International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 6 Issue 3 PP 152-157, 2025
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V6I3P119

Original Article

Master Data Management in Multi-Cloud Environments: A Survey with Operational Evidence from Banking and Insurance Deployments

Kuladeep Sandra
Independent Researcher, USA.

Received On: 19/07/2025 ***Revised On: 07/08/2025*** ***Accepted On: 09/09/2025*** ***Published On: 22/09/2025***

Abstract *- Master Data Management (MDM) is exponentially harder in multi-cloud, multi-region, multi-domain environments than in the single-cloud monolithic settings for which classical MDM patterns were designed. We survey the evolution of MDM patterns in light of operational evidence drawn from a hybrid on-premises and multi-cloud deployment spanning 6 business units, 14 source systems, and roughly 4 million source identifiers that consolidate to approximately 2.5 million unique entities. We examine four architectural patterns (centralized hub, distributed registry, federated masters, and hybrid) and report that data residency constraints make centralized hubs operationally infeasible for our context, while federated masters with reconciliation contracts have proven viable. We report production entity resolution results: 92% precision and 45% recall for deterministic rules, 87% and 62% for probabilistic Fellegi-Sunter matching, and 89% precision with 76% recall for a learned Siamese model, with cross-region privacy constraints reducing matching accuracy by 3 to 5 percent. We describe a compliance audit that discovered Personally Identifiable Information (PII) in 14 unknown analytical tables, requiring 40TB of deletion and a rebuild of the pseudonymization pipeline. We discuss governance structures that succeed where centralized governance creates bottlenecks, and identify open research challenges in private entity resolution, cross-organizational governance, and validation of MDM correctness under GDPR and CCPA constraints.*

Keywords *- Master Data Management, Entity Resolution, Multi-Cloud Architecture, Data Governance, Privacy Constraints, Data Residency, Record Linkage.*

1. Introduction

Master data is the connective tissue of an enterprise. Customers, accounts, products, suppliers, and locations are the foundational entities that multiple business processes depend on, and when these entities are inconsistent across systems, every downstream artifact becomes suspect. Analytics counts diverge between reports. Customers receive duplicate communications. Risk calculations rest on records that no two systems agree about. The discipline of Master Data Management (MDM) exists to prevent this drift by maintaining authoritative records (golden records) of the entities that matter most.

MDM is hard even in the single-cloud, single-region case. In multi-cloud, multi-region, multi-organizational-domain environments, it becomes a fundamentally different problem. The deployment spans 6 business units across retail banking, commercial banking, and insurance, with 14 source systems each carrying its own definition of what a 'customer' is. Cross-unit data products (a customer lifetime value model that spans divisions, a unified risk profile, a regulatory submission that must reconcile counts) are routinely blocked not by technical limitations but by entity definition disputes. In year three of operation, a regulatory audit discovered PII in 14 unknown analytical tables, forcing the deletion of 40TB of intermediate data and a rebuild of the entire pseudonymization pipeline. These are not exceptional events; they are the operational reality of MDM at scale.

The infrastructure context illustrates the constraints. The organization operates an on-premises lakehouse built on Apache Iceberg and Spark, running on 24 Ceph storage nodes and 32 bare-metal Kubernetes nodes. They also operate cloud analytics workloads across multiple public cloud providers. Insurance regulations in several jurisdictions require that policy data remain on-premises or in specific cloud regions, and GDPR imposes hard limits on where European personal data may reside. The classical MDM assumption (a centralized hub that consolidates everything) is simply illegal. The architectural question is therefore not 'where do we put the golden record?' but 'how do we maintain coherent master data when no single location is permitted to hold all of it?'

Multi-cloud MDM is harder than traditional MDM along several axes simultaneously. Network latency between a hub and geographically distributed source systems makes synchronous reconciliation impractical. Data residency constraints prevent consolidation. Eventual consistency in multi-region replication introduces windows during which master data is momentarily inconsistent. Organizational complexity compounds the technical difficulty: on-premises systems are governed by IT under slow change cycles, while cloud systems iterate rapidly with frequent schema changes,

and the two cultures clash at every reconciliation boundary. Finally, multi-source conflicts are not always resolvable on technical grounds. When System A says a customer is an individual, System B says a customer is a contact at an organization, and System C says a customer is a household sharing a policy, no algorithm can decide which definition is 'correct' without a business judgment.

This paper organizes around three applied research questions. (RQ1) What MDM architectural patterns are viable in multi-cloud environments, and where do centralized and distributed approaches succeed or fail? (RQ2) How can entity resolution be scaled across distributed systems under privacy constraints that forbid full dataset scanning and PII export across regions? (RQ3) What governance structures enable efficient entity definition and conflict resolution across organizational domains without creating centralized bottlenecks? These questions are addressed in turn, grounding the analysis in operational evidence from a deployment of more than 500,000 tables across 6 business units in 3 time zones.

2. The MDM Challenge in Multi-Cloud Environments

Master data is distinguished from operational data by three properties. It is low in volume (millions of records, not billions). It is high in impact, because many downstream processes depend on its consistency. And it is high in cost-of-error: an inconsistent customer record produces wasted marketing spend, a failed risk calculation, or a regulatory finding. Classical MDM solves this by designating a system of record, reconciling inputs from multiple sources, and publishing golden records to dependents. The assumption is that consolidation is permitted.

2.1. Regulatory Constraints on Data Residency

In the deployment environment, consolidation is not permitted. GDPR requires that personal data remain in specific jurisdictions. CCPA grants California residents access and deletion rights that depend on data being identifiable as theirs. Insurance regulations in several countries mandate that policy data reside on-premises or in designated cloud regions. The organization cannot copy European customer records to a North American hub for matching. It cannot export Californian records to an EU region. It cannot store insurance master data exclusively in a public cloud. The architectural consequence is that master data must be distributed by construction, with per-region golden records and explicit synchronization rules.

2.2. Source System Landscape and Entity Definition Conflicts

The source system landscape compounds the difficulty. Three systems illustrate the spectrum. System A is a 25-year-old ERP that runs daily batch processes and defines a customer as an individual person identified by a legal ID. System B is a modern event-driven SaaS CRM that defines a customer as a contact record at an organization, identified by company tax ID. System C is an insurance underwriting platform that defines a customer as a policyholder, which may be either an individual or an organization, identified by an underwriting reference. Whether the same real-world party appears across all three systems depends entirely on the question being asked. For billing, they may be three distinct entities. For consolidated risk, they may be one. For marketing, the answer depends on consent.

Across the 6 business units, the numbers tell the same story. Retail Banking holds 2 million customer records, primarily individuals identified by government ID. Commercial Banking holds 500,000 customer records that are organizations identified by tax ID. Insurance holds 1.8 million customer records organized as households of policyholders identified by policy number. A cross-unit analysis of 'customer lifetime value' must reconcile roughly 4 million source identifiers down to 2.5 million actual unique entities, and that figure depends on whether one accepts that an insurance policyholder counts as a 'customer' at all—a question on which the business units have never fully agreed.

Entity definition disputes recur in predictable patterns. The customer scope dispute asks whether retail, commercial, and insurance records refer to the same population. The individual-versus-household dispute pits Commercial and Insurance (which both naturally model households) against Retail (which models individuals). Temporal disputes ask whether master data should track historical changes (names, addresses, contacts) or only the current state. Consent disputes ask how to honor a GDPR or CCPA erasure request when other business units retain a legitimate need for the record. None of these questions has a purely technical answer. The operational impact of entity disputes is severe and measurable: a customer risk profile product was blocked for three weeks while Risk and Marketing argued over the entity definition; two reports on customer churn produced different numbers because one used the Retail definition and the other a broader cross-divisional definition; when a regulator asked the routine question 'how many customers do we have?' the resulting internal investigation discovered 14 conflicting definitions of customer in active use across the enterprise.

3. MDM Architectural Patterns in Multi-Cloud

Four architectural patterns recur in the MDM literature, and direct operational experience has been gained with all four.

3.1. Centralized Hub Pattern

The centralized hub is the classical pattern: source systems feed into a single system of record that publishes golden records to dependents. Its strengths are clarity (a single version of truth), centralized reconciliation logic, and a clear audit trail. Its weaknesses are equally clear: the hub is a throughput bottleneck and a single point of failure, and in this environment it is structurally illegal because consolidating GDPR-protected and insurance-regulated data in one region violates residency requirements. The first attempt at a centralized customer master hub failed for exactly this reason. The hub became both a data residency violation and a delivery bottleneck simultaneously.

3.2. Distributed Registry Pattern

The distributed registry pattern decouples governance from storage. A central metadata registry maintains entity definitions and reconciliation rules, while source systems retain local master copies of the actual data. The registry directs reconciliation but does not store the records themselves. This pattern fits multi-cloud well because metadata can be centralized without violating residency. The second attempt used a distributed registry, and it worked acceptably for entities with clear ownership (Commercial customers owned by Commercial Banking, for instance), but it broke down on contested entities, particularly household data shared between Commercial and Insurance, where neither domain accepted the other's reconciliation rules.

3.3. Federated Masters Pattern

Federated masters push autonomy further. Each domain owns the master data for its entities, domains commit to synchronization contracts, and cross-domain matching is performed at query time rather than in source systems. There is no global golden record; instead, there is a reconciled view assembled when needed. The strengths are domain autonomy, clear ownership, and natural fit with multi-cloud residency (since each domain can place its data in its preferred region). The weaknesses are reconciliation complexity at query time and larger inconsistency windows. The third attempt moved to federated masters: Retail owns the individual customer master of 2 million records, Commercial owns the organization master of 500,000 legal entities, and Insurance owns the household master of 1.8 million policyholder groups. Reconciliation occurs in a separate customer 360 view rather than in source systems. This is the pattern that has worked best.

3.4. Hybrid Pattern

The hybrid pattern combines a centralized hub for low-volume, high-criticality entities with a registry for everything else. Commercial Banking uses this approach: roughly 500 'global accounts' (multinational corporations whose relationships span every division) are managed in a centralized hub, while the remaining 499,500 commercial customers live in the distributed registry. The strength is that strong governance is concentrated where it matters most. The weakness is the complexity of running two parallel governance models and the recurring edge cases of when an entity should graduate from one to the other. For the regulated, multi-jurisdiction environment, federated masters dominate; for an organization without residency constraints, a centralized hub may still be the right answer.

4. Entity Resolution at Scale with Privacy Constraints

Entity resolution is the technical core of MDM. Given an entity E in System A and an entity F in System B, the task is to determine whether E and F represent the same real-world thing. Three classes of approaches dominate the literature. Deterministic rules apply explicit logical conditions: if name matches and address matches and date of birth matches, then the records refer to the same customer. Probabilistic approaches, of which the Fellegi-Sunter model is canonical, compute likelihood scores for match versus non-match based on the agreement of multiple attributes. Learning-based approaches, including modern neural architectures, train on labeled pairs to produce match probabilities directly. The multi-cloud privacy constraint changes the problem fundamentally: the organization cannot export PII across regions for matching. Moving California customer records to an EU server, or moving European records to a North American region, would violate CCPA and GDPR respectively. Matching must occur within privacy boundaries, which means the matching algorithm itself must be designed around the constraint that fully-identifying attributes never cross regional borders.

4.1. Privacy-Preserving Entity Resolution Pipeline

The pipeline addresses this with a six-step process. First, identifiers (name, address, date of birth) are tokenized locally within each region. Second, hash signatures are computed locally and are deterministic but non-reversible. Third, only hash signatures are exchanged across regions, never raw PII. Fourth, matching occurs on hash signatures combined with de-identified attributes such as the last four digits of a phone number or a zip code. Fifth, candidate matches are returned to the regions with confidence scores. Sixth, regions confirm matches independently without ever sharing full PII. This structure preserves regulatory compliance at the cost of some matching accuracy.

4.2. Blocking and Learned Matching Models

Scale is the second major constraint. With 4 million source identifiers across 14 source systems, naive pairwise matching would require 16 trillion comparisons. At one microsecond per comparison this is computationally infeasible, and in any case the vast majority of comparisons would be obvious non-matches. Scale is addressed with blocking and sketching: records are first grouped by region, by industry segment, and by age band, reducing the comparison space from 16 trillion to roughly 2 billion. Within each block, a learned matching model is applied. The learned model is a Siamese LSTM network trained on 50,000 human-verified labeled pairs that takes tokenized names, encoded addresses, and phone-number features as input and emits a match probability.

The results illustrate the precision-recall tradeoff sharply. Deterministic rules achieve the highest precision at 92% but miss more than half of true matches at 45% recall. Probabilistic matching trades roughly five percentage points of precision for seventeen points of recall. The learned model recovers most of the deterministic precision (89%) while substantially improving recall to 76%. The remaining 14% false negatives in the learned model trigger manual review by a human steward. Privacy constraints reduce all of these accuracy figures by three to five percent compared to a hypothetical no-constraint baseline, because fully-identifying attributes such as Social Security numbers and full legal names cannot participate in cross-region matching.

4.3. Pseudonymization and Temporal Consistency

Successfully matched entities receive a pseudonymous identifier. A retail customer R123, a commercial organization C456, and an insurance household I789 may all receive a single pseudonymous ID such as CUST-2M-89342, which enables a customer 360 view without revealing individual identities outside their authorized boundaries. The pseudonymization step adds 2 to 3 seconds of latency to a typical reconciled view, because reconstructing individual records from pseudonymous identifiers requires joins back to source systems. Three challenges recur: feedback loops require disciplined logging so that human-rejected matches improve future model versions; temporal consistency requires explicit match-history tracking so that previously-matched entities do not silently unmatch; and consent under CCPA's right to be forgotten requires marking records as archived rather than deleting them outright, since the audit trail must survive the erasure.

5. Governance Patterns

The hardest MDM problems are not technical. They are governance problems disguised as technical problems. Who decides what a customer is? Who owns the master record when two divisions have legitimate claims? Who arbitrates when entity definitions conflict? These questions cannot be answered by an algorithm, and they cannot be deferred indefinitely without paralyzing data product development.

5.1. Centralized Governance Approach

Centralized governance places entity definitions in the hands of a single MDM team. Definitions are debated once, codified, and published, and all domains are required to use them. The strengths are consistency and auditability. The weaknesses are speed and fit: definitions change infrequently, one-size-fits-all definitions rarely match domain-specific needs, and the central team becomes a bottleneck for any new entity or any change to an existing one. This was observed firsthand: a debate over whether to model the customer as an individual, a household, or a legal entity ran for forty weeks of central deliberation before a working compromise was reached.

5.2. Federated Governance Approach

Federated governance pushes definitions entirely to domains. Each business unit defines its own entities, no central authority intervenes, and reconciliation is the consumer's problem. The strengths are speed and fit. The weaknesses are inconsistency and waste: when every domain defines a customer differently, cross-domain analytics become impossible without per-product reconciliation rules. The second governance attempt was federated, and while individual domains moved fast, cross-domain products stalled.

5.3. Distributed Governance with Alignment Incentives

Distributed governance with alignment incentives is the pattern that has worked. Domains own their entity definitions but commit to a shared reconciliation contract: explicit rules for how their definitions map to a common pseudonymous identifier space, and a steering committee for unresolved disputes. Steward-led resolution of typical disputes takes one to two weeks, an order of magnitude faster than the forty-week centralized debate, because the decision authority is delegated to people close to the business reality. Disputes that cannot be resolved at the steward level escalate to a business steering committee that meets monthly.

The dispute faced most often was dual-claim ownership. Commercial Banking would claim a large organization belonged in the Commercial master because of its strategic value as a corporate account. Insurance would claim the same organization belonged in the Insurance master because of a large employer group policy. Centralized resolution would have meant the MDM team picking one division and creating resentment in the other. Federated resolution would have meant both divisions proceeding independently and inconsistently. The distributed-with-escalation approach produced a third answer: dual ownership. The organization lives in both Commercial and Insurance masters, with explicit reconciliation between them. This creates more reconciliation overhead but produces better business outcomes, because both divisions can serve the customer effectively without negotiating cross-divisional access on every transaction.

6. Operational Experience: Compliance Challenges

In year three of operation, a regulatory body audited PII handling. The audit discovered PII in 14 analytical tables that were not authorized to contain it—tables built for aggregate analytics, such as a customer age distribution table, which had inadvertently retained email and phone number columns that were unnecessary for the analytical purpose and created a compliance exposure. The root cause was traceable to the entity resolution pipeline. Intermediate tables that joined source records for matching had been intended as temporary, but query performance optimizations had made them persistent, and PII had leaked from the operational layer into the analytical layer.

6.1. Remediation Process

Remediation proceeded in five steps. First, all tables containing PII were cataloged and the 14 unauthorized cases in the analytical layer confirmed. Second, those tables were classified as Confidential and PII-bearing in the catalog. Third, 40TB of unnecessary historical intermediate data was deleted. Fourth, the reconciliation pipeline was rebuilt to emit only pseudonymous identifiers into the analytical layer, never raw PII. Fifth, attribute-based access control through Apache Ranger was implemented to restrict the remaining authorized PII access to specific roles. The affected analytics were re-run on pseudonymized data and business logic was verified to produce equivalent results.

Four lessons compress the remediation experience. First, privacy constraints are not optional governance overlays; they drive technical architecture decisively. Second, intermediate-table sprawl is a recurring failure mode and requires explicit lifecycle governance, not informal cleanup. Third, pseudonymization introduces real query latency (2 to

3 seconds for a customer 360 view) because reconstructing individual records requires joins back to authoritative sources. Fourth, the GDPR right to be forgotten is operationally complex: removing an individual from master data requires removing them from every derived table, reprocessing affected analytics, and notifying downstream consumers, none of which can be done instantly.

6.2. Gdpr and Ccpa Implementation

The GDPR implementation handles erasure through master data versioning and lineage tracking. Historical records remain visible for audit, but current PII is suppressed after erasure. Lineage tracking identifies every table and report dependent on the affected customer, and an erasure workflow marks the record deleted in the master, invalidates dependent views, and notifies stakeholders. Each erasure request requires roughly 15 minutes of manual review, 5 to 10 minutes of processing, and 30 minutes of impact assessment, for an average fully-loaded cost of $150 per request. CCPA adds parallel obligations for access requests and opt-out-of-sale, where the operational difficulty is that 'sale' is not clearly defined and intra-enterprise data sharing between Retail and Insurance can plausibly be interpreted either way. The organization logs all sharing events, disables sharing for opted-out customers, and audits quarterly.

7. Research Challenges and Future Directions

Several research challenges remain open. The first is the measurement of MDM quality. There is no standard benchmark for entity resolution accuracy at enterprise scale. A useful benchmark would consist of perhaps 100,000 entities with ground-truth matches and non-matches, evaluated on precision and recall. The challenge is that ground truth is expensive (manual verification by subject matter experts costs roughly $10 to $50 per pair depending on complexity) and that real customer data cannot be published for benchmarking. Synthetic benchmarks that preserve matching difficulty without revealing real identity patterns are an open problem.

7.1. Private Entity Resolution at Scale

The second challenge is private entity resolution at scale. Secure multi-party computation and homomorphic encryption offer cryptographic guarantees but at computational costs that make them impractical for the comparison volumes faced. Hash-based matching combined with manual review of edge cases is the practical compromise, but it accepts the 3 to 5 percent accuracy loss documented earlier. Algorithms that close this gap without full PII exchange would substantially improve cross-region MDM.

7.2. Cross-Organizational Governance

The third challenge is governance across organizational boundaries. MDM within a single organization is hard; MDM across organizations is harder. Consider a retailer, a logistics provider, and a financial institution that wish to share a customer master for supply chain finance. They have different governance standards, different security postures, and different privacy tolerances. Federated masters spanning organizational boundaries are an active area of research, and the governance structures that would enable them remain underdeveloped.

7.3. MDM Correctness and Real-Time Matching

The fourth challenge is validating MDM correctness. Without external validation, errors accumulate silently and the customer 360 view becomes unreliable over time. Self-validating matching algorithms that compute meaningful confidence bounds without external verification would change the economics of MDM operation. The fifth is the move from batch to real-time MDM: nightly reconciliation must give way to matching that completes within 100 milliseconds when a new entity arrives, which requires hierarchical approaches combining fast learned blocking with slower precise matching for uncertain cases. The sixth is MDM in federated learning contexts, where organizations want shared analytics without sharing raw data, and federated entity resolution would let them compute shared matches without exposing identities.

8. Conclusion

Master Data Management in multi-cloud environments is operationally complex but necessary. Organizations with distributed infrastructure, multi-jurisdictional compliance obligations, and divergent business unit definitions cannot avoid the entity definition and reconciliation work; the only choice is whether to address it deliberately or to pay its costs in delayed products and audit findings. Centralized hub patterns, despite their elegance in textbooks, fail in this environment because data residency makes consolidation infeasible. Distributed and federated patterns are more viable, with federated masters proving to be the strongest fit for multi-jurisdictional regulated workloads.

Entity resolution is both a technical challenge and a governance challenge, and both dimensions require ongoing investment. The learned matching model achieves 89% precision and 76% recall under privacy constraints that reduce accuracy by 3 to 5 percent compared to an unconstrained baseline. Compliance with GDPR and CCPA is not optional; it shapes architecture from the ground up, as the 40TB remediation and pseudonymization rebuild made painfully clear. Distributed governance with explicit escalation outperforms both centralized and federated alternatives in this experience, but no governance pattern eliminates the need for human judgment on contested entities.

The findings come from banking and insurance and reflect a hybrid on-premises plus multi-cloud architecture under specific regulatory constraints. Manufacturing, healthcare, and pure-cloud organizations will face different tradeoffs, and the patterns recommended should be evaluated against local constraints rather than adopted uncritically. Future work in benchmark development, private matching algorithms, cross-organizational governance, and real-time matching at scale would substantially advance the state of practice. MDM is not solved, and in multi-cloud environments it is unlikely ever to be solved in a single

closed-form way, but the patterns surveyed here provide a defensible foundation for organizations navigating the same constraints.

Conflicts of Interest

The author declares that there is no conflict of interest concerning the publishing of this paper.

Acknowledgements

The author thanks the MDM platform and governance teams whose production deployments and operational metrics made this survey possible.

References

[1] P. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin, Germany: Springer, 2012. https://scholar.google.com/scholar?q=Data+Matching+Concepts+and+Techniques+for+Record+Linkage+Christen

[2] H. Köpcke, A. Thor, and E. Rahm, "Evaluation of entity resolution approaches on real-world match problems," Proc. VLDB Endowment, vol. 3, no. 1, pp. 484–493, 2010. https://scholar.google.com/scholar?q=Evaluation+of+entity+resolution+approaches+on+real-world+match+problems

[3] F. Provost, M. Allen, and S. Rogers, Practical Master Data Management. Sebastopol, CA: O'Reilly Media, 2018. https://scholar.google.com/scholar?q=Practical+Master+Data+Management

[4] E. A. Durham, M. Kantarcioglu, Y. Xue, C. Toth, M. Kuzu, and B. Malin, "Composite Bloom filters for secure record linkage," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 12, pp. 2956–2968, 2014. https://scholar.google.com/scholar?q=Composite+Bloom+filters+for+secure+record+linkage

[5] J. Vaidya, C. Clifton, and M. Zhu, Privacy-Preserving Data Mining. Boston, MA: Springer, 2005. https://scholar.google.com/scholar?q=Privacy-Preserving+Data+Mining+Vaidya

[6] D. Loshin, Master Data Management. Burlington, MA: Morgan Kaufmann, 2010. https://scholar.google.com/scholar?q=Master+Data+Management+Loshin

[7] W. H. Inmon, B. O'Neil, and L. Fryman, Master Data Management and Enterprise Information Management. Basking Ridge, NJ: Technics Publications, 2008. https://scholar.google.com/scholar?q=Master+Data+Management+and+Enterprise+Information+Management+Inmon

[8] DAMA International, DAMA-DMBoK: Data Management Body of Knowledge, 2nd ed. Basking Ridge, NJ: Technics Publications, 2017https://scholar.google.com/scholar?q=DAMA-DMBoK+Data+Management+Body+of+Knowledge.

[9] P. P. Tallon, "Corporate governance of big data: Perspectives on value, risk, and cost," IEEE Computer, vol. 46, no. 6, pp. 32–38, 2013. https://scholar.google.com/scholar?q=Corporate+governance+of+big+data+Tallon

[10] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," Journal of the American Statistical Association, vol. 64, no. 328, pp. 1183–1210, 1969. https://scholar.google.com/scholar?q=A+theory+for+record+linkage+Fellegi+Sunter

[11] Apache Software Foundation, "Apache Iceberg table format specification v2," Technical Documentation, 2024. [Online]. Available: https://iceberg.apache.org/spec/https://scholar.google.com/scholar?q=Apache+Iceberg+table+format+specification

[12] S. Shankar, R. Garcia, J. M. Hellerstein, and A. G. Parameswaran, "Operationalizing machine learning: Challenges and best practices," IEEE Software, vol. 41, no. 2, pp. 42–51, 2024. https://scholar.google.com/scholar?q=Operationalizing+machine+learning+Challenges+and+best+practices

[13] M. Stonebraker and I. F. Ilyas, "Data integration: The current status and the way forward," IEEE Data Engineering Bulletin, vol. 41, no. 2, pp. 3–9, 2018. https://scholar.google.com/scholar?q=Data+integration+The+current+status+and+the+way+forward