INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Data governance in smart factory: Effective metadata management

Mahendraprabu Sundarraj
maheeprabu@gmail.com
Sun Power Corporation, United States

Rajkamal Mahamuni Natarajan
rajkamalmn@gmail.com
Bazaarvoice Inc., Texas, United States

ABSTRACT

Most of the enterprises recognized the importance of Data governance and started data governance programs either at the Enterprise at the individual Business units level. Data governance councils at various levels in an enterprise define and enforce data quality and security, using policies, standards, and Procedures. The success of Data governance program is heavily relying on people or team who is governing the data governance council. However, Data governance is best implemented by leveraging people, process, and Technology. In the past, subject matter experts of individual data domains maintained the metadata, which is a critical component of data governance. Efficient and automated metadata management, which can be established today by leveraging technology and process not just SMEs, has the potential to mitigate the risk of people dependency. Smart factories, which are heavily automated, generates data at a scale and speed which were never seen before by manufacturing industry, and face the crisis to maintain the data quality, and to make the data available for analytics for any further use. Data generated by smart factories are diverse and are mostly stored in distributed systems, which further increases the complexity of data governance through metadata management. Efficient metadata management, mostly automated, can help smart factories to achieve Data governance goals, and help to provide data as a service. This paper discusses the shortcomings of Hybrid data governance model – acknowledged as the better model for data governance by industry – and proposes a system architecture which has all the benefits of the hybrid model in addition to improvements that are necessary for a smart factory.

Keywords— *Smart Factory, Data Governance, Metadata management*

1. INTRODUCTION

Smart Factory, a result of Industry 4.0, is a heavily automated production facility with highly optimized operating efficiency and resource utilization. Goals of the smart factory are achieved by leveraging the data produced by Edge devices, which are smart devices with native storage and compute. Edge devices produce data at a scale and speed which were never seen before by the industrial world. Though the edge devices, essential data sources in a smart factory setup, physically exist at the factory site, data produced by Edge devices are continuously cleansed, enriched, stored and analyzed at various network levels such as On-Premise network, private enterprise cloud, and public cloud. Availability of Data and accessibility to data are the critical differentiators of industry 4.0 compared to its predecessors. Smart factories are supported by hybrid cloud architecture, comprises of both on-premise and cloud applications. Various Edge device networks, diverse data types (Structured and unstructured data), ever increasing data growth, and Data stores hosted in complex hybrid cloud architecture create challenges in maintaining Data quality, security, and availability in smart factories.

Almost all the factory organizations started the data governance program either at an enterprise or at the individual department level to ensure data quality and availability. Among all the issues Data governance is capable of addressing, significant data issues come from inconsistent data definitions, data types, and inability to trace data source [1]. Metadata – Data about data – can address all inconsistencies in Data definitions and types. Given the complex mix of data sources and Data formats in smart factory, Metadata management is critical, and it is not a one-off activity but a continuous activity. Metadata is usually defined and maintained by Subject Matter Experts (SME) of the data domain. However, advancements such as machine learning and artificial intelligence have opened the possibility of automating metadata management to a great extent.

This paper discusses the importance of Data governance in Smart factory, and strategy and best practices to automate metadata management. The main section of this paper has three sub-sections - literature review of Data governance and automated metadata management, smart factory architecture and associated data issue challenges, and the architecture to automate metadata management in a smart factory context. Much has been written about data governance and the importance of metadata management in general, and Data governance cannot be restricted to only metadata management. However, Effective metadata management plays a significant role in establishing, enforcing, and maintaining the data governance in a smart factory, which has highly distributed IT architecture and relies on data quality for its success not just for compliance.

2. DATA GOVERNANCE AND AUTOMATED METADATA MANAGEMENT

2.1 Data Governance

After the collapse of Enron, which led to Sarbanes-Oxley (SOX) Act, also known as the "Public Company Accounting Reform and Investor Protection Act" or "Corporate and Auditing Accountability, Responsibility, and Transparency Act" [3]. Data governance along with IT governance gained attention. IT controls audit, a result of the SOX Act, made data quality as an essential objective for all organizations. Most of the organizations started the data governance programs to prepare themselves for meeting the IT compliance standards set by the SOX Act.

Data Quality can be defined by data correctness, completeness, and consistency across all databases, and business rules conformance [2]. Data quality was not just a legal requirement (as expected in SOX), but also essential to ensure to convert data into useful insights. Data governance is defined, enforced, and monitored by Data governance councils. Data governance councils consist of three types of members- Data Trustee, Data steward, and Data custodian [1]. Figure 1 shows the responsibility of each of the member type.

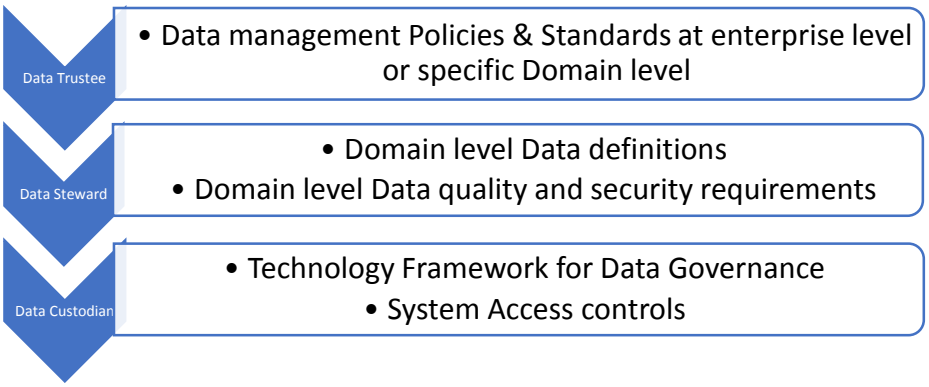

Fig. 1: Data governance council member type and responsibility

Data governance is not just a group of people, but also rules, responsibilities, and enforcement methods [4] to achieve Data quality. Data governance is defined, enforced, and monitored by Data governance councils. Based on the types of data governance decisions need to be taken by Data governance bodies, data governance is either federated or centralized [7]. In some cases, Data governance is maintained within the silo data domains without any relationship with other data domains in the enterprise. Three types of Data governance models are given below.

- Silo Data governance
- Centralized Data governance
- Federated Data governance

2.1.1 Silo data governance: Data governance policies are defined as specific to a data domain with the help of domain-specific data stewards or experts. Keeping Data governance at individual data domain level is mostly the first step towards achieving enterprise-level data governance. In rare cases, Silo data governance continue to exist even after establishing centralized data governance if the data domain is different from other data domains at the enterprise level, and a significant amount of data governance decisions unique to the data domain need to be made.

2.1.2 Centralized data governance: Centralized data governance body defines the data governance standards and enforces on all other data domains that exist in the organization. Since a single entity makes much of the data governance decisions, it brings in much predictability and expedites the improvement in data quality. However, centralized data governance may not cause significant improvements if the data domains of the organizations are highly diverse and it is challenging to define standard data compliance requirements.

2.1.3 Federated data governance: Federated data governance provides balanced model between Centralized and Silo data governance models. Federated data governance model has data governance bodies both at center and individual data domains. Data compliance requirements are classified into two categories – Enterprise level requirements and Data domain specific requirements. Enterprise level data compliance decisions are made at center and individual domain specific decisions are taken at individual data domain level.

All the above data governance models heavily rely on executive level sponsorship. Automation of enforcing and monitoring data governance can significantly increase the success of data governance program in an enterprise.

2.2 Automated Metadata Management

If Metadata, which describes the data, is persisted and managed accurately, Data quality will be improved significantly [5]. Metadata helps to cleanse the data at data generation itself, increase query execution speed, and help much faster data discovery. Sample metadata schema, not a complete one, is shown below describes the data from multiple systems or domains such as ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), HRMS (Human Resource Management System) and PLM (Product Lifecycle Management).

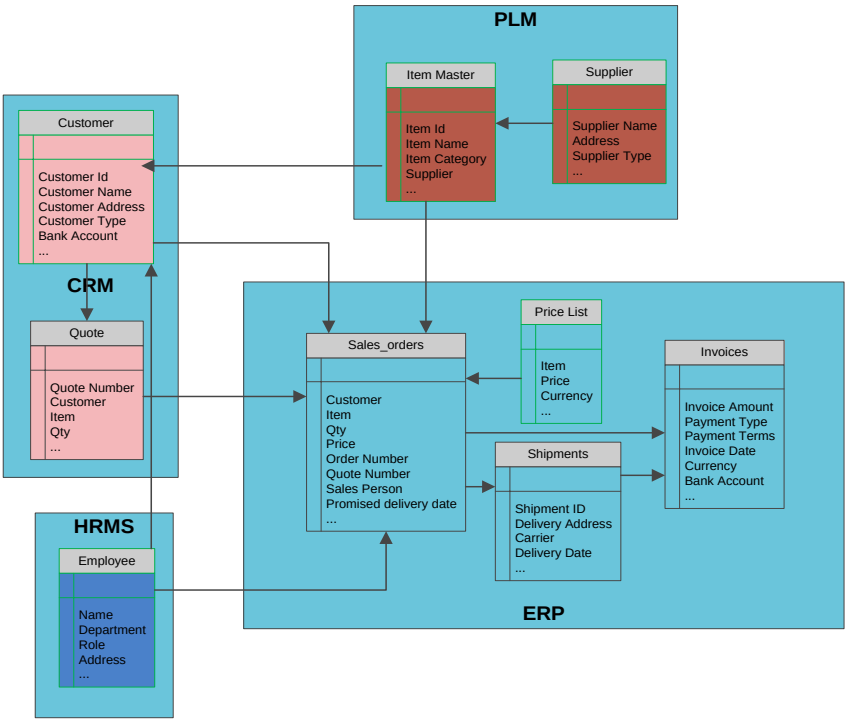

Fig. 2: Sample subset of the enterprise data schema

As shown above, metadata not only describes the data sets but also contains the relationship between these data sets, which is essential to understand data lineage. Given the domain expertise needed to understand each data set from a specific domain or system (ERP, CRM, HRMS or PLM), Subject matter experts of each domain or Data stewards were responsible for metadata management. Multiple data stewards collaborate and manage the metadata at the enterprise level through data governance councils. However, when organizations face the exponential data growth (as it happens in IoT environment), and data being continuously transformed, mined, and persisted in multiple formats, dependency on people (SMEs) to manage the metadata will not be a sustainable process. Automating metadata becomes the necessity.

Data governance can be simplified into four steps [1]:

(a) Form the team and define RACI (Responsible, Accountable, Consulted, and Informed)
(b) Identify Data Domains
(c) Identify Critical Data elements
(d) Identify control and monitoring

Out of the above four steps, three of them (Step 2, 3, and 4) are dependent on the metadata repository.

In general, Metadata repository stores metadata from various sources such as Databases, Documents, APIs or web services, Web applications, and third party (supplier, customer, and trading partner)[6].

3. SMART FACTORY IT ARCHITECTURE

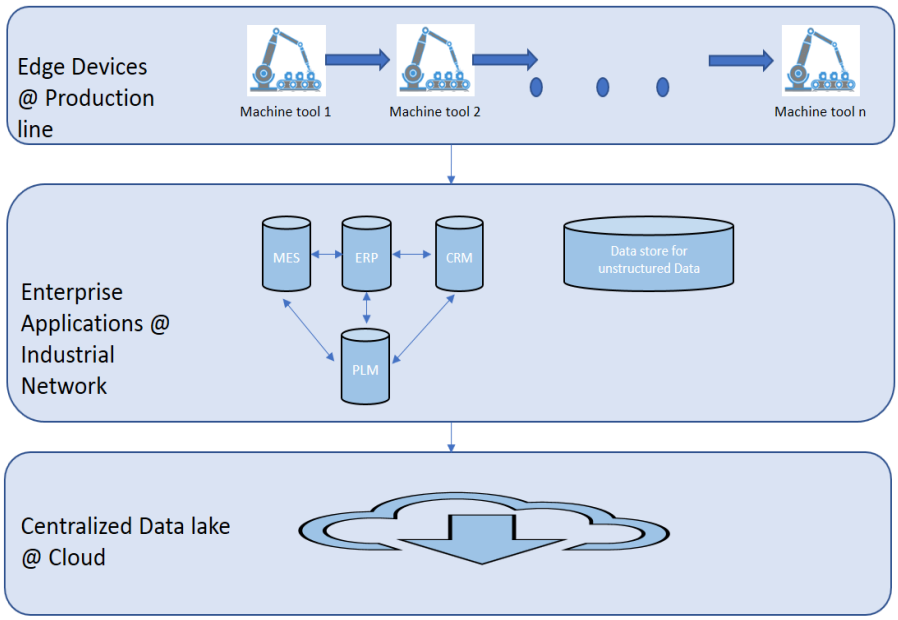

Fig. 3: Simplifies architecture of smart factory and data flow

3.1 Edge Devices @Production line

IoT revolution has introduced smart devices with native storage and compute. Most of the new-age Barcode scanners, RFID scanners, and cameras come with native memory and compute. Vendors of these smart devices define the format and structure of the data generated by these devices. Not all the smart devices in each production line come from the same vendor, leading to diversity in data formats and lack of standardization. [10]

3.2 Enterprise applications @Industrial network

Business applications such as MES (Manufacturing Execution System), ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), and PLM (Product Lifecycle Management) uses machine-generated data. These business applications are purpose-built on multiple technology platforms. Given the various business use cases and technology differences between these business applications, machine-generated data is divided and stored in different formats and structures in all the applications mentioned above — not a single business application stores all the machine data generated in the smart factory. Data from multiple systems must be merged in a meaningful way, and analytics such as Operation reports and Dashboards (Data visualization) must be created to get a complete end to end data visibility. Apart from the structured data stored in business applications, there are data storage solutions at the industrial network level to store an enormous amount of unstructured data such as images and unstructured machine logs generated in a smart factory environment. Factory organizations analyze the unstructured data frequently to improve continuously.

3.3 Centralized Data Lake @Cloud

Given the diversity of Data and distributed storage, Factory organizations find it challenging to get any meaningful insights out of all data combined. Before the cloud, Organizations used to undertake costly and un-scalable data warehousing projects to combine data from different business applications and find insights to make data-driven decisions. Cloud provides almost unlimited computation capacity and cheap storage to host and analyze all data an organization generates and helps organizations to create data lakes. Data lakes, unlike data warehouses, allow storing any data format in its original form. Distributed query processing, APIs, and numerous server-less features available in the cloud make data retrieval from Data Lake an easy task. However, allowing the data to be stored in any format in a distributed manner makes data governance complex. Ensuring the Data quality through governance on business applications is a quantifiable task since there is an upper limit on data storage and compute of these silo business applications. However enforcing the data quality on a highly distributed environment such as a smart factory, which is built on IoT and cloud services, is an ever-changing task. Hence purpose-built tools are needed to handle data governance in the smart factory.

4. PROPOSED ARCHITECTURE - EFFECTIVE METADATA MANAGEMENT IN THE SMART FACTORY

Automated metadata management for the streaming data helps to measure run-time data quality, cleanse the data in run-time, and help to improve the data retrieval rate significantly.

As explained above in the "Smart Factory Architecture" section, smart edge devices produce data in various formats and metadata for each type of device is unique. Hence these devices are valuable data sources for metadata repository.
There are three popular metadata management models. [8] All the models have one common goal of giving a single access point to all data consumers.

(a) Centralized model: Centralized metadata repository with batch synchronization with data sources
(b) Distributed model: metadata is left with data sources; Real-time access to metadata is given via a metadata engine.
(c) Hybrid (Combination of Centralized and distributed models): Built to leverage strengths of the above models and solve the weaknesses of the above models.
(d) Hybrid architecture still has few areas to be improved further. Below are the areas to improve and how the proposed architecture helps is written Question and answer format.

- If consumers face an error due to wrong metadata, how the metadata repository corrects automatically, reducing the need for manual intervention?

The proposed architecture provides a process to auto-correct inconsistencies in metadata in almost real-time, reducing the manual intervention significantly.

- As per the hybrid model, the availability of metadata in real-time is dependent on the availability of data sources (given real-time access). Real-time access to metadata stored at data sources puts a computing load on data sources. In smart factories, Edge computing usually is thin and cannot be exposed to serve random real-time metadata queries. Performance is a concern here.

Proposed architecture proposes event-driven metadata push from edge devices rather than exposing the edge network for real-time metadata access. Event-driven data push is a low-latency data replication method, and it is a sustainable way to provide close to real-time metadata access.

- Enforcing the enterprise-level data governance policies from metadata repository is an opportunity, not often highlighted as an advantage.

Proposed architecture includes a set of APIs to enforce enterprise-level data governance or quality policies at data sources level. Small to medium level data cleanup rules can be easily automated, leaving only transformational changes to be done in a manual route.

Below are the improvements which were done over hybrid architecture in the proposed architecture.

(a) Auto correction of metadata inconsistencies
(b) Event-driven metadata push – Sustainable way to get close to real-time metadata access
(c) Ability to enforce enterprise metadata governance policies to distributed systems

Figure 4 explains the proposed system architecture to automate the metadata management in a smart factory environment.

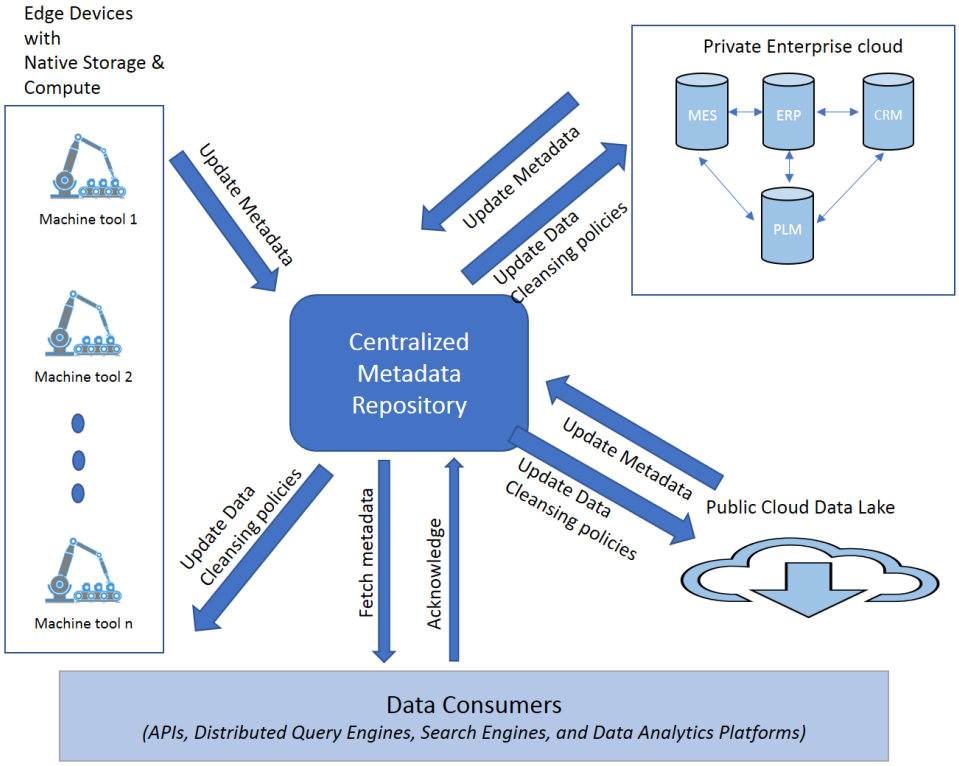

Fig. 4: Proposed Architecture for automated metadata management

The critical component of the above architecture is CMR (Centralized Metadata Repository) which consolidates and persists all the metadata of specific data domains. CMR is not just a data store. It has two subcomponents.

(a) Data store: To store metadata and revisions
(b) Application Programming Interfaces (APIs): To collect the metadata, to maintain the metadata quality, and to make metadata available to metadata consumers in a secured manner.

Data consumers are typically APIs (web services), Distributed query engines, search engines, and Data analytics platforms. They are registered and recognized by CMR to avoid data leaks within the enterprise and outside the enterprise.
Data sources which are shown in the above diagram interact with Centralized Metadata Repository (CMR) through two types of interfaces – "Update Metadata" and "Update Cleansing Policies". These interfaces help to automate metadata management.

(c) Update metadata: Data source updates the metadata (metadata of the data it stores or generates) to CMR frequently. Metadata can be changed due to a variety of reasons such as new Data source addition, upgrade on existing edge devices or business applications, and business-driven enhancements on data sources. Automating the "Update metadata" interface helps to keep CMR as accurate as possible, and it is an essential first step in automating the metadata management.

Below diagram explains the automated process flow following the "Update Metadata interface" and is followed by the process flow description.

- Data source updates the CMR if any change in metadata
- CMR receives the metadata update and creates a new revision. CMR maintains all the revisions of metadata of each data source. Maintaining the revisions of metadata helps to prepare compliance reports and to revert to previous versions in case of erroneous changes in the latest version.
- Data consumer platforms fetch metadata from CMR frequently
- Data consumers optionally send the response signal – success or failure – to CMR.
- Response received from consumers is used to validate the accuracy of metadata and to resolve the inconsistencies automatically.

(d) **Update Cleansing policies:** CMR is the data source for data cleansing policies, and update on those policies should be fed to relevant data sources. Update on these policies is needed for the below reasons.

- To fix the root cause of the recurring data compliance issue at data sources.
- To reflect the new update on enterprise-level data governance policies.
- "Update on cleansing policies" interface can be scheduled to run in batch mode or on-demand when there is an update on enterprise-level data governance policies.

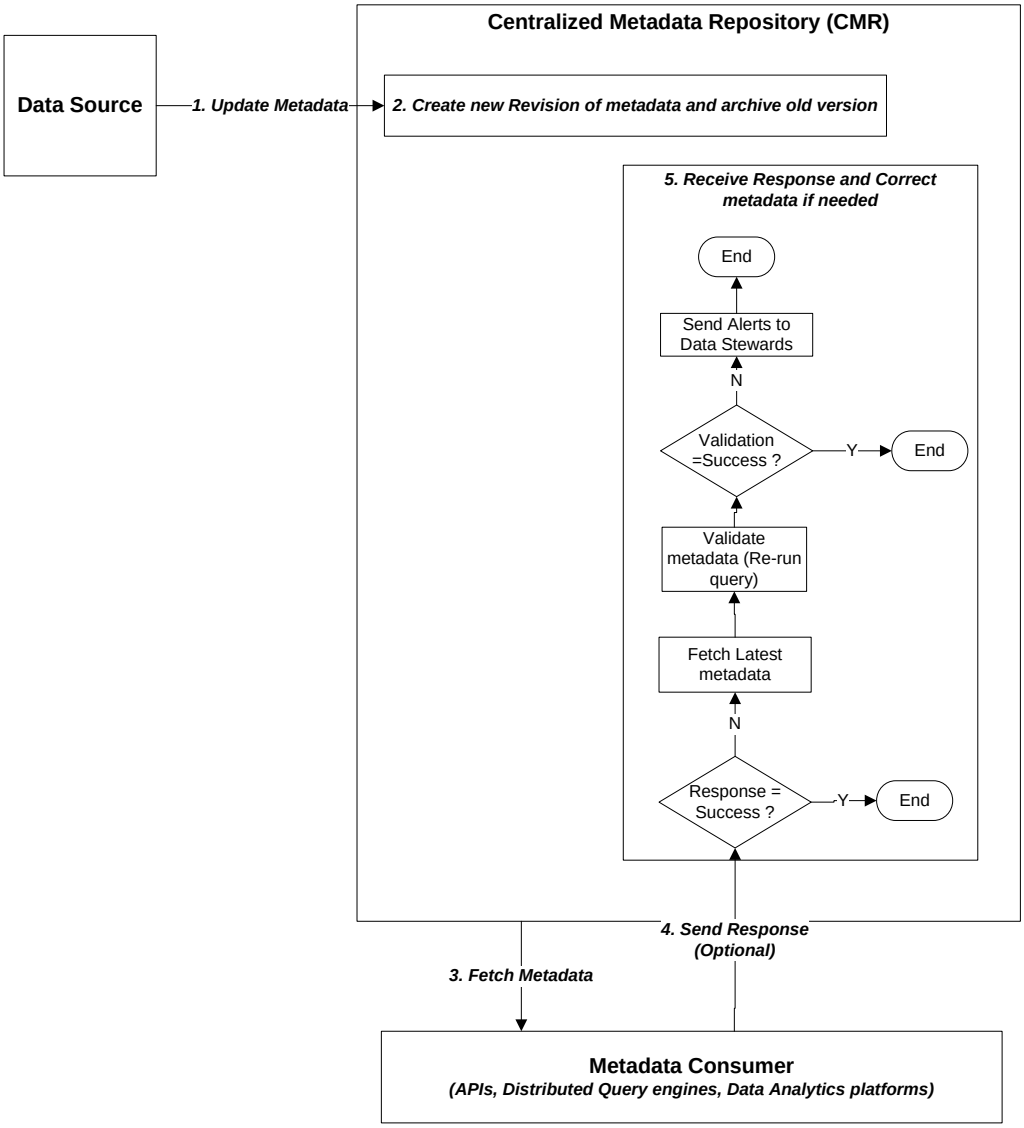

Fig. 4: Logic for "Update metadata Interface" process flow

Data consumers, leveraging CMR architecture explained above, fetches the data from distributed data systems (edge, private and public cloud systems) as if they query from one database [9]. Proposed architecture not only improves the data consumer's experience but also provides faster query performance. Since the data is left to stay in the original form in the data lake, effort and time to prepare data are significantly reduced, enabling low-latency architecture.

5. RESULT

A prototype was created based on the proposed architecture for a single-factory organization and results were recorded.

Centralized metadata repository did not only help to improve the metadata accuracy by 50% but also indirectly helped to reduce the edge layer compute load, which used to be a concern due to real-time metadata fetches unpredictable queries from data consumers.

Time to enforce enterprise level cleansing policies was used to be more than 90 days. After implementing the new model, enterprise-level data cleansing policies were enabled in less than a week. New policy or code changes are reviewed and approved by the Change approval board, which consists of data stewards, custodians, and owners.

6. CONCLUSION

Based on the results recorded, the proposed architecture was found to be ideal for industrial manufacturing organizations. Below are the key takeaways.

- Federated data governance is suitable for distributed systems than any other model.
- Microservices based architecture (used in CMR APIs) brings in adaptability and scalability
- Decoupled data integration patterns (between data generators and consumers) enables Low-latency architecture
- Monitoring helps to keep the cost lower and issues under control.

Every organization's IT landscape and their goals are different. Hence the proposed architecture should be customized according to the needs of an organization.

7. REFERENCES

[1] https://www.dataversity.net/data-governance-demystified-lessons-from-the-trenches/

[2] Bair, J 2004. Practical Data Quality: Sophistication Levels, viewed 25 Mar 2006
[3] https://en.wikipedia.org/wiki/Sarbanes%E2%80%93Oxley_Act
[4] http://www.datagovernance.com/the-basic-information/
[5] Marco, D. (2000). Building and Managing the Metadata Repository: A Full Lifecycle Guide. Wiley. ISBN 978-0471355236.
[6] David Marco and Michael Jennings (2004). Universal Meta Data Models. Wiley. ISBN 9780764571596
[7] http://www.datagovernance.com/choosing-governance-models/
[8] https://www.dag.com/hubfs/Metadata_Architectures_Whitepaper.pdf
[9] http://www.datavirtualizationblog.com/3-stage-evolution-data-federation/
[10] https://www.iiconsortium.org/pdf/Introduction_to_Edge_Computing_in_IIoT_2018-06-18.pdf

