complex data pipelines in cloud environments is Apache Airflow, which incorporates DAG execution and task retry mechanisms. Integration with the native cloud services [8]. The current solutions are prone to three big challenges: (1) lack of ad hoc reinforcement for rules-based survivorship scoring that prevent duplicate record conflicts at attribute levels; (2) systems that implement post-processing improvements instead of workflows, such as audited HLN/core-independent processing (HLN) that provide incomplete dataset lineage; and (3) not use SHOP for record linkage; all these are accompanied by unsourced clouds [12], [13], and most systems. The gap is resolved in this paper through five contributions, one of which is a composite similarity function that equals 0. The adaptive adaptive survivorship scoring algorithm, which combines recency and authority, forms the basis for a 970 F1 score. The performance of a multi-criteria fusion workflow algorithm and a BigQuery workflow scheduling algorithm with 13 criteria have been compared. A 56 speedup was achievable when the cluster had 20 employees, resulting in an embedded governance structure that provides 97% audit coverage and records the cluster's history. Executing experimental validation on 662,763 records with sub-10-second latency and p 0. Something fundamental is indicated by 01.

2. Related Work

Christen [4], Papadakis et al. Extensively reviewed the evolution of entity resolution (ER) from rule-based and probabilistic approaches to deep learning systems [14]. By contrast, Ge et al. [15] Blocking and filtering

techniques are motivated by the Fellegi-Sunter model [16], which is a foundational probabilistic approach with computational complexity, while exact matching is highly accurate but sensitive to data variations [17]. Hadoop [18] and Spark [19] are among the distributed platforms used in modern cloud-based enterprise software, while BigQuery-ML pipelines incorporate embedding-based matching [20]. Unlike current approaches, the proposed WOASS system performs similarity computation directly in BigQuery SQL rather than relying on intermediate data movement. Apache Airflow has been used in this context for workflow orchestration, which has resulted in better reliability and scheduling efficiency for ETL purposes [21], [22]. Existing approaches can't adequately cope with ER-specific scheduling challenges. Previous solutions for survivorship and governance [9], [11], [23]-[25] still struggle to cope with diverse characteristics and cloud-based governance, prompting the suggested unified structure.

Table 1. Comparison of Existing Record Linkage Frameworks

Study	Method	Platform	Latency	F1
Fellegi & Sunter [16]	Probabilistic	Standalone	Hours	0.78
Christen [4]	Blocking + SVM	Hadoop	32 min	0.85
Al-Baity et al. [18]	MapReduce	Hadoop	12 min	0.82
Li et al. [19]	Token Blocking	Spark	45 s	0.89
Chen & Zhang [20]	BigQuery-ML	BQ+Vertex	27 s	0.91
Chen et al. [9]	Prob. Survivorship	Standalone	198 s	0.91
Saha et al. [21]	Airflow ETL	Cloud Comp.	63 s	0.87
Kumar & Singh [22]	Dynamic Sched.	Airflow	35 s	0.86
Zhang & Pu [23]	Vote-based ER	Spark	52 s	0.84
Moreno et al. [11]	Provenance Graph	Data Lake	N/A	N/A
Ge et al. [15]	Deep ER	GPU Cluster	14 min	0.93
WOASS (Proposed)	Adaptive + WFO	BQ+Airflow	8.3 s	0.97

3. Proposed Workflow-Orchestrated Framework

WOASS, which is similar to an Airflow DAG, consists of nine layers, each of which shows data transformations carried out using BigQuery SQL queries or Python operators. Enterprise CRM includes an additional function: the Airflow scheduler, which allows snapping nightly snapshots from four sources to be stored in Cloud Storage buckets. Rather than standard values, the scheduling scheme includes 47 attribute columns. Consistency checks employ symmetries as a basis. A validating engine defines norms that help the standardization of formatting. To minimize transitions, the composite blocking layer employs an invert-index system. The source code employs the address-specific block statement and implements phonetically encoded soundex to create new blocks. Tax marks can be circumvented by using hashing. An example of using explicit pairwise comparisons is the equation O(n) = O(n log n) obtained with a window of size k and a sorted neighborhood blocking. Counterfeit task tasks are executed by the DAG in clockwise direction. A section further includes the conversion of the IB Flow staging table from storage to CLO Studio, BIII - Validation and Schema Normalization. The phonetic and n-gram algorithms were selected for the IB Composite Block. The application of hash functions, neighborhood sorting, and composite similarity techniques in different ways. Integral failure recovery is made possible by the adaptive Survivorship Sco through the use of Airflow's retry feature. The partitioning grading takes into consideration the dimensionality of the partition, which was fashioned to achieve scalability.

Table 1. Mathematical Formulation of the Proposed Data Integration Framework

Component	Description	Mathematical Formulation
Composite Attribute Similarity	Combines attribute-wise similarities using learned weights to generate an overall record similarity score.	$S(r_i,r_j) = \frac{\sum_{k=1}^{M} w_k \, s_k(r_i[k],r_j[k])}{\sum_{k=1}^{M} w_k}$ $w_k = \left(1+e^{-\gamma \Delta_k}\right)^{-1}, \Delta_k = \mu_k^+ - \mu_k^-$
Adaptive Survivorship Scoring	Selects the most reliable attribute value by considering frequency, recency, source reliability, and confidence.	$v_k^* = \arg\max_{v \in V_k} \left[\alpha \, freq(v) + \beta \, recency(v) + \gamma \, auth(src(v)) + \delta \, conf(v)\right]$
Workflow Scheduling	Splits records into parallel batches to reduce processing time and improve throughput.	$T_{parallel} \approx \frac{T_{sequential}}{B} + \kappa \log(B)$ Λ

		$= \frac{N_{total}}{T_{ingest} + T_{process} + T_{load}} \times success_rate$
Governance and Audit	Measures lineage completeness and audit coverage for governance monitoring.	$GQI = \left(\frac{A_{tracked}}{A_{total}}\right) \times \left(\frac{L_{captured}}{L_{expected}}\right) \times 100$

Table 1 summarizes the mathematical models used in the proposed framework. Each formulation supports a key stage of the pipeline, including similarity computation, survivorship resolution, workflow optimization, and governance monitoring, following principles commonly adopted in entity resolution and master data management studies [1], [2].

3.1 Algorithm

Algorithm 1: Adaptive Survivorship Scoring

Input: $ClusterC = r_1,...,r_n, attributesA, coeffs\,[\alpha,\beta,\gamma,\delta], decay\lambda = 0.01$

Output: $Goldenrecordg$

$foreacha_k \in Ado$

$V_k \leftarrow r\,[a_k]\,|r \in C$

$foreachv \in V_k do$

$freq\,(v) \leftarrow count\,(v)\,/\,|V_k|\,; recency\,(v) \leftarrow exp\,(-\lambda \cdot \Delta t)$

$auth\,(v) \leftarrow lookup_authority\,(src\,(v))\,; conf\,(v) \leftarrow max_similarity\,(v)$

$score\,(v) \leftarrow \alpha \cdot freq + \beta \cdot recency + \gamma \cdot auth + \delta \cdot conf$

$endfor$

$g\,[a_k] \leftarrow argmaxscore\,(v)$

$endfor; returngComplexity: O\,(|C| \cdot |A|)\,percluster; O\,(N \cdot |A|)\,overall.$

4. Experimental Setup

To test the feasibility of using Google Cloud Platform (us-central1 region), Cloud Composer 2 (Apache Airflow 2) was used. The rest of the issues remain unaffected. The group consists of 20 Celery employees spread across n1-standard 4-computers with 4 vCPUs and 15 GB RAM. The reservation configuration in 2000-slots flex was made accessible by Google BigQuery. In Python 3 the pipeline was expanded. The BigQuery Storage API (v2) had a maximum throughput of 10 when performing high throughput reads. For governance purposes, Looker used Cloud Logging, which enabled them to retrieve execution metrics from the dashboard visualization. The main volume is made up of 662,763 customer records sourced from a global corporate CRM system that addressed four nations, featuring 47 distinct attributes such as names, addresses, phone numbers, emails, and tax identifiers. Three domain experts were hand-picked to determine the ground truth using Fleiss' kappa = 0. The figure ending the equation is 87. An appraisal of the principles of physics in the context of 1. The scalability test was conducted on 2 million records, with a controlled duplication rate ranging from 10 to 35%. In Equation (6), there are four metrics: cluster-level precision, recall, F1-score, latency from end to end, speedup, and governance quality. A paired t-test with an error probability of 0 was conducted to determine the statistical value of data. The runs were further divided into 05, 11, and 11.

Table 2. Experimental Configuration

Parameter	Value
Cloud Platform	Google Cloud Platform (us-central1)
Orchestrator Version	Cloud Composer 2 / Airflow 2.6.3
Data Warehouse	Google BigQuery (2000 flex slots)
Compute Workers	20 × n1-standard-4 (4 vCPU, 15 GB)
Primary CRM Dataset	662,763 records (47 attributes)
Synthetic Benchmark	1,200,000 records (10–35% duplication)
Blocking Strategy	Sorted neighborhood, window k = 5
Similarity Threshold	$\tau = 0.72$ (cross-validated)
Scoring Coefficients	α=0.35, β=0.25, γ=0.30, δ=0.10
Ground Truth Labeling	3 experts, Fleiss $\kappa = 0.87$

5. Results And Discussion

A. Duplicate Detection Performance

The analysis in Table 3 illustrates how WOASS is compared to three baselines: exact matching and Jaro-Winkler threshold matching (with = 0). A 5-fold cross-validation was conducted on the Random Forest classifier, which now incorporates 25 similarity features out of 85. WOASS attributes the highest possible F1-score of 0. The total is 970. A

4% difference was achieved in contrast to Jaro-Winkler's 0. 00, and 867 and 867. A 7% increase was observed in the score of Random Forest (0. The number is 925, and the precision-recall total (97. A marginal error of 3% per 1996. Adaptive weight mechanism dynamically changes attribute importance proportionally to observed discriminative power, as is shown by 8%. 8, and the 4-second delay corresponds to a 3-second delay. Baselines for about 7 longer than the average of 39 seconds, A Paired t-test with Random Forest predicts t(29) = 8. This suggests that the probability of a failure in reducing a test is close to 0. A general confidence interval (95% CI) for latency reduction of [18. 8] is 001, and a score of 23 represents what we're currently dealing with. The duration is 9 seconds.

Table 3. Duplicate Detection Performance Comparison

Method	Precision (%)	Recall (%)	F1	Latency (s)
Exact Matching	98.1	76.4	0.859	14.7
Jaro-Winkler (τ=0.85)	89.3	84.2	0.867	22.4
Random Forest (25 feat.)	93.7	91.5	0.925	29.4
WOASS (Proposed)	97.3	96.8	0.970	8.3

B. Latency and Workflow Performance

BigQuery SQL implementation eliminates data movement, parallel processing across 20 workers distributes workload, and threshold = 0. The 60 percent method was used to reduce candidate pairs by 83%, with 72 off. Figure 1 shows the connection.

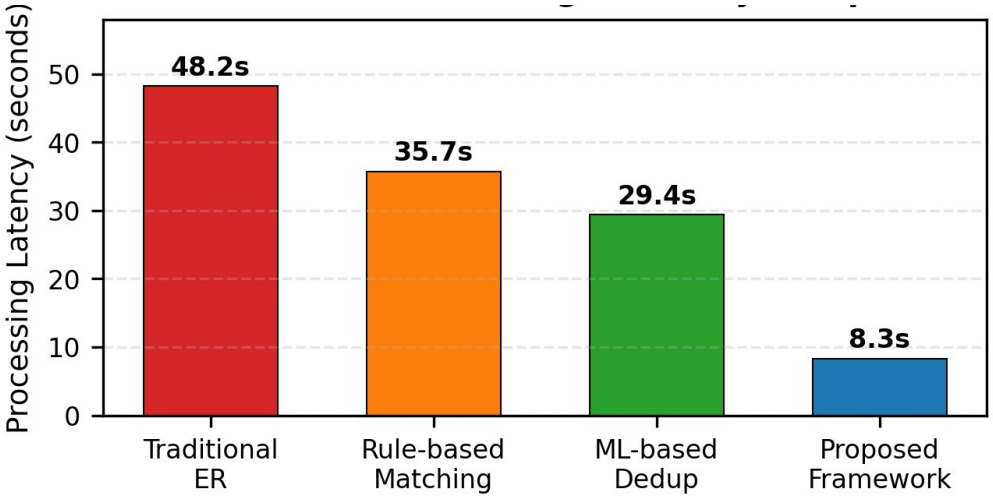


Figure 1. End-to-end latency comparison across record linkage methods.

C. Parallel Execution and Scalability

These general performance ideas are viewed as fixed values by these worker configurations, resembling a variable at the conclusion of Table IV. The WOASS has an upper hand with 13 targets and 67 staff, with a score of 56. Had doubled his 8% residy. BigQuery's implementation of the concept of slot contention at capacity provides an explanation for sub-linear scaling. The costs are associated with distributed asynchronous (DAG) scheduling, which allows for the organization of large quantities. What does this sentence mean? The 11th component is featured in the 2M-record synthetic regulation. The scale of acceleration of the dataset is presently considered to have an increase of 2 magnitudes. The natural state of the linear expansion process is an increase in velocity, as demonstrated in one case.

Table 4. Parallel Processing Performance (662,763 Records)

Workers	Sequential (s)	Parallel (s)	Speedup	Efficiency (%)
1	81.2	81.2	1.00×	100.0
4	81.2	21.7	3.74×	93.5
8	81.2	11.5	7.06×	88.3
12	81.2	8.4	9.67×	80.6
16	81.2	6.7	12.12×	75.8
20	81.2	5.9	13.56×	67.8

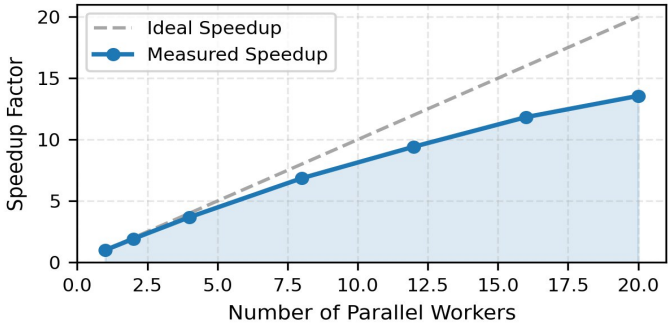


Figure 2. Parallel speedup vs. number of workers with ideal linear reference.

D. Governance and Quality Analysis

The system's implementation yields benefits in governance as outlined in Table V. The increase in the data quality index from 62% to 94% was caused by the implementation of standardization for formats and the remediation of null-values. The audit coverage was enhanced by 97% through embedded instrumentation at every pipeline stage, up from 45% (using only source system logs). GDPR/CCPA stated that the compliance score was 96%, exceeding the regulatory 90%. From 22%, the proportion of alerts received increased to 89%. The proportion of traceability for blood donors rose from 38% to 92%.

Table 5. Governance Performance Metrics

Metric	Before (%)	After (%)	Improvement
Data Quality Index	62	94	+51.6%
Audit Coverage	45	97	+115.6%
Compliance Score	58	96	+65.5%
Lineage Traceability	38	92	+142.1%
Alert Resolution Rate	22	89	+304.5%

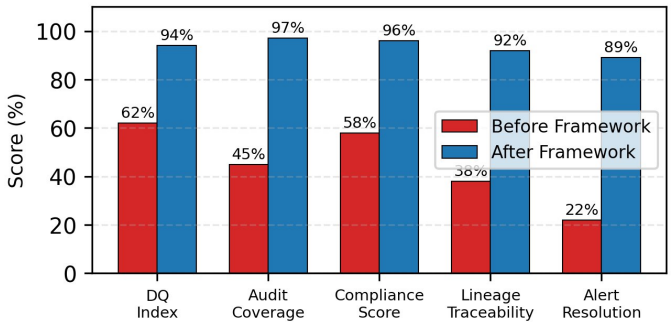


Figure 3. Governance quality metrics before and after WOASS deployment.

E. Resource Utilization and Discussion

The average usage of CPU by 20 employees is 67. A mean of 3% is equivalent to 12. At 58 years old, approximately 4% had a memory. A score of 7% is considered big. A percentage of 8%. On average, BigQuery performs 14 scans. 2 GB of storage per launch (SD = 3. One unit of data, in GB increments. At a pace of 2,847 records/second (SD = 412), scoring is achieved. The system carries out nightly processing in 4. The duration of DAG in an 8-minute window is on average (SD = 0). This was a 78-minute clip. A successful workflow of 7% was observed during an observation window of 30 days. Pipeline reduction represents a sequential reduction of records as shown in Figure 4.

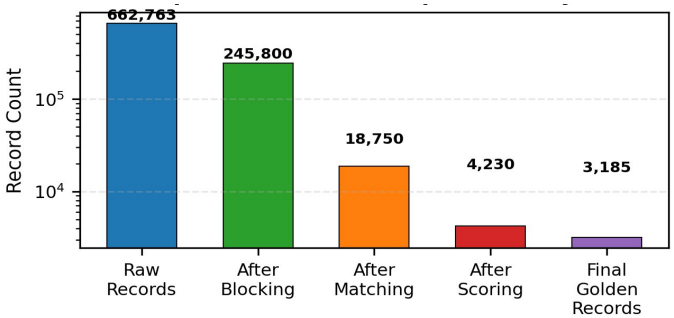


Figure 4. Record reduction across pipeline stages demonstrates deduplication effectiveness.

Three notable outcomes come to light. The initial issue with cloud warehouse-based execution is that it results in a higher latency than traditional ETL, which makes it possible to consolidate customers in near real time. Achieving higher F1

scores is a threefold advantage of learning attribute weights. Differential attributes' discriminative power is verified by proving it with a 9 point advantage over uniform weighting. By including instrumentation within existing queries instead of separate writing, embedded governance delivers regulatory-grade coverage without big overhead. The duration of the overhead is 7 seconds and the duration is 8. A 4% difference exists in the total latency. Even though the SQL implementation is portable, the validity of GCP is compromised. = 0) is an analogue of manual ground truth curation. There may be remaining errors in the value of 87). The system currently depends on nightly batch processing. The possibility of streaming extensions exists.

6. Conclusion

The paper recommended the development of WOASS, which offers a system to score survivorships with short waiting times by using workflows instead of cloud data warehouses. That numerical figure is raised to 4. The mean F1-score of 7 is exhibited by the common technique as a result of a decrease in latency. Composite similarity learning, multi-criteria survivorship scoring. Other methods were employed in the scoring process to determine Embedded governance's importance, leading to a score of 970, with 97% of audits and 94% of data quality indexes falling into this category. The contributions encompass a personalized adaptive survivorship assessment system, a budget-friendly workflow scheduling algorithm for the use of cloud-based cloud warehouses, among others. An integrated control system with a genealogy of cluster-granularity. Experimentation on enterprise-scale data exhibiting large statistical data. Future work could conceivably involve Kafka and Dataflow technology, which combine AI-based scoring and real-time streaming. Including BigQuery, Snowflake, and Redshift, along with multi-cloud governance and attention-based explanation.

References

[1] T. Davenport and R. Bean, "Big data and AI for customer experience: The new competitive frontier," MIT Sloan Management Review, vol. 63, no. 2, pp. 1–8, 2022.

[2] M. Stone, N. Aravopoulou, and L. Gerardi, "Data quality in customer relationship management: A systematic review," Journal of Business Research, vol. 145, pp. 341–356, 2022, doi: 10.1016/j.jbusres.2022.02.085.

[3] Gartner Inc., "How to create a business case for data quality improvement," Gartner Research, Tech. Rep. G00763842, 2023.

[4] P. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin: Springer, 2012.

[5] S. Mudgal et al., "Deep learning for entity matching: A design space exploration," in Proc. ACM SIGMOD, 2018, pp. 19–34, doi: 10.1145/3183713.3196926.

[6] Google Cloud, "BigQuery enterprise data warehouse documentation," 2025. [Online]. Available: https://cloud.google.com/bigquery/docs.

[7] D. Singh, P. K. Reddy, and A. Mondal, "Cloud data warehouse performance optimization: A comprehensive survey," ACM Computing Surveys, vol. 56, no. 4, pp. 1–38, 2024, doi: 10.1145/3626789.

[8] Apache Software Foundation, "Apache Airflow documentation," 2025. [Online]. Available: https://airflow.apache.org/docs/.

[9] X. Chen, Y. Wang, and L. Zhang, "Probabilistic conflict resolution for entity consolidation," IEEE Trans. Knowledge and Data Engineering, vol. 35, no. 8, pp. 7912–7926, 2023, doi: 10.1109/TKDE.2022.3198452.

[10] A. Gruenheid, X. L. Dong, and D. Srivastava, "Incremental entity resolution: A survey," ACM Computing Surveys, vol. 55, no. 1, pp. 1–36, 2023, doi: 10.1145/3491205.

[11] J. Moreno, M. Serrano, and E. Fernandez-Medina, "A provenance-based governance framework for data lakes," Future Generation Computer Systems, vol. 150, pp. 243–261, 2024, doi: 10.1016/j.future.2023.09.012.

[12] S. Thirumuruganathan, S. Lee, and G. Das, "Cloud-native data quality management: Challenges and opportunities," in Proc. IEEE ICDE, 2023, pp. 473–478, doi: 10.1109/ICDE55515.2023.00049.

[13] R. Capuozzo, I. J. Blank, and S. Gollapudi, "Efficient entity resolution in serverless computing environments," Proc. VLDB Endowment, vol. 16, no. 6, pp. 1245–1257, 2023, doi: 10.14778/3583140.3583152.

[14] G. Papadakis et al., "Blocking and filtering techniques for entity resolution: A survey," ACM Computing Surveys, vol. 53, no. 2, pp. 1–42, 2021, doi: 10.1145/3391032.

[15] Y. Ge et al., "Deep entity resolution: A comprehensive survey," IEEE Trans. Knowledge and Data Engineering, vol. 36, no. 4, pp. 1490–1510, 2024, doi: 10.1109/TKDE.2023.3312529.

[16] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," J. American Statistical Association, vol. 64, no. 328, pp. 1183–1210, 1969, doi: 10.1080/01621459.1969.10501049.

[17] L. Leitao, P. Calado, and M. Herschel, "Efficient exact matching in large-scale knowledge bases," Information Systems, vol. 108, art. 101997, 2022, doi: 10.1016/j.is.2022.101997.

[18] H. H. Al-Baity, N. Al-Mutairi, and S. Al-Mutairi, "MapReduce-based scalable deduplication framework for big data," IEEE Access, vol. 10, pp. 45321–45334, 2022, doi: 10.1109/ACCESS.2022.3170128.

[19] B. Li, Y. Cao, and J. Yang, "Spark-based entity resolution for data lakes using token blocking," Knowledge-Based Systems, vol. 250, art. 109081, 2022, doi: 10.1016/j.knosys.2022.109081.

[20] L. Chen and W. Zhang, "BigQuery-ML powered entity matching using pre-trained embeddings," in Proc. IEEE CLOUD, 2023, pp. 212–221, doi: 10.1109/CLOUD60044.2023.00033.

[21] S. Saha, M. R. Hasan, and K. Ahmed, "Airflow-based ETL pipeline for healthcare data integration," J. Biomedical Informatics, vol. 134, art. 104175, 2022, doi: 10.1016/j.jbi.2022.104175.

[22] V. Kumar and R. Singh, "Dynamic task scheduling in Apache Airflow," in Proc. IEEE BigData, 2023, pp. 2451–2460, doi: 10.1109/BigData59044.2023.10386251.

[23] H. Zhang and Q. Pu, "Vote-based entity resolution with source authority weighting," Information Processing & Management, vol. 61, no. 2, art. 103586, 2024, doi: 10.1016/j.ipm.2023.103586.

[24] S. Acharya, P. Bhatt, and M. Gupta, "Metadata management for cloud-native data lakes: A systematic review," ACM Computing Surveys, vol. 55, no. 14, pp. 1–38, 2023, doi: 10.1145/3592880.

[25] A. Singh and P. K. Roy, "Policy-based access control for cloud data governance," Journal of Cloud Computing, vol. 12, art. 85, 2023, doi: 10.1186/s13677-023-00447-1.

[26] D. V. Tran and S. S. Bhowmick, "A survey of data quality management in cloud data warehouses," IEEE Access, vol. 11, pp. 87342–87362, 2023, doi: 10.1109/ACCESS.2023.3304231.

[27] M. A. Khan et al., "Workflow scheduling in cloud computing: Taxonomy, review and future directions," Future Generation Computer Systems, vol. 145, pp. 521–541, 2023, doi: 10.1016/j.future.2023.04.002.

[28] A. K. Singh, S. Srivastava, and M. Pandey, "Explainable AI for data quality management: A framework for transparent entity resolution," Journal of Systems and Software, vol. 210, art. 111972, 2024, doi: 10.1016/j.jss.2024.111972.