Data Governance in Data Mesh Infrastructures: the Saxo Bank Case Study

Divya Joshi, Sheetal Pratik and Madhu Podila

December 4, 2021

Joshi, Divya. Pratik, Sheetal. & Rao, MP. (2021). Data Governance in Data Mesh Infrastructures: The Saxo Bank Case Study. In *Proceedings of The 21st International Conference on Electronic Business* (pp. xxx-xxx). ICEB'21, Nanjing, China, December 3-7, 2021.

Data Governance in Data Mesh Infrastructures: The Saxo Bank Case Study

(Enter one for initial submission: Full Paper)

Divya Joshi*, Thoughtworks Technologies, Gurgaon, India, divyajoshi1001@yahoo.com
Sheetal Pratik, Saxo Bank, Gurgaon, India, sheetal.pratik@gmail.com
Madhu Podila Rao, Thoughtworks Technologies, Hyderabad, India, pmsrao@gmail.com

ABSTRACT

Data governance (DG) is the management of data in a manner that the value of data is maximised and data related risks are minimised. Three aspects of DG are data catalogue, data quality, and data ownership and these aim to provide transparency, foster trust, and manage access and control the data. DG solution involves change management and alignment of incentives and mere technology is not enough to address this. In this paper we aim to provide a holistic view of data governance that is a synthesis of academic and practitioner viewpoints and conclude by giving an example of a pilot case study (Saxo Bank) where authors worked on tech and cultural interventions to address the data governance challenges.

Keywords: Data related risks, Trust in data, Change management and alignment of incentives, Governance

*Corresponding author

INTRODUCTION

The modern organisations have evolved and are marked by some salient characteristics such as partner ecosystem, advent of platforms that require seamless exchange of data with other organisations, and desire to innovate. This requires flexibility and easier access to data. Moreover, the regulatory requirements are complex and tightening giving rise to closer control of data, and obligation to protect the security and privacy of the data subjects. Besides, data is growing humongously and may be lodged in disparate systems in the organisations. This can lead to silos resulting in obstacles to compliance and poor decision making.

If the data governance is lax, there are risks of data breach and regulatory violation that can result in financial implications via lawsuits and regulatory fines. IBM estimates the global average cost of a data breach as $4.24M and the average cost per lost or stolen record in a data breach as $148 M (Ponemon, 2018). As per Dharni (2018) Google ended up paying a fine of £50 million to the EU, Uber $148 M, and Marriott International £110 to regulators for non-compliance.

Bischoff (2019) states that organisations not only pay financially for data breach- but it's also a huge brand hit and loss of reputation. A study of companies trading on the NASDAQ found that the share price of the companies with a data breach was an average of 13% lower than the Index three years after the incident. On the other hand, too tight data governance stifles productivity and innovation. Hence Data governance is extremely important in the right flavor- not too lax and not too tight.

Furthermore, the data landscape is changing- data is moving to the cloud and older relational structures are not able to keep up. Organisational structures are evolving- more and more business teams have analysts, and the core data team is confined to data infra. Then, different functions in the organisation are using data to different means and the core data team is not able to satisfy the demand of switching business contexts for which they have no expertise. This leads to data latency and business is looking to lower tech barriers so that they have quicker access to data for decision making.

That is where the concept of Data Mesh comes in. Dehghani (2020) defines that Data mesh is a decentralized socio-technical approach to manage analytical data at scale. However, this demand for data democratisation puts pressure on centralised governance. If data were to be democratised, governance must be federated.

PROPOSAL

How do we intend to approach the problem?

For a long time, governance has been a centralised and a tightly controlled function. If this were to be federated, there needs to be a specific plan rooted in the principles of DATSIS: discoverable, addressable, trustworthy, self-describing, inter-operable & secure data, as described by Balnojan (2019). Data governance must keep a delicate balance with just enough governance to avoid the loopholes but encourage productivity and address demands of decentralisation. Good federated data governance can answer security and governance vows by providing efficient and transparent controls over data that the regulatory environment obligates and also support federated delivery by making data visible and trustworthy. This can facilitate decentralised data ownership.

Federated Data Governance in data mesh infrastructure

To inspire trust in data, the data governance strategy should address three key aspects: discoverability, security, and accountability. This can be done by putting in place federated data governance based on data catalog and data quality framework. The idea is that the organisations should understand the data they hold.

Data ownership is another key part of this equation. This reduces silos and need for human touch points. We envisage that federated data governance should consist of below elements.

Data Catalogue (DC)

Data catalogs are descriptions of data and serve as an inventory of metadata giving users the information necessary to evaluate data accessibility, health, and location. With self-service business intelligence, data catalog is a powerful tool for data management and data governance. Catalogues assist in locating relevant data and charting clear data representation hence help a user decide how she can use data. The constituents of DC are technical metadata, ownership information, data lineage, and Business Glossary. Technical metadata describes how the data is organized and the structure of the data objects such as tables, events, objects, attributes with their types and lengths, indexes, and connections. Ownership Information captures the relationship and origin of data. Data lineage (DL) facilitates distributed discovery- one can get the right information at the right time and draw connections between data assets with lineage. Business Glossary (BG) establishes common understanding of business terms in the organization hence prevents misunderstanding by misinterpretation.

Data Quality (DQ)

While there are divergent views, we think DQ is a subset of Data governance. Self-service DQ includes ability to define and implement quality rules. A gap in DQ leads to lack of credibility of data and loss of business opportunities owing to Dark data assets. In this model, end users are empowered to export, report, and edit the quality rules. Net-net, people who manage data actually understand the data they manage, hence breaking the silos.

Data Ownership (DO)

The federated data governance requires change management and alignment of incentives and mere technology is not enough to address this. Hence data ownership is a necessary requirement. This facilitates systemic change and helps create data culture, enabled by personas such as Data Steward, Data Product Manager, and Data Domain Owner.

Although Data Mesh has become a popular concept in the industry, there is only one scientific paper “Data-Driven Information Systems: The Data Mesh Paradigm Shift” (Machado et al, 2021), and no contributions in the scientific community on Data Governance in Data Mesh Infrastructure. This contribution is important to build a framework for this emerging concept and this will allow organizations and practitioners to design and implement federated data governance within the Data Mesh paradigm.

POTENTIAL IMPACT AND CHALLENGES: FEDERATED DATA GOVERNANCE

Maximizing the Impact: What are the business cases?

Some very clear business cases that emerge for federated data governance: expansion of partner ecosystem (such as open banking, marketplaces, and many more), regulatory requirements and compliance, innovation (ability to experiment by reducing time to market, monetization of data assets), improving user experience, reducing the dependence on core data team hence lowering the tech barriers in the organisation.

Challenges: Why is this so Hard?

For attaining intelligent federated data governance, an organisation must solve some crucial challenges that are a mix of technical and social, making it so hard. Firstly, agreement and compliance to the governance in a dynamic policy environment is a moving target. Then, getting business and technology on the same page is very important because the need for self-serve tools and reorganisation is best driven by business, not tech. Furthermore, seamless integration of tooling with the existing data ecosystem is tough. In our experience, available tools may not map exactly to the organisational needs. Selecting the right tool befitting the unique needs and the seamless integration become extremely important. Then creating data culture and change management is a humongous task. Culture and change management are usually the hardest part in any equation. It requires partnership at multiple levels and buy in from the top management. A mesh of structural, procedural, and relational interventions can be a potent remedy.

DEMONSTRATION CASE: OUR IMPLEMENTATION OF DATA GOVERNANCE: SAXO BANK DATA WORKBENCH PLATFORM

Problem Statement/ Business problem

At Saxo Bank, the focus is to democratize trading and empower clients with information and agility to act. Thus, good quality and real-time data is imperative to facilitate clients, while managing risk and compliance. Whilst enterprise-wide adoption of Kafka helped this, the need to govern, catalog, improve, and publish the data quality transparently became more acute. This made Saxo an ideal case study. The hypothesis was that the sound data governance will:

First, enhance data quality by providing data governance, data cataloging, & data quality management in the platform as-a-service. This should be measurable via improvement in data quality measure for data assets under governance.
Second, reduce time to market by ensuring that data assets are discoverable and searchable, cataloguing data assets, enabling transparency, and providing easy to use self-service tools to domain teams.
Third, facilitate integrations by following internationally agreed standards for data assets. So that it is possible to avoid integration issues by expanding schema to associate the data products and data elements with industry standard business ontologies. Fourth, ensure 'Saxo as first WLC (White Label Customer) principles' in all operations.

As is state analysis and Stakeholder Engagement
The journey started with understanding the vision and the key drivers that led to data governance implementation in the organization. Next step was to assess the current state before diving into any solutioning. Saxo had Kafka as the authoritative data source platform. The solution needed to enable governance for the data products published on Kafka. Conceptual diagram of 'As is state' is as following:

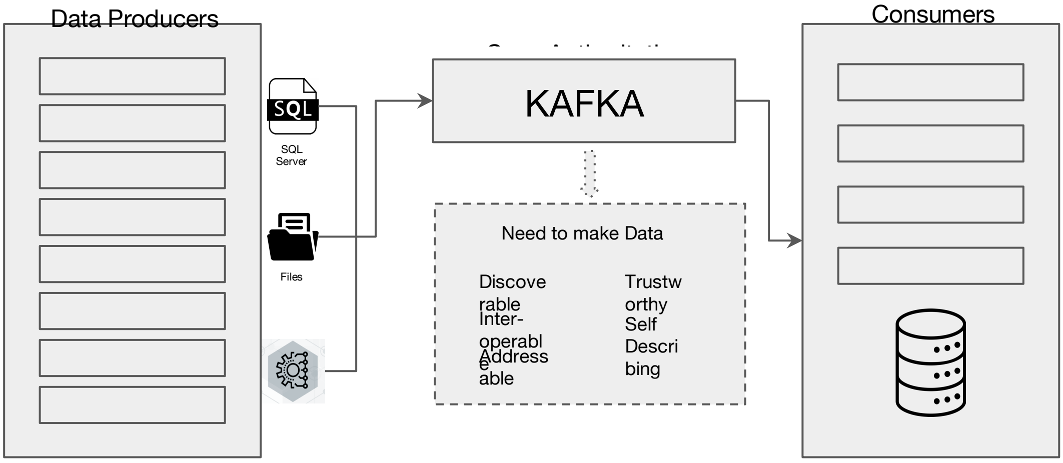

Figure 1: Conceptual Diagram of As is state

This understanding helped create the functional features roadmap. This was supplemented by workshops with key stakeholders and persona Interviews with different data stakeholders to document the pain points and goals. The resulting feature matrix contained expectations from the data governance solution with priority and criticality of the features.

Priorities for Solutioning Principles
Business, process, technical, and financial priorities were considered. Business priorities included world class Saxo experience, packaging and selling asset management products, introducing money making logic and subscription models to the clients, Industrializing the Saxo solutions wholesale offering. Process wise, federated governance and self-serve platform were the priority. Technical priorities were alignment with Data Mesh and current kafka implementation. Financial priorities were assessing total cost of ownership and reducing operation cost.

Solution Recommendation
Based on the above factors, open-source LinkedIn Data hub for metadata and GreatExpectation for Data Quality were shortlisted. The conceptual framework looked

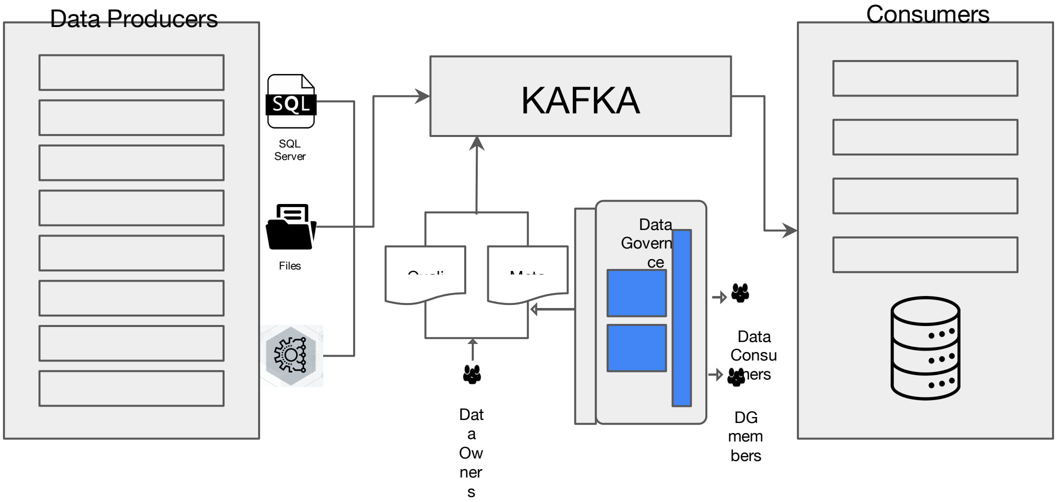

Figure 2: Conceptual Diagram of the Proposed Solution

Tool Selection Approach
The selection needed to be contextual to the organization, data mesh, domain ownership, and light touch governance framework. Below are the details of the steps of the process, hence a feature matrix reflecting the gap analysis in the existing data ecosystem, stakeholder priorities, and feature expectations from the tool was created. Then, an analysis of in-house vs COTS vs open source was done. For commercial tools, evaluation of organizational maturity is important to get a clarity on what out of a long list of COTS features will be used. Other important parameters were flexibility in customization, cost of supporting the tool, and deployment options. In the case of open-source tools, first consideration was the current features supported, then the community behind the tool, roadmap visibility, architecture, and extensibility of the tools were evaluated.
An extended analysis on the shortlisted tools was done by performing the spikes (for open-source tools). Below were the tools shortlisted and analysed:

Data Catalog: Collibra, Alation, Data.world, Zeenea, Marquez, Amundsen, and Great Expectations
Data Quality Informatica IDQ, Trillium DQ, AWS Deequ, Apache Griffin, LinedIn Data Hub

Implementation Approach
A two parts approach was taken: Installation & configuration of the tools in the existing ecosystem and implementation of the Data Catalog and Data Quality Solution.

For integration of the tool, one important consideration was that Saxo was the first organisation to implement Linkedin DataHub in the Kubernetes environment thus the open-source libraries were extended to deploy in the Kubernetes cluster.
Furthermore, the DevOps pipelines for deployment of the tool were created with automation- for a self-serve capability. This was a step to leverage the product upgrades with minimal effort. Then, a forking/synching strategy was created to keep up to date with new features. Finally, the observability was implemented by exposing the logs and metrics from all the services to ELK platform and the alerts for failures/errors/warnings were configured.

For implementation of Data Catalog and Data Quality solution the key principle of push-based metadata acquisition was followed, in-line with the data mesh principles. i.e., The domain teams are responsible for publishing the metadata (armed with necessary self-serve capabilities) and they know their data better. The open-source tools were extended for containerization and the same was installed & configured in the Azure AKS environment. Metadata onboarding efforts were minimized because, otherwise, bringing the datasets into the Kafka platform requires an on-boarding process. The solution extended the onboarding process to collect additional information (such as dataset ownership etc.), reducing the efforts to onboard metadata to minimum. The integration layer was created to extract the metadata from the onboarding templates/forms, the schema/structure was extracted from the Avro/Proto schema definitions and transformed the same into the expected format/structure of the product to fill the data catalog.

Data Catalog
Implementation of the data catalog included bringing the technical metadata, ownership, lineage, and business glossary, pushing the metadata repository through defined API/interfaces, and making the same available through discovery interface.
For Technical Metadata, the dataset information was extracted from the dataset (topic) definition files and the structure/definition of the data elements from the associated schemas and pushed to the metadata repository via product API/interface. Then the

templates were enriched with ownership information and the same were extracted and associated with the dataset in the metadata model.

LDH provided the support for data lineage. Dataset level lineage was implemented to help visualize/navigate the lineage of a dataset with the parent and with the derived datasets. The lineage was re-worked to foster better representation by a combination of extending the onboarding templates and discovering the Kafka DataHub connector configuration, and a mix of user provided and discovered information was used for building the upstream relationships and the same was pushed to the metadata repository.

Business Glossary

Implementation of the Business Glossary presented unique issues: Linkedin DataHub didn't support the business glossary. Saxo team was keen on getting business terms as the first-class citizen to enhance the metadata and to resolve data quality issues that arise due to inconsistent nomenclature. The kafka onboarding already had FIBO definitions for the data elements while onboarding the schema. Building on the existing Kafka schema onboarding process, design of the business glossary was proposed. Next step was to Introduce and onboard the new entity into the LinkedinDataHub entities for business glossary).

First, new entities were designed and implemented to model business glossary (glossary nodes, terms, and relationships) in the open-source product. Saxo uses the proto format to define the schema and it is a rich source of business concepts/definitions. Solution enhanced the proto options to capture additional metadata in the schema (proto definition files) that in turn helped enrich the glossary. This enabled users to discover the datasets through business terms/concepts and discover the relationship b/n terms.

Data Quality

For Data Quality implementation, the DSL was defined to facilitate the data producers to create the quality rules and pushed these to the respective domain repository. Pipelines were built to provision the automated jobs based on the onboarded rule definitions to perform the data quality checks and emit the quality results to designated Kafka Topic.Quality results/metrics are collected from Kafka topics and then pushed to metrics stores (elasticsearch) and the same metrics are made available to producers/consumers through Kibana dashboard. This allowed domain/product teams to focus on documenting the quality rules and not worry about the building of the data quality jobs. Dashboards/metrics were provided to empower users.

Data Domain Ownership

Saxo Bank is committed to assimilating data governance across the organization. Milestones in this journey included formulation of DG committee, defining roles and responsibilities, and creation of domain ownership model (domain teams are responsible for owning and managing the data). The responsibility started from documenting the data products, classifying the data appropriately, and enabling publishing the data products via kafka to the consumers. Domain teams now own their respective repositories that manage the life cycle of the data. Each data product contains the ownership definitions (both business and technical). While onboarding the data products, teams define the schemas, relate it with the business terms, and classify the dataset attributes.

Solution Outcomes

As Consumer, user can discover the data products, view the metadata, view data lineage to understand the data flow better, view quality snapshot along with metadata and able to access to quality notifications/alerts for the datasets of interest

As a data product owner, a user can link data products/elements to the business terms, that encourages the consumption and avoids data inconsistencies, define information classification at data element level, and profile the data before onboarding to Kafka.

Net-net, Data Workbench platform of Saxo enables the domain teams to publish transparent and trustworthy data for the consumers and will pave the way to onboarding of new partners with increased confidence. This is a big milestone in the journey of open banking.

CONCLUSION

In our view good federated data governance is critical because poor practices in controls have significant negative impacts on stakeholders and wider society and on the other hand, excessive controls stifle productivity. Federated data governance inspired by data mesh principles is scalable, self-served, and self-owned. Clear ownership gives better control on quality.

We think the issue merits a broader view than just technology and is cross-cutting, involving organisational incentives. This presents a unique opportunity to use technology and ways of working to quantify and drive out risks and bias in data driven solutions. Data mesh is an answer to these. Overall, good solutions require diverse multi-disciplinary teams, tools that teams can use autonomously, enabling scalable governance models and tooling. The benefits are safe, data-driven innovation at scale, better controls and compliances, and better outcomes for individuals and society.

ACKNOWLEDGMENT

We are grateful to Michael Cusumano, Sloan Management Review Distinguished Professor and Deputy Dean at the MIT Sloan School of Management, Joseph Doyle, Erwin H. Schell Professor of Management and Applied Economics, MIT Sloan School of Management, Maribel Yasmina Santos, University of Minho, Centro Algoritmi Guimarães, Portugal, Allan R. Tate, Executive

Chair, MIT Sloan CIO Symposium, Alison Gerlach, Managing Partner, Breakthrough Ventures, Bob Lonadier, Senior Product Manager, VMWare, Martin Fowler, Chief Scientist, thoughtworks, Smita Joshi, Lead consultant, thoughtworks for their support.

REFERENCES

Bischoff, P. (2019). How data breaches affect stock market share prices. Retrieved from https://www.comparitech.com/blog/information-security/data-breach-share-price-analysis/. (Accessed on Oct 25, 2021).

Sven Balnojan: Data Mesh Applied, Retrieved from towardsdatascience.com, https://towardsdatascience.com/data-mesh-applied-21bed87876f2, (Accessed on Oct 25, 2021)

Dharni, A. D. S. (2020). Data privacy compliance using COBIT 2019 and development of MISAM audit caselet. Concordia University of Edmonton. pp.13-14. Edmonton, Alberta.

Dehghani, Z. (2020). Data Mesh Principles and Logical Architecture, Retrieved from https://martinfowler.com/articles/data-mesh-principles.html. (Accessed on Oct 25, 2021).

Machado, I., Costa, C., & Santos, M. Y. (2021). Data-Driven Information Systems: The Data Mesh Paradigm Shift. 29th International Conference on Information Systems Development (ISD2021 Valencia. Spain).

Ponemon, L. (2018). Cost of a data breach study: global overview. Benchmark research sponsored by IBM Security independently conducted by Ponemon Institute LLC. Retrieved from https://www.ibm.com/downloads/cas/OJDVQGRY (Accessed on Oct 25, 2021).