Data Strategy for a Business Domain: An Application of Data Mesh

Gopal "Sharath" Sharathchandra
SVP, Financial Solutions | Feb 16, 2023

BACKGROUND

There have been various articles about enterprise data strategy. In this white paper, I discuss the data strategy for a business domain within the enterprise. By "business domain", I refer to a business area within an enterprise which could either be a line of business (LOB) having a P&L or a shared service such as finance, HR, IT or modeling and analytics. To save space, I will use the terms "domain" and "business domain" interchangeably. While the discussion centers around the modeling and analytics (MA) domain and is based on my experience as a domain Chief Data Officer (CDO) in the financial services industry, the key points carry over to any business domain, across industries, that uses data and analytics in driving its business goals (e.g., LOBs, finance, marketing, sales, HR).

I view the data strategy outlined here as best practice based on my experience and discussions with peers and others in the industry. Further, since the time of my experience, I became aware of the literature around the data mesh concept and now recognize the considerable overlap between it and the data strategy discussed below. As a result, I view this data strategy as an application of data mesh.

DATA STRATEGY HIGHLGHTS

According to Gartner, "A data strategy is a highly dynamic process employed to support the acquisition, organization, analysis, and delivery of data in support of business objectives."

Other definitions describe data strategy in terms of the tools, processes, people, and rules. In the below discussion, I focus on data strategy as it connects directly to the business objectives of the domain. My experience in the MA domain may provide a slight "defensive" slant to the discussion but the salient elements should carry over to an "offensive" data strategy as well. In addition, I have chosen not to focus on technology (infrastructure, tools, etc.) even though technology is obviously an important part of designing and executing a data strategy. My experience is that the domain CDO is a businessperson who needs to articulate and establish the data capabilities needed for domain success. While he or she will make decisions regarding technology to meet these capabilities, such decisions are usually in concert with and informed by their IT partners. I have discussed the CDO role for a domain at some length separately.

The MA domain typically produces a lot of data and provides it to the rest of the enterprise and, in many cases, externally as well. Given the importance of this data and the fact that it is produced from the output of complex models and analytic processes that are developed by and well understood only within the MA domain, it is important that the domain take full ownership of such data. The domain CDO, with single point accountability to meet the domain's data needs, is then the owner of the data for the domain.

In order to fully own the data that the domain creates and provides to other domains or externally, the following are key components of the domain's data strategy, all of which need to be owned by the domain:

1. Single data source for all domain processes
2. Data curation and data pipeline creation
3. Data governance of domain data
4. Data quality process

5. Datamart with self-service for domain provided data
6. Data culture instilled across domain

Each of these components is discussed in more detail below and, at the end, I discuss their alignment with a data mesh.

1. SINGLE DATA SOURCE FOR DOMAIN

It is important that the domain construct a single data source from which to consume for all its needs. This is particularly true of a MA domain that supports a number of different modeling and analytical processes. For example, in my experience in financial services, the MA domain is often responsible for credit risk modeling, stress testing of capital against extreme economic scenarios, loss forecasting and reserving and various risk and portfolio management analytics, all typically calculated at the level of individual loans. It is essential that all these processes use the same data source so that they are working with the same definitions, assumptions, lineage, system of record, etc. for their data so that there are no downstream issues with reconciling data in different reports even though the data may appear correct individually.

Many people have their own story to tell of meetings where reports were presented that were inconsistent even while the authors were claiming to use the same data, leading to confusion and even friction.

Fig. 1 shows the data flows within a large company. On the far left, is the point of data capture—if the company is a bank and the data pertains to a loan made to a customer, it is typically captured in a loan document either by the customer or by a bank representative. This data then flows from the point of capture to what is known as the source system. There can be multiple source systems, each one aligned to a business domain, product or other organizational grouping, and each source system is typically a System of Record (SOR) for some or all of the data fields in it.

The data from the source systems then flow to the Enterprise Data Warehouse (EDW) landing first in a staging area. The data is then normalized, transformed, and integrated. They are then made available for consumption across the enterprise through the consumption layer of the EDW. The right side of Fig. 1 shows a domain that uses data from the EDW to then create a domain data source (DDS). The domain then uses data from the DDS for various modeling and analytics processes and the resulting output of these processes is provided to other domains through a datamart. Finally, as shown, there could be several such domains, each creating their DDS and providing output to other domains through a datamart.

Figure 1

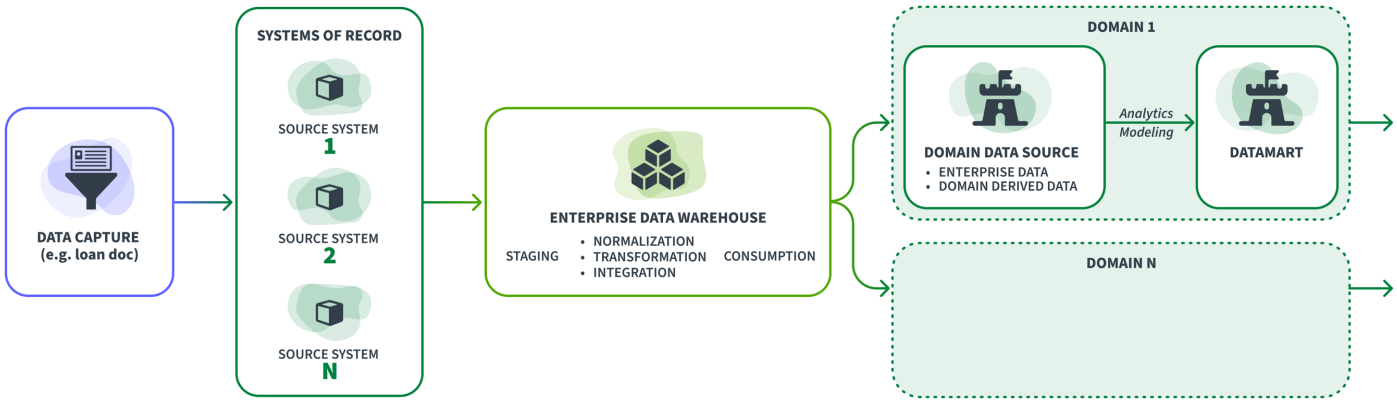


Fig. 2 shows the data sets and internal processes of the MA domain in more detail. As shown in the left-hand side of Fig. 2, the DDS usually comprises two distinct datasets:

i. Model development dataset - for training models
ii. Model execution dataset - for executing models

The model development dataset is designed to contain a number of variables (features) that are created with the purpose of being candidates for use in a family of models along with necessary metadata such as definition, assumptions, lineage, system of record (SOR) etc.

It is similar to a feature store in the context of MLOps. New variables may be added to this dataset as and when they are identified for use in training models. Once the models are finalized, the subset of these variables that are in the final model versions used for execution are then added to the model execution dataset. In the case of financial models, which involve data over time, models are developed on historical data and are then executed using current data typically to predict future behavior or outcomes i.e. model development data tends to be historical data while model execution data is typically current data. As a result, a given time period's model execution data (current data) will become part of the model development data in the next period since each period's current data then becomes the next period's historical data.

Figure 2

2. DATA CURATION AND DATA PIPELINES

The MA domain prepares data and uses it to train models and, once these models are trained, the data that they are trained on are then used to execute the models as well as to monitor their performance. The feedback loop from model monitoring is then used to improve the models as well as to adjust them to new data.

MLOps is essentially this process with a strong focus on automation. The data consumed by the domain are usually made up of data created outside the domain and used widely across the enterprise ("Enterprise Data"), data created within the domain that is derived from Enterprise Data for use within the domain typically to feed the models ("Domain Derived Data" or DDD) and, in some cases, external data. The domain obtains Enterprise Data from the EDW and uses it without any modification while the DDD is created by modifying data fields from the EDW and, in many cases, combining them with other data fields to derive the inputs specifically needed by the models. The analytical processes in the MA domain may also require all these types of data fields.

An example of Enterprise Data would be the FICO score (of a borrower) which is a data field that is obtained from an external source based on the borrower's information and is typically widely used across the enterprise. An example of a DDD field related to this Enterprise Data field would be the attribute FICO < 660. This data field is an indicator variable (i.e. it is 1 if FICO is less than 660 and 0 otherwise) that is derived from the FICO score data field and is often used to identify borrowers with weaker credit quality. This indicator variable may be used in one or more models in the MA domain by itself or it may also be combined with other data fields to create a new data field, which would then also be a DDD, before use by the models. However, because their use is specific to models in the MA domain, they are less likely to be used across the enterprise. Further, different models in the MA domain may require more than one such indicator variable (e.g. FICO < 700, FICO < 660, FICO < 620) for their specific modeling purposes, necessitating the creation of additional DDDs.

The domain CDO needs to be responsible for the creation and maintenance of the DDS. This provides the domain CDO function the speed and agility to respond to changing business conditions. The alternative of having the DDS be created and maintained by the Enterprise Data Office (EDO) or another domain is generally sub-optimal as it would put the domain's priorities in competition with those of the EDO or other domains with no guarantee that the response to the domain's needs would be timely. Pursuant to this, the domain will need to own the curation of the data in the DDS, that includes Enterprise Data and DDD, as well as the building of the respective data pipelines to support that. These pipelines, for the most part, would be from the EDW, which the EDO is responsible for.

List of Acronyms

CDO: Chief Data Officer
MA: Modeling & Analytics
EDO: Enterprise Data Office
SOR: System of Record
EDW: Enterprise Data Warehouse
DDS: Domain Data Source
DDD: Domain Derived Data
ESG: Environmental, Social & Governance
DaaP: Data as a Product

3. DATA GOVERNANCE

The domain needs to be responsible for data governance of all the data it has created. In the case of the MA domain, these are mainly from the output of complex models and analytic processes that the domain has developed and has intimate knowledge of. No other domain is in a better position to govern this data. The data produced by the domain also includes DDD which is part of the input data feeding models and processes. As shown in Fig. 2, it also includes the large volume of intermediate output data which is data produced along the way to calculating the final output or, in many cases, is a by-product. For example, if the final output is the cumulative loss expected on a 5-year loan, the intermediate output may be the probability of default in each of the 5 years of the loan's life. The intermediate output is used to support a variety of analytics that not only help provide intuition into the final output but also drive various analytical processes and is often as valuable as the final output.

For effective data governance, the domain CDO will need to establish a data governance body with purview over all data used and produced by the domain. This body should be chaired by the domain CDO or their head of data governance and, to ensure that it has the necessary clout, it needs to have the backing of domain leadership as well as of the enterprise CDO. The body will typically have representation from the business domains that own the source data, the EDO, IT, finance, risk and also include domain staff such as data scientists/modelers (henceforth, I will use the term modelers to include data scientists as well), data engineers and control experts, among others. The domain CDO will need to develop a charter with clear roles and responsibilities for the data governance body. He or she should also leverage pre-existing data governance to avoid duplication of oversight as it is often the case that Enterprise Data is already being governed by the EDO. Further, if some of the inputs to the DDDs are effectively being governed elsewhere in the enterprise, the domain can build on that for its data governance.

Along with setting up a data governance body, the domain CDO also needs to establish a data governance process that involves regular meetings of the body in order to ensure timely identification and resolution of all matters pertaining to domain produced data and the associated metadata. Such matters would include data quality, data definitions, assumptions, lineage, SOR, data taxonomy as well as access control policies. Data taxonomy can also include, for example, classification of assumptions about missing data, also known as imputation (Fig. 2)—whether an assumption is local (made in the context of a model or an analytical process) or global (made for all models and processes). Assumptions like these about data, whether made upstream in other domains or within the MA domain, constitute part of the metadata and need to be exposed to governance. The data governance process would be subject to periodic review by internal audit as well as, if applicable, external audit and regulators.

One important area of focus for the data governance body is to ensure consistency between the data used to train models and the data used to execute the same models (Fig. 2). Modelers usually experiment with a number of data fields (features) in developing their models and, in some cases, sourcing the data fields themselves. Once the model is finalized, validated and put into production, the data fields that are in the final version need to be made available through data pipelines that are also in production so that the model's execution in production can be supported. It is very important that the data pipelines provide data in production that is consistent with the data used to train the models where consistency refers to the data having similar values as well as similar metadata including, in many cases, the same lineage. This is to ensure the model's accuracy, so

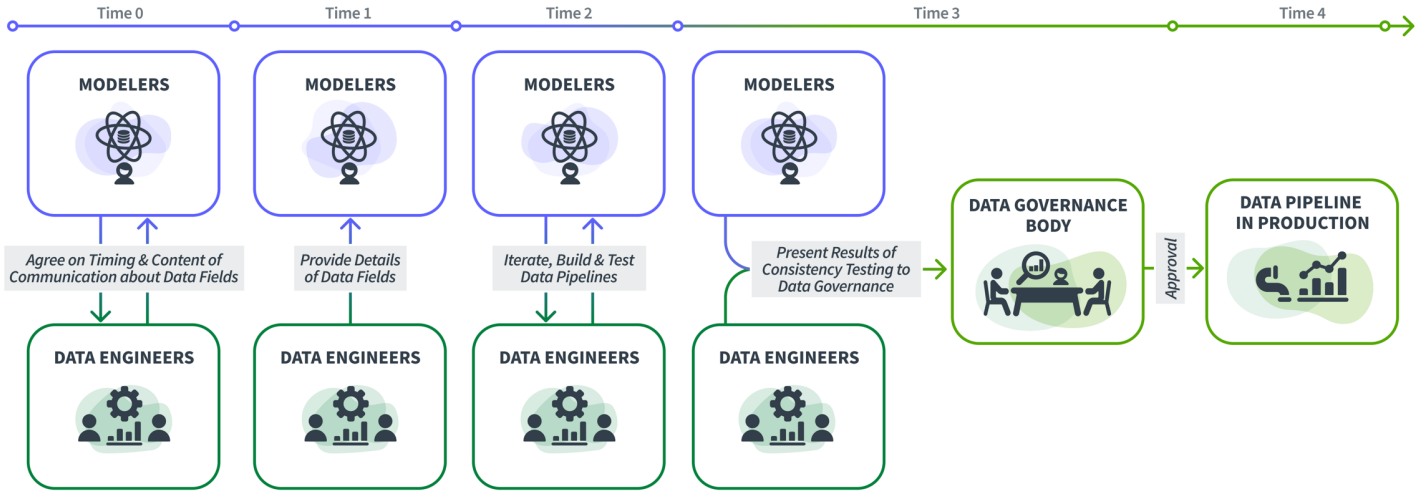
that the model's forecasts in production are based on the same data that the model was trained on. Note that data consistency as referred to here is different from the concept of "data drift," which usually refers to changes in the statistical distribution of the data from the time the model was trained to when it is executed (though some definitions of data drift also include changes driven by coding or infrastructure which would fall under data consistency).

How does one ensure data consistency? First, there needs to be a process that formalizes communication between modelers and data engineers sufficiently early in the model training process. As noted earlier, in order to do their jobs well, modelers need the freedom to experiment with a large number of data fields in a "sandbox" type environment before they settle on a smaller subset that becomes part of the final model. Data engineers, on the other hand, need advance notification of the data fields that will be used by the finalized model in production along with the necessary metadata in order that they have sufficient time to build the requisite data pipelines in production. Since models typically need to be executed right after they are productionalized, the challenge, therefore, is to devise a communication process that balances the needs of the modelers and the data engineers.

Such a process can be as follows. When a modeler is able to determine that a data field that they are using to train the model is likely, with some agreed upon level of confidence, to end up in the final version of the model, they will then communicate this to the data engineering team so that the latter can start work on designing and building the data pipeline needed to deliver this data field in production. Once the pipeline is built, the data from the pipeline will then be compared with the training data as part of the overall testing of the pipeline and the results of such testing will then be presented to the data governance body whose approval is required before the pipeline can be used in production. This ensures that consistent data is available in production in a timely manner. If modelers wait until the model is finalized to have this communication with the data engineering team, it may not always provide adequate time for the data pipelines to be ready when the model needs to be executed in production.

Fig. 3 illustrates such a process.

Figure 3



4. DATA QUALITY PROCESS

The domain CDO function needs to leverage its knowledge of the domain and partner with data consumers across the domain and with other domains in establishing a data quality process. This will typically take the form of design and placement of data quality controls such as:

- Movement: Is the data flowing from one location to the next without getting corrupted?
- Range: Is the data within normal and acceptable ranges?
- Trending: Are the trends in the data over time consistent with intuition?
- Relationship: Is the relationship between two or more data fields changing?

These controls will usually have ranges of acceptable values and thresholds for what is considered out-of-bounds. Regarding the location of these controls, it is generally preferable to upstream more controls (i.e. have the controls be preventative rather than detective), however, not all such controls can be upstreamed. The domain data governance body will need to oversee and approve the results of the data quality process.

When the results indicate abnormalities with the data, the data engineering team in the domain CDO function will typically need to research and identify the underlying causes. The results may surface genuine changes in the data over time but could also indicate issues either upstream or with the data pipelines.

The term "data observability" is increasingly used in the context of data quality with the distinction between data observability and data quality stated by one source as "..Data quality aims to ensure more accurate, more reliable data. Data observability seeks to ensure the quality and reliability of the entire data delivery system..".

Proponents of data observability state that it involves continuous and proactive monitoring of data quality factors as well as of items like lineage and schema. While lineage and schema are typically not monitored proactively within a data quality process, they do get reviewed if the research identifying the root cause of the data anomaly uncovered leads to them. Also, the placement of preventative controls in a data quality process, that was mentioned earlier, is intended to make the monitoring be more proactive. That said, data observability has the potential to make a data quality process be more timely, thorough, and efficient in detecting and fixing anomalies and that is, in particular, through the use of automated monitoring and root cause analysis and, increasingly, AI/ML to propose data quality thresholds.

A robust and well controlled data quality process is critical particularly for a domain whose data is A robust and well controlled data quality process is critical particularly for a domain whose data is consumed by external partners, regulators or are used in financial statements. The process also needs the flexibility to quickly adapt to new data with different characteristics. For example, at the onset of Covid-19 in 2020, financial institutions had to gather or create new data in order to better understand the credit risk of borrowers. This was due to typical credit quality indicators such as delinquencies becoming less useful since borrowers were permitted by the CARES Act to stop making payments on loans. The data quality process then needed to quickly accommodate this new data and ensure their quality.

5. ANALYTICS AND VISUALIZATION

Analytics are key to extracting the insights that are in the data. Peter Sondergaard, then at Gartner, famously said in 2011 "Information is the oil of the 21st century, and analytics is the combustion

engine." A well-designed analytics capability can be a multiplier for the value provided by the data and its impact on the quality of business decisions. As shown in Fig. 2, the analytics and visualization capabilities are positioned such that they can be applied to both the intermediate and final data outputs of the domain's processes.

There are at least two considerations when designing an analytics capability to unlock the most intuition and insights from the data. First, the capability needs to provide increasingly deeper views of the data as well as views from different perspectives. For example, if the domain makes use of time series data to drive business decisions, the use of waterfall charts that show the changes in data from one time period to the next as well as the attribution of such changes to changes in various factors over these same time periods can be insightful. Similarly, if the data is aggregated from different components, then the capability to readily break these apart and understand how changes over time in each of these components are contributing to changes in the aggregate can also be quite useful. A popular example of this today is understanding how inflation rates in the different components of the consumer price index are driving the overall inflation rate and what might be the overall forecast of the inflation rate based on the trajectories of the components.

Second, the analytics capability needs to include visualization and enable ad hoc exploration. As is well understood, the brain is much better able to draw insights and develop intuition when the data are represented visually. A good visualization capability takes advantage of this by making skillful use of colors, shapes, sizes, backgrounds, etc. as well as of different types of graphs and, further, it also allows the user to easily interact with it by, for example, filtering and drilling down further. Ad hoc exploration capabilities are very important since they enable timely pursuit of new lines of inquiry that are thrown up during the interaction with analytic tools. Ad hoc capabilities also increasingly include the use of natural language processing (NLP) techniques that lets less technical users query the data with commonly used words which broadens the universe of those interacting with the data and can lead to greater democratization of the insights. Ad hoc analytics, as with any other analytics, should be subject to appropriate governance, which is usually different from data governance, when used in business decision making.

6. DATAMART WITH SELF SERVICE

When providing output data to other domains, it makes the data more useful by providing it through a datamart structured such that the various data outputs are integrated per a data model that When providing output data to other domains, it makes the data more useful by providing it through a datamart structured such that the various data outputs are integrated per a data model that reflects the business uses of the data. The advantage of this is that users obtain data that is already organized to meet their needs instead of having to do so themselves. To take an example from financial services, quantities such as average or expected credit loss, credit losses and revenue under different stress scenarios as well as related analytics that are the output from different processes can all be linked together at the level of an individual loan so that the user can readily obtain all of these quantities for any loan or set of loans. A self-serve capability then makes this datamart easier to use and more accessible.

7. DATA CULTURE

Instilling a data culture in the domain is one of the most important aspects of the domain CDO's data strategy and is critical to the success not only of the

strategy itself but also to advancing the domain's objectives. While there have been various articles written about data culture, I will focus on what I have found to be important for a domain such as MA.

First, the domain CDO needs the full support not only of the domain leadership but also of the next level up leadership which is often important in ensuring the necessary resourcing. Further, it is particularly important that the data engineering and data governance functions in the domain have similar stature to that of the modeling and analytic functions. Next, for the data strategy to be well executed, the modelers and the data engineers need to partner effectively with each other, and this can happen best only when each group understands the other's business. In particular, this requires that the modelers appreciate the importance of good quality data - what constitutes good data quality and what does it take to achieve that – and the data engineers understand how the models are using the data and which data fields are most critical. I have found that, once this two-way understanding was developed, both parties were extremely eager to partner with each other and modelers, in particular, were often proactive in developing data quality rules and in identifying data issues. Finally, there needs to be a minimum level of data literacy across the key functions. This is also a key responsibility of the domain CDO and one that lays the foundation for the domain's data culture.

RELATIONSHIP TO DATA MESH

By now, you will likely have noticed several points of commonality between what I have laid out above and a data mesh. The key aspects of a data mesh are:

A. Domain ownership of data and data pipelines
B. Data as a Product (DaaP) view by domain
C. Data interoperability across domains, enabled by federated data governance
D. Common infrastructure for domain-agnostic activities, leveraged by all domains

This white paper's discussion on domain data strategy clearly refers to A. While I do not use the term Data as a Product, a domain data strategy that emphasizes data governance and data quality along with greater accessibility and democratization of output data is primarily about ensuring that the data provided by the domain meets the consumer's needs, which is consistent with B. Items C and D are not explicitly discussed in the white paper but they follow from items A and B directly – if every domain owns their data and provides it as a product to other domains, then there needs to be interoperability across domains with the governance that ensures this being federated across domains (C). It is also efficient for domains to use a common infrastructure (D) to support domain agnostic activities, since domains can then outsource such work and instead focus on domain specific work for which they have the best skills. The principles in this white paper outlining a domain data strategy can, therefore, be seen as an example of the data mesh concept in practice (In Fig.1, the source systems on the left side reside within different domains, so that domains are both supplying data to the EDW as well as consuming data from it).

When I read Zhamak Dehghani's paper, which was after leaving my domain CDO role, I was immediately struck by its vision of a federation of more autonomous domains each owning their data and creating data products in a decentralized manner. I also recognized the overlap between the data mesh and my perspective, detailed in this white paper, derived from working within a domain. That said, I will note the following.

First, while the data mesh envisages a federation of

autonomous domains setting rules for all of them and sharing common infrastructure where it is efficient, the reality is that most organizations, particularly financial institutions, tend to have a centralized EDO that, along with setting enterprise standards and policies for data, is also responsible for much of the enterprise infrastructure including centralized parts such as the EDW. This does not necessarily preclude the idea of a data mesh but it means that, in such organizations, domain autonomy needs to be earned and requires demonstrating the maturity of the domain in data matters particularly with respect to data governance and data quality processes. There are also often well justified concerns about domains creating data silos with all of their deleterious results and it requires clear communication to overcome such misgivings (which are, in some cases, political too) in order for a domain to take ownership of data curation and data pipeline building for the data it uses and produces.

Next, not all domains have a similar level of understanding of their data needs and/or a similar level of sophistication in their capability to meet the same within the domain. They may not have a domain CDO or, even if they do, may lack data engineering talent. This criticism has been leveled at the concept of a data mesh i.e. that it is not very useful in practice as most domains are not capable of greater autonomy and it has been constructively suggested to instead have a piecemeal implementation of a data mesh. Similar to this suggestion, a domain can position itself along a continuum on the path to fully owning its data based on the level of its data maturity and capability i.e. it does not have to be all or nothing with regard to domain autonomy. As the domain's sophistication and confidence in its capabilities grow, it can then occupy a more advanced position on this continuum. Even for more sophisticated domains such as MA, there often first needs to be a recognition internal to the domain that the complex data needs of the domain are better met by a domain CDO function that owns the data for the domain.

Finally, for a domain such as MA which produces a large volume of modeled data and analytics that are consumed in other domains and even externally, it is very important that the domain ensure data governance over all of the data it produces and consumes in order to be able to stand behind this data. While a federated data governance is a key part of the data mesh concept, the emphasis on a robust domain data governance process needs to be very strong for domain such as MA.

APPLICABILITY TO OTHER DOMAINS

As noted earlier, the above discussion of a domain data strategy applies to other domains which involve the use of data and analytics in their business decision making. These would include, among others, finance, marketing, sales and HR. Some of these domains e.g. marketing also make use of models in addition to analytics.

The MA domain in financial services has tended to more "defensive" in terms of data strategy as it is often responsible for measuring risk and supports regulatory and financial reporting objectives such as stress testing of capital and loss reserving. That has started to change in recent years though, in most cases, the domain's data strategy is still likely to tilt defensive.

Domains such as marketing and sales, on the other hand, tend to have more "offensive" data strategies, as they seek to use data to uncover new business opportunities or broaden existing ones. The above discussion applies across all kinds of domains and aspects of the data strategy such as data quality and data governance can be calibrated to the needs of the domain e.g., financial reporting typically has a very high bar for data quality since the data feeds

financial statements that are reviewed by external stakeholders such as investors, analysts, and regulators.

CONCLUSION

In this white paper, I have focused on the domain data strategy. Business domains increasingly consume and produce a great deal of data and it is important to recognize their ownership of the data that they produce. The goal should, therefore, be to empower domains so that they have the capability to act with speed and agility in meeting their business objectives, which is also recognized by the data mesh concept. A domain's ability to design and execute a domain data strategy is key to the domain's and the overall enterprise's success.