Summary Insights: Data Governance and Classification

RFG Perspective: In today's hybrid multi-cloud, data-driven world, the integrity and role of data – an enterprise's greatest asset – has never been more important. Nonetheless, strong, comprehensive data governance practices are not yet fully in place across all domains. Sadly, only a small percentage of the world's data is automated and up-to-date, resulting in multiple governance gaps. Moreover, data owners and custodians do not always have the full transparency they need to execute their data governance directives and policies. Business and IT executives must address the growing concern that data governance best practices are not implemented everywhere, resulting in inadequate controls and causing inconsistent compliance with data privacy regulations and corporate data security policies.

INTRODUCTION

One of the axioms of IT is that data is the new oil – meaning that a company's data assets are a very valuable resource that can be analyzed to produce better business outcomes. Finding efficient and structured ways to govern company and personal data is "Job One" for data governance teams.

However, customers are facing the reality that the current state-of-the-art means that most data governance and classification tasks are being done manually and via spreadsheets. This approach is inefficient – and can introduce errors and delays into the classification process itself. That has become even more difficult in the face of the COVID-19 crisis and the New Normal work-from-home (WFH) conditions, in which employees who fled their offices access vital, mission-critical data from their homes, often without the standard data controls found in corporate offices.

Automation of the data-classification process will be an important next step in speeding up the tagging and classification of large amounts of data for subsequent analysis. Enterprise executives really want to find – and to use – good AI/ML software tools that could help them to achieve faster data classification and better data governance. Those tools should be used to manage data across data centers, edge devices, enterprise networks, hybrid clouds and multi-cloud deployments.

Many business executives are not confident that the software tools they have are adequate to deal with this demanding task, especially in highly distributed WFH environments. For this reason, software vendors must pay attention to a wave of customer concerns about data governance. These software firms must address these concerns by supporting standardized interfaces and by raising customer awareness about software tools' capabilities and patterns of deployments.

RFG 100 and Insights from Participating RFG 100 Executives

On Dec. 2, 2020, RFG facilitated a videoconference on "Data Governance and Classification." The panelists on the call were:

- Chris Doris, VP, Morgan Stanley
- Peggy Tsai, VP Solutions, BigID
- Mike Davis, CISO, Alliant Group
- Jean S. Bozman, President, Cloud Architects LLC

During this videoconference, RFG 100 executives and panelists discussed the business-critical issue of re-aligning data governance and classification to address the changes that occurred with the rapid shift to WFH last spring and accelerated migrations to hybrid cloud and multi-cloud environments. In this research document, we include their key findings about the challenges that are associated with data classification challenges and data governance exposures.

RFG Analysis of Poll Results

We polled the RFG100 executives about their data governance requirements. The poll results show that enterprises are struggling with updating their data governance programs. More specifically, many enterprises are scrambling to keep up with all the regulatory requirements imposed upon them by government agencies around the world.

Until there are better interfaces between AI/ML software and customers' data governance tools, organizations will not be able to comply with the oversight of the regulatory agencies. If they do not comply, they will be exposed to fines imposed by governmental regulations – or to lawsuits brought by individuals who believe that their data privacy and PII (personally identifiable information) have not been adequately protected.

Poll Results

The December poll of RFG100 conference attendees turned up the following results:

- Nearly 45% said their companies had made substantial progress in organizing data, and in establishing multi-site data-governance policies across their enterprise.
- The same respondents reported that "there is much more work to do," meaning that such efforts will require more effort – and more time – to fully implement.
- Respondents cited multiple reasons for working to improve data governance during the current crisis, including:

- Ensuring a consistent data retention approach across the enterprise (67%)
- Protecting PII – personally identifiable information (56%)
- Setting and maintaining common standards across toolsets (44%)
- Ensuring data quality across the enterprise (44%)
- Addressing data aging (39%)
- Enabling data tagging, especially at point of ingestion (39%)
- Tracing data lineage to ensure data integrity (28%)

- Two-thirds of the RFG100 respondents stated that their data definitions are in place. However, 22% indicated that the definitions have not been fully implemented throughout their organization.

ANALYSIS

Enterprises must have a consistent, unified view of all of the data they are managing, regardless of where the data is stored or what format it is in. Yet, the reality is that enterprise executives are always accountable and responsible for the data being managed throughout their organization – regardless of where it resides.

It would be nice to assume that cloud service providers (CSPs) assume a level of responsibility for data governance – and they do assume some responsibility, according to the contract signed by the customer and the CSP. However, some enterprises are still seeing some degree of difficulty when they ask their cloud provider to make customized connections to data that customers believe are needed to gain better control of their data governance practices in the hybrid cloud.

That's why it is pragmatic to assume that cloud providers will respond more quickly to requests from large enterprises, due to long-standing business relationships and size of company. Examples would include requests for applying the CSP's AI/ML software to the enterprise's existing set of data governance software tools. Alternatively, CSPs could offer AI/ML services to bridge any gaps relating to end-to-end data governance.

Another key is to demonstrate that the enterprise and the CSP have agreed to have shared objectives in applying consistent data governance practices for data, no matter where it is stored. Many of the financial services firms have gotten together under the banner of the EDM [Enterprise Data Management] Council to address their requirements. The council consists of a number of financial services firms and software and services firms that focus on data standards, best practices and the development of training and certification programs. RFG100 executives said this

approach is working – and should be expanded to include more companies – generating more corporate input to the emerging EDM-certified standards.

Automation of data classification – manual process vs. automated.
Automating the process of data classification and data labeling is a "must" because manual processes take so much longer – delaying the improvements in data governance that businesses need to achieve.

To complicate matters, the process of application development is rapidly transforming to a cloud centric, container-based landscape of "born on the cloud" applications. The emerging areas of DataOps, DevOps, and FinOps (cloud financial management) – to name three examples – depend upon careful review of data classification, data cleansing and data governance procedures.

Data types are also important to recognize, in any data governance policy. Most of the applications developed using cloud-centric technologies generate object-based data, which is stored differently than the widely deployed block-based and file-based data. That is why classification and governance of data in object databases is an additional, and different kind of, challenge that governance groups and software tools must address. Thus, customers and CSPs must support a combination of object, block and file data types in any enterprise-wide or hybrid cloud data plan.

Finding the right software tools to automate the process is a necessary part of automating data classification. However, governance tools are still maturing. At best existing tools can automate 70-80 percent of the classification effort. The rest of the process is still manual. It is this current state that results in some organizations reporting that they are still looking for automation tools – and better interfaces to those tools – that fit their business' needs. Clearly, there are many software tools in the marketplace, so IT organizations must work closely with their organization's business units to identify – and to agree on – which software tools best fit their analytics and business needs.

Data Sprawl and Simplicity.

Data sprawl is a major challenge for many organizations, as dozens –or even hundreds – of copies of databases and applications proliferated throughout the campuses and business units. Examples include Microsoft's SQL Server database and Microsoft Sharepoint databases, instances of Oracle relational databases, and copies of SAP ERP software. These replicated instances complicate the data governance efforts, especially since multiple studies show there are many more copies of databases and files than management believes there to be. The amount of collected, retained and copied data should be minimized to the least amount possible to achieve best practices for data governance throughout the organization.

Over the past 15 years, there has been a push to have organizations put data custodians or stewards in place to address these and other relevant data issues. Yet less than 10 percent of enterprises have such individuals on staff. This is an unfortunate situation, because data stewards are the staffers who address enterprise-wide data classification, definition, cataloging, ownership, sprawl and usage of data.

Another data challenge is that data lakes can morph into what some are calling "data swamps" if data is not cleansed, managed and properly labeled – all on an ongoing basis. As part of this cleanup process, "data silos" – developed by different business units within an enterprise – are to be avoided, because they prevent enterprise-wide data searches and data management. The process of drifting into a data swamp may take years – and it may also take years to separate the datasets into manageable segments.

Working with Data from Microsoft Office 365 and SaaS Providers

Microsoft Office 365 – a cloud service that supports and delivers Microsoft Office – and SaaS software are widely adopted by enterprises worldwide. But familiarity is not enough for many enterprises. RFG100 panelists said that many organizations are looking for "best practices" that can applied to using the management tools associated with these well-known cloud services. Specifically, customers are looking for better ways to integrate Office 365 and other SaaS services with their organization's existing portfolio of data management software products.

Regulatory Requirements

Regulatory compliance is a fact of life, as more and more enterprise workloads migrate into hybrid clouds. Enterprises find that complying with many regulations is not an option, but a requirement, which continues to get more and more complex as additional countries create their own security and governance policies and standards. Examples include SEC compliance, data privacy compliance (e.g., the European Union's GDPR and California's CCPA), and data-location requirements in a geography (e.g., Germany and Canada).

In the U.S., one approach is to ensure compliance with U.S. federal security standards, as certified by the CSA (Cloud Security Alliance) and by FedRAMP (Federal Risk and Authorization Management Program). RFG100 respondents see the support of high-level clearances as a necessary part of overall, end-to-end data protection in corporate and federal-agency environments.

SUMMARY

Data governance is gaining importance in the wake of COVID-19 and WFH use of multiple devices and cloud services. This is a global concern, although specific regulatory bodies are making policies that apply to specific geographies (e.g., countries or regions).

Many enterprises have data governance policies that are adapting to the New Normal computing environment that emerged in the spring of 2020. Given the increased focus on classifying and cleansing data – it is clear that automation, AI and ML will all play major roles in promulgating the new data classifications and data standards in a consistent way, across organizations – and reaching into the cloud.

RFG POV: As the year 2020 neared an end, CxOs were well-aware of the need to scale up their data governance efforts – and to do so as rapidly as possible to catch up to the current operational realities. However, thoughtful adoption of technologies and best practices will be required in order to ensure that data governance is applied consistently and securely across large and complex enterprise environments. Otherwise, data catalogs, data dictionaries and data repositories would be incomplete and would lack the controls, integrity and transparency required for satisfying corporate and governmental compliance.

Additional relevant research and consulting services are available. Interested readers should contact Client Services to arrange further discussion or interview with Mr. Cal Braunstein, CEO and Executive Director of Research. Jean S. Bozman, President of Cloud Architects Advisors LLC, co-authored this report.