International Journal of Management, IT & Engineering
Vol. 10 Issue 02, February 2020
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Achieving MDM on Data Lakes

VenkatRamana Rapolu*

Abstract

Master Data Management (MDM) is a Process or methodology to create a Golden copy of your record by integrating Data from different Sources so that this Data can be consumed Enterprise wide; MDM is possible on Transactional Data and Not on agile Data. Since it consumes Data from different Sources and if becomes huge Volume Transational Data and its hard to achieve MDM on RDBMS; Why can't we take advantage by implementing MDM on Lakes ? As last several years many organizations have ingested Data intoHadoop for their Analytic Usecases.

Is it possible to implement MDM on Data Lakes? Do we achieve Significant benefits by Mastering Data on Lakes? How to reduce Data redundancy on Lakes?

Keywords:

Master Data Management (MDM);
Data Quality;
Data Governance;
Data Lakes;
TransactionalDataManagement.

Author correspondence:

Venkat Ramana Rapolu,
MDM, Data Analytics, Data Lakes
Bentonville, USA
Email: rapoluramana@gmail.com

1. Introduction

MDM is a Process with Technology to ensure we have Accurate Data and gives 360 view of the record. As MDM is possible on Transactional data which is Structured, has Meta data or Attributes to Govern. Last several years many Organizations are using Hadoop for their Analytic Usecases.

So how to implement MDM on Data Lakes? Yes its possible to implement MDM on Lakes by Addressing below issues ::

Issue 1 :: What if Data is Non Transactional? Unstructured (Log Data) etc can we implement MDM?

a. No we cannot implement on Unstructured Data like Logs or Non Transactional Data.

Resolution ::

The best practice is apply MDM selectively in data lakes. It's most likely not necessary for every data set. However, there are many situations where, even if the data doesn't seem like the typical MDM fodder (e.g. ERP, CRM etc.) you need to categorize it and match it against master records.

Issue 2::Will MDM Force You Into A "Schema On Write" Situation, Which Slows Us Down?

As Data Lakes main benefit is "schema on read"; So if you are implementing MDM means are we doing "Schema on Write"?

Resolution: In this MDM on Lakes we cannot dump data into lakes; So Data which has Relationship can be ingested into Lakes. If Lakes Data is consumed by Data Scientists thenwe need schema and its meta-data to understand the business meaning and relationship between each data element. On this Metadata Attributes we can create business rules and algorithms.

* Doctorate Program, Linguistics Program Studies, Udayana University Denpasar, Bali-Indonesia (9 pt)

Issue 3 ::Can we Create Master Data Entities From Unstructured Data Like Tweets?

Resolution: As discussed it should be Transactional Data (Not Agile or Unstructured); Certain types of unstructured data are impossible to match and cleanse with MDM processes. So, you can ignore those from an MDM perspective. However, the best practice is to assess the data at the ingestion stage and figure out if the analytics process will benefit from applying MDM.

Issue 4:: Can MDM Systems Even Work With Cloud-Based Data Lakes?

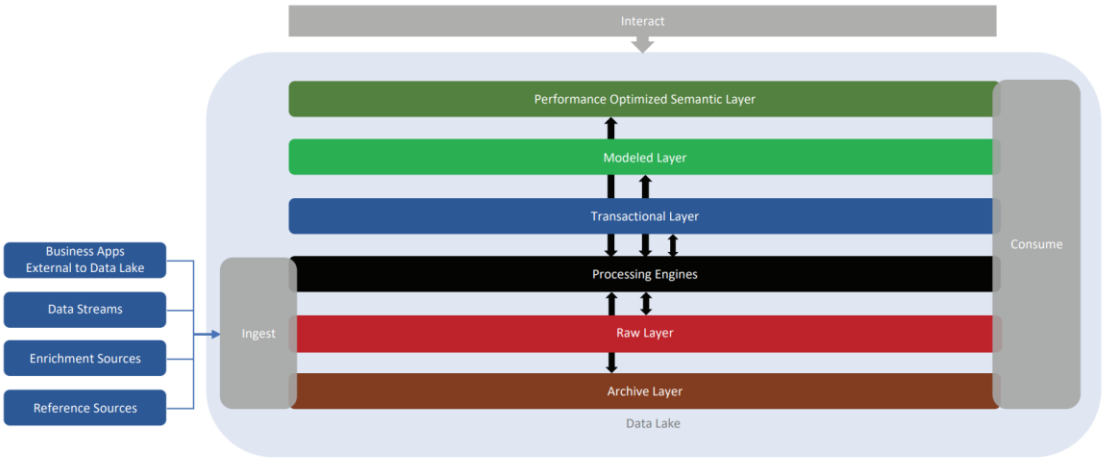
MDM systems were never meant to work with the kind of free range data we have in our data lakes. MDM is primarily used for data warehouses and data marts. It will be a total nightmare to integrate an MDM system with your big data, if it's even possible.

Resolution: If it involves data, there will be a Hadoop solution for it. This applies to MDM. You don't necessarily have to extend an existing enterprise MDM solution into a cloud-based data lake. However, you should look at the possibility of applying the same MDM standards to big data in the cloud as well as in your own data center.

4. Conclusion

Architectingthe Value of a Data Lake

Data lakes can provide flexible ingest methods, flexible storage methods, coexistent transactional and analytical workloads, and performant data accessibility, all of which enable numerous business use cases, agile Multi Domain MDM among them.



ENHANCED MDM IMPLEMENTATION STYLES THROUGH DATA LAKES

Traditional MDM implementation styles include the Registry hub, Coexistence style, Consolidated approach and the Centralized Transactional hub. Although there are pros and cons associated with each approach, the data lake delivers new enhancements to these styles that can extend and improve technology capabilities and business agility.

REGISTRY The output of a Registry MDM approach would be accessible via the semantic layer through the Interact or Consume zones and would potentially source input data from the Modeled, Raw and Archive layers by leveraging data processing methods from the Processing zone.

CONSOLIDATED This MDM and Data lake pairing is essentially an organic Consolidated MDM style in which all data is available, whether copied or native, to a central location. The main value of this pairing is that additional copying is eliminated due to the preexisting coexistence of all the data within the same

ecosystem with minimal processing required due to the 'touch early and touch once' approach described above.

COEXISTENCE The coexistence style, in which the source systems are updated with mastered data, is also greatly simplified through the advantages of having all data organically coexistent without additional effort, but more importantly, the output of the mastering methods need only to store data in the modeled, raw and/ or semantic layers within the data lake itself rather than undergoing additional costly and slow ETL processes to transfer data externally to the data lake.

CENTRALIZED The Centralized, or Transactional, MDM style routes all references or calls to data that is mastered under the authority of the MDM governance program, are made to a single central location potentially through REST calls, SQL queries, etc. And like the other styles, we realize the advantages of all processes sharing and re-using the same data stores and processing engines.

With some or all the transactional and analytical data workloads co-existing in the data lake, any combination of these approaches in a multi MDM domain environment is organic, natural, and to some degree automatic.

References

The main references on MDM implementation on Data Lakes are:
[1] https://blog.centurylink.com/data-lakes-and-master-data-management/
[2] https://www.blue-granite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses
[3] https://blog.stibosystems.com/4-common-master-data-management-implementation-styles