ISSN: 2959-6386 (Online), Vol. 2, Issue 1, April 2023

Journal of Knowledge Learning and Science Technology

journal homepage: https://jklst.org/index.php/home

Master Data Management – Disruptive Modern Architecture

Subhodip Pal[1], Taniya Pal[2]

[1]Independent Researcher and Director, Management Consulting Firm, Chicago, IL, USA
[2]Western Governors University, Utah

Abstract

Master Data Management (MDM) concept came into the mainstream during the mid 2000s, to generate a single Golden profile of the customer. As the Enterprise IT architecture started adopting and integrating various Commercial Products for CRM, ERP, Finance, HR, Supply Change Management etc., along with other homegrown custom apps, the information around the Customer started drifting across the various applications which resulted in various inefficiencies in day-to-day business operation. Master Data Management tools and technologies provide a way to perform identity resolution and survive latest and greatest information. Traditionally, the MDM tools always adopt a lean approach that uses minimal attributions to identify the smallest amount of (master) data with the biggest influence on business outcomes like Name, Phone, SSN, Email and Address. These constitute only less than 1% of the enterprise data. In the age of generative AI, there is a greater need to understand the context and the relationship of ALL data. This paper explores an alternate approach to mitigate the shortcomings of the currently available MDM tools

Keywords: MDM, Graph MDM, MDM in Data Warehouse, Snowflake Native Apps, Data Warehouse

Article Information:
Article history: *Received:*11/02/2023 *Accepted:* 25/03/2023 *Online:* 30/04/2023 *Published*: 30/04/2023
DOI: https://doi.org/10.60087/jklst.vol2.n1.p114
[i]Correspondence author: Subhodip Pal **Email: subhodip.pal@gmail.com**

1. Introduction

Master Data Management as the name suggests, only attempts to solve only a fraction of Enterprise data in a siloed application which is typically CapEx and OpEx intensive. Although it is a critical element of IT Architecture, yet there is a continuous need from a broader range of stakeholders, who are seeking the fast, scalable, nimble, single, consolidated, cross-referenced, trusted, enterprise-wide view of ALL data (and not just Master Data, which is typically <1 %), with data lineage to source, for self-service BI and Analytics. Also, CxOs wants to leverage that single, consolidated and trusted data set to train the AI & ML models which includes Master Data, Reference Data and Transactional Data.

Obviously ,some players in the industry has acknowledged the issue and tried to mitigate the problem by introducing a Multi-Domain MDM and providing a blank slate for Data Modelling(like Semarchy ,Informatica Multi Domain) rather than prescribing a pre-built Data Model (Siebel UCM, Reltio) but majority of these tools were built on old architecture based on the principle of storing data in Relational Databases and lifting data out of these data stores , processing in batches in the CPU(Servers) with limited static Memory and then storing back the results into the Relational Database. This doesn't scale and presents an ever-growing problem with the exponential increase in data volume and breadth of attribution. Apart from that it presents challenges with data security both at rest and in motion.

2. Literature Review

Master Data Management (MDM) is a centralized system that maintains golden data across all systems within an organization [1] [2]. It allows for effective communication and data centralization in large organizations with multiple applications and teams working on customer, product, and supplier data [3] [4]. MDM ensures that any system updated with new information will be sent to MDM through events, which then communicates with and updates the sub-systems within the organization's ecosystem [5]. However, operational inefficiencies and scalability constraints have been discovered in the traditional MDM system, resulting in a tedious process of identifying failures, reprocessing event information, and re-synchronizing sub-systems. In this paper, we have highlighted the issues and the potential solution.

3. Methodology

This section will describe the current MDM Architecture and proposed future state.

3.1 *Traditional MDM Architecture*

It has traditional three tier architecture, with a Relational Database as a Datastore. The Application Servers lifts and processes the data in chunks based on the pre allocated CPU and Memory on the Application layer and the UI layer is for Data Stewardship and Governance.

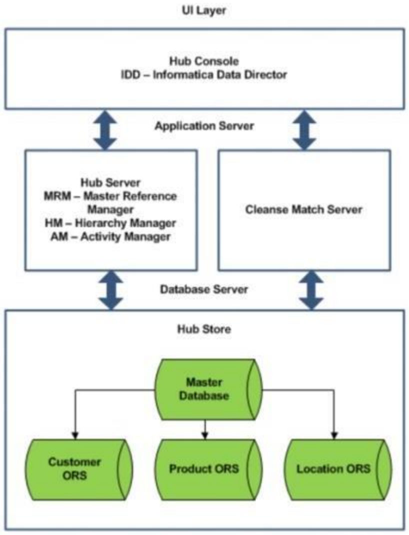


Fig:1[20]

This architecture faces significant challenges when trying to scale with the increase in Data Volume and securing the data across multiple layers and while in motion between these layers. Also, there is minimal capability to process semi-structured and unstructured data like JSON, Parquet etc. natively. There is very minimal data observability capability to monitor and trace individual records, apart from the text based log files.

3.2 *MDM – Traditional Data Engineering*

Before proposing the revised approach, it’s important to understand how the Matching, Merging and Survivorship works.

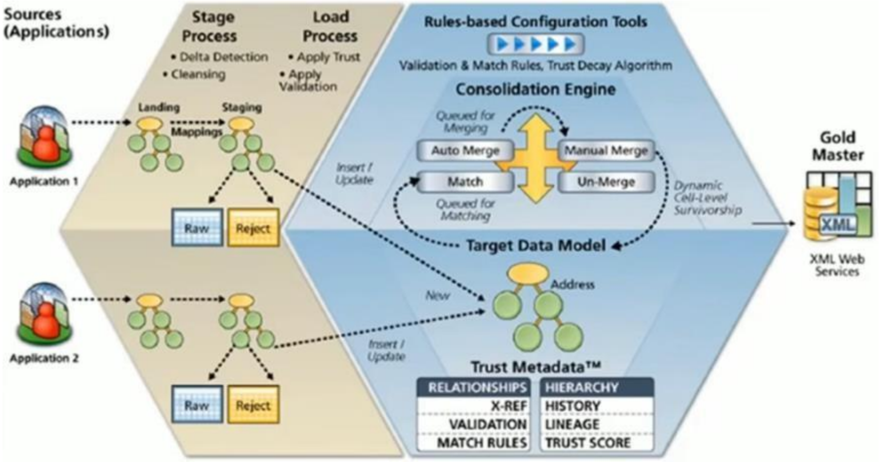


Fig :2 [20]

Here are the high-level steps:

1. Staging – Pull and store the New Batch (Delta) of Data from each of the Sources
2. Loading – Perform Data Quality check and load to Base Tables
3. Tokenize – Perform Clustering
4. Match – Perform deduplication
5. Merge – Consolidate the similar records
6. Survivorship – Pick and choose the attribution from the Merged Entities based on pre-defined rules
7. Publish – Publish Golden Survived records to downstream.

The most computationally intensive part of the process is Step 3-6.

Why do we need to perform Tokenization /Clustering – In order to deduplicate the data we need to try and match each pair of data and evaluate if they are similar. The number of Unique pairs for N records are (N *(N-1) /2) unique pair.

So, It grows exponentially with the increase in Record Volume (N). Clustering process identifies and restrict the pairs that actually need a fuzzy comparison within themselves.

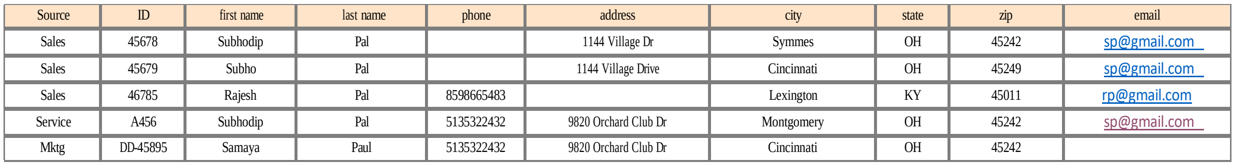
Source records containing person or company data fields	unique pairs
2,000,000,000	1,999,999,999,000,000,000
2,000,000	1,999,999,000,000
65,000,000	2,112,499,967,500,000

This is more of an art, than science and there is no specialized prescribed algorithm and varies based on the dataset.

What is Deduplication and Fuzzy Match –

A fuzzy match is a process to evaluate the similarity of two or more records based on the combination of records attribution. The human brain does this very well, just not so efficient and fast for millions of records, hence the need for an automated system to emulate the human brain.

In the example below which records would you match and why? How many people do you count? Are there any households? What fields are you using?



Source	ID	first name	last name	phone	address	city	state	zip	email
Sales	45678	Subhodip	Pal		1144 Village Dr	Symmes	OH	45242	sp@gmail.com
Sales	45679	Subho	Pal		1144 Village Drive	Cincinnati	OH	45249	sp@gmail.com
Sales	46785	Rajesh	Pal	8598665483		Lexington	KY	45011	rp@gmail.com
Service	A456	Subhodip	Pal	5135322432	9820 Orchard Club Dr	Montgomery	OH	45242	sp@gmail.com
Mktg	DD-45895	Samaya	Paul	5135322432	9820 Orchard Club Dr	Cincinnati	OH	45242	

Why do we need De-Duplication?

- Many systems have people and company data but stored as names and contact info fields
- Even if they have unique IDs the IDs would typically be for one system and there can be overlaps across systems
- The names will have differences due to
  - Nicknames
  - Dropped info (e.g., middle name, initials, Jr/Sr, INC, LLC, …
  - Misspellings
- Stripped accents (e.g., Chloe vs Chloé, Dorfener vs Dörfener, …)
- Mis-ordered Name fields (e.g., Subhodip Pal vs Pal Subhodip)

So, matching people and companies is necessary to enable merging data from different records / sources to find the same person for example, who purchased a car, was marketed to and had a vehicle serviced.By capturing or creating unique IDs from records involving the same person ,all the sales and service records for a given person can be reported on. Matching can also be used to identify records that may be different but belong to the same household to consolidate loyalty points.

3.4 MDM – Algorithms for Fuzzy Match

- Distance algorithms- functions that calculate how “far apart” two strings are based on how many differences between strings.
  - Well known Algorithms has been around since the 1960s like Jaro–Winkler distance, Levenshtein distance [19]
- Phonetic algorithms- functions that change string into a simplified phonetic representation
  - Soundex, Metaphone - Phonetic Algorithms
- Nickname files – a list of common nicknames for a given first name
  - Nicknames are locale specific (i.e., language and region)
- Frequency distribution files - The frequency of names and “noise” words
  - Distribution frequency of names is also locale specific
- Company designators – a list of busines abbreviations commonly used in business names
- Transliteration – Representing names from one script into another (e.g., Kanji to Latin)

Merging and Survivorship – Merging is just linking two or more similar records based on the similarity score and threshold of tolerance level. Survivorship process is picking the latest and greatest attribution among the Matched and Merged entities based on pre-defined business rules like Importance of Source, Recency, Frequency etc.

3.4 MDM – Modern Approach

In order to mitigate the challenges around the traditional MDM tools and the associated architecture, here is the proposed re-design

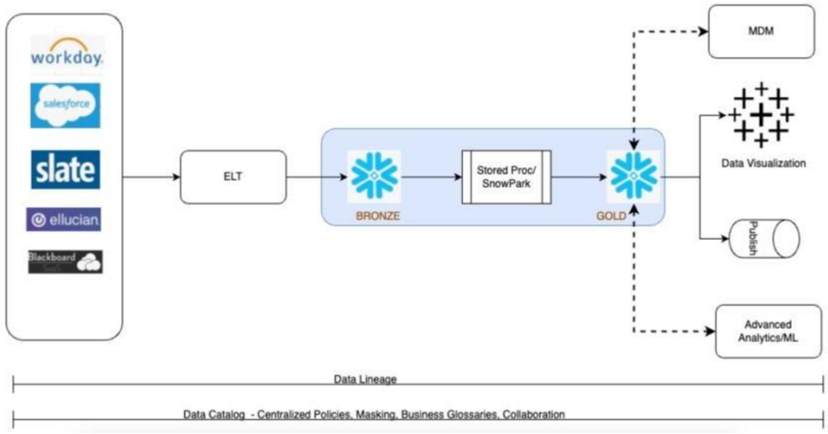


Fig:3

As depicted above, with the evolution of technology and capability to perform complex Data Engineering and ML/AI inside the Warehouse, we can perform these MDM functionalities inside the modern Cloud Datawarehouse like Snowflake.

Leveraging either modern Integration tools like IDMC, Fivetran etc. or using Native Connectors built on Snowpark or similar technologies we can bring the data in Raw format to the Data Warehouse i.e., performing an ELT and not ETL and then build a pipeline within the Datawarehouse to transform the data into a Canonical Data Model, generate Cluster Key and stream the incremental changes to the Matching Engine.
The Matching Engine is nothing but a Containerized set of known and available algorithms, packaged together based on the functional need. In the Snowflake ecosystem, it will be a Snowpark Container Service. Once the match score is determined, then the Merge and Survivorship is a typical Data Engineering job.

Additionally, there is an option to plug in Commercially available traditional MDM tools or Matching Algorithms to the ecosystem, if necessary. Here is a reference architecture on the Azure ecosystem

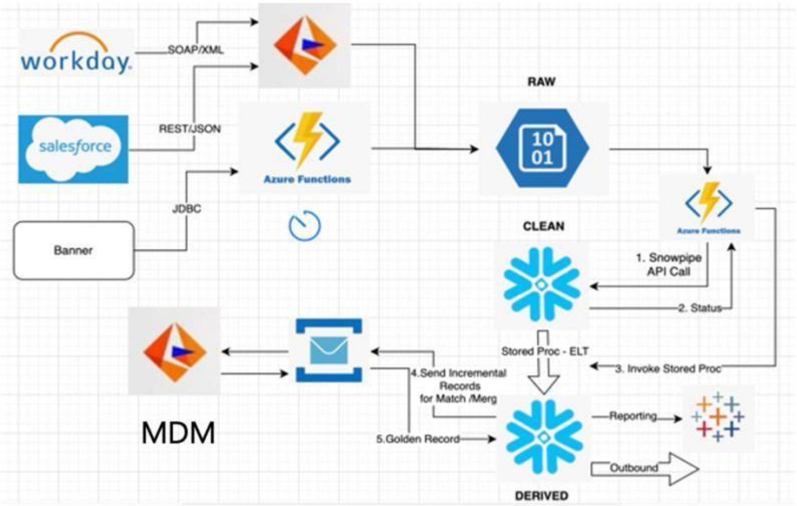


Fig :4

4. Results and Discussion

This modern approach addresses quite a few shortcomings of traditional MDM tools:

1. Scalability – As these jobs are now executed in a modern infinitely on-demand scalable system, the performance bottle necks are eliminated. Co-location of data and compute within the same Virtual Warehouse increases the performance.
2. Security – As the data is not lifted and shifted across multiple tiers, the security risk is reduced.
3. Semi-Structured /Unstructured Data – All data types/loads are supported in these modern platforms, hence the elimination of the need to pre-process data.
4. Cost – As these modern applications are priced base on Consumption, the TCO is drastically reduced as it eliminates the inefficiencies around the sizing for accommodating peak traffic.
5. Single Platform for ALL data – There is no need to have a satellite system to handle the De-Duplication of the Master data and then build complex process to synchronize it back on top of the transactional data every time the data drifts or new data is generated
6. Analytics and AI – Now that all data is under one platform with lineage back to sources and Raw Data, business can perform live Visualization, train AI models all within one system.
7. Centralized Governance – All data can be governed centrally – Lineage, Access Control, Masking, Materialized Views etc.
8. Stewardship – Cloud Native, mobile friendly, user friendly Data Stewardship apps can be built using vertically integrated tools like Streamlit

More research needs to be done on the MDM domain to study if the traditional Matching algorithms can be replaced with Neural Network based matching and if rules-based Survivorship can be replaced by ML Ops or RNN based Neural Network with feedback mechanism to continuously evolve based on Data Steward activities.

Another field of study, that needs greater attention is the application Graph Databases and Graph Algorithms on the field of Deduplication and Survivorship.

5. Conclusion

This paper has explored a modern MDM approach which could substitute traditional MDM tools which cannot handle modern data volume across the industry. The modern disruptive technology approach to perform MDM inside the data warehouse will help reduce difficulties associated with traditional MDM. Moreover, this has the potential of reducing overall IT architecture complexity and TCO by eliminating unnecessary complex system. The TCO for the modern architecture has been proven to be 80X -100 X lower in limited pilot deployments in the Higher Education and Automobile Industry.

References

[1] Wu, K., & Chen, J. (2023). Cargo operations of Express Air. Engineering Advances, 3(4), 337–341. https://doi.org/10.26855/ea.2023.08.012

[2] Wu, K. (2023). Creating panoramic images using ORB feature detection and RANSAC-based image alignment. Advances in Computer and Communication, 4(4), 220–224. https://doi.org/10.26855/acc.2023.08.002

[3] Liu, S., Wu, K., Jiang, C. X., Huang, B., & Ma, D. (2023). Financial Time-Series Forecasting: towards synergizing performance and interpretability within a hybrid machine learning approach. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2401.00534

[4] Wu, K., & Chi, K. (2024). Enhanced E-commerce Customer Engagement: A Comprehensive Three-Tiered Recommendation System. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 348-359. https://doi.org/10.60087/jklst.vol2.n2.p359

[5] hasan, M. R. (2024). Revitalizing the Electric Grid: A Machine Learning Paradigm for Ensuring Stability in the U.S.A. Journal of Computer Science and Technology Studies, 6(1), 142-154. https://doi.org/10.32996/jcsts.2024.6.1.15

[6] MD Rokibul Hasan, &Janatul Ferdous. (2024). Dominance of AI and Machine Learning Technique in Hybrid Movie Recommendation System Applying Text-to-number Conversion and Cosine Similarity Approaches. Journal of Computer Science and Technology Studies, 6(1), 94-102. https://doi.org/10.32996/jcsts.2024.6.1.10

[7] Singla, A. (2023). Machine Learning Operations (MLOps): Challenges and Strategies. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 333-340. https://doi.org/10.60087/jklst.vol2.n3.p340

[8] Singla, A., & Chavalmane, S. (2023). Automating Model Deployment: From Training to Production. Journal of

Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 340-347. https://doi.org/10.60087/jklst.vol2.n3.p347

[9] Msekelwa, P. Z. (2023). Beyond The Borders Global Collaboration in Open Distance Education through Virtual Exchanges. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(2), 1-13. https://doi.org/10.60087/jklst.vol2.n2.p12

[10] Msekelwa, P. Z. (2023). DATA DRIVEN PEDAGOGY: LEVERAGING ANALYTICS FOR EFFECTIVE E-LEARNING STRATEGIES. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 1(1), 55-68. https://doi.org/10.60087/jklst.vol1.n.p12

[11] Ahmed, M. T., Islam, M., & Rana, . M. S. . (2023). Climate Change and Environmental Security in Bangladesh: A Gender Perspective . Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 1(1), 18-24. https://doi.org/10.60087/hckggn20

[12] Islam, M., & Rana, M. S. (2023). CONTAMINANT IDENTIFICATION IN WATER BY MICROBIAL BIOSENSORS: A REVIEW. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 1(1), 25-33. https://doi.org/10.60087/jgrkv103

[13] slam, M. (2023). BRIEF REVIEW ON ALGAE BASED BIOFUEL. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 1(1), 46-54. https://doi.org/10.60087/7xz85292

[14] hasan, M. R. (2024). Revitalizing the Electric Grid: A Machine Learning Paradigm for Ensuring Stability in the U.S.A. Journal of Computer Science and Technology Studies, 6(1), 142-154. https://doi.org/10.32996/jcsts.2024.6.1.15

[15] Hasan, M. R. (2023). NetSuite's Next Frontier: Leveraging AI for Business Growth. International Journal of Science, Engineering and Technology, Volume 11 Issue 6. Retrieved from: https://www.ijset.in/volume-11-issue-6/

[16] Singla, A., Sharma, D., & Vashisth, S. (2017). Data connectivity in flights using visible light communication. In 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN) (pp. 71-74). Gurgaon, India. https://doi.org/10.1109/IC3TSN.2017.82844537

[16] Lin, F., et al. (2020). Predicting Remediations for Hardware Failures in Large-Scale Datacenters. In 2020 50th Annual IEEE-IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S)* (pp. 13-16). Valencia, Spain. https://doi.org/10.1109/DSN-S50200.2020.000168

[17] Sullhan, N., & Singh, T. (2007). Blended services & enabling seamless lifestyle. In 2007 International Conference on IP Multimedia Subsystem Architecture and Applications (pp. 1-5). Bangalore, India. https://doi.org/10.1109/IMSAA.2007.45590859

[18]. Building for scale. (n.d.). https://scholar.google.com/citations?view_op=view_citation&hl=en&user=jwVmi8AAAAJ&citation_for_view=jwV-mi8AAAAJ:zYLM7Y9cAGgC10

[19]. Tannga, M. J., Rahman, S., & Hasniati. (2017). COMPARATIVE ANALYSIS OF LEVENSHTEIN DISTANCE ALGORITHM AND JARO WINKLER FOR TEXT DOCUMENT PLAGIARISM DETECTION APPLICATIONS. JTRISTE, 4(2), 44-54. Retrieved from https://jurnal.kharisma.ac.id/jtriste/article/view/29

[20] https://www.informatica.com/content/dam/informatica-com/en/collateral/reference-architecture/mdm-reference-architecture_4660.pdf

[21] Vemuri, N. V. N. (2023). Enhancing Human-Robot Collaboration in Industry 4.0 with AI-driven HRI. *Power System Technology*, *47*(4), 341-358. Doi: https://doi.org/10.52783/pst.196

[22]. Vemuri, N., Thaneeru, N., & Tatikonda, V. M. (2023). Smart Farming Revolution: Harnessing IoT for Enhanced Agricultural Yield and Sustainability. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, 2(2), 143-148. DOI: https://doi.org/10.60087/jklst.vol2.n2.p148*