WHITE PAPER

SAS®: A Comprehensive Approach to Big Data Governance, Data Management and Analytics

SUNIL SOARES
JULY 2018

INFORMATION ASSET, LLC

Contents

About the Author

Sunil Soares is the founder and Managing Partner of Information Asset, a consulting firm focused on data governance and enterprise data management. He is the author of several books, including *Selling Information Governance to the Business*; *Big Data Governance*; *Data Governance Tools* and *The Chief Data Officer Handbook for Data Governance*. For more information, please visit information-asset.com.

Today we are drowning in data from sources such as social media, call transcripts, smart meters and digital pictures. IDC forecasts that by 2025 the global datasphere will grow to 163 zettabytes (a trillion gigabytes). That's 10 times the data generated in 2016. All this data will unlock unique user experiences and a new world of business opportunities.[1]

This deluge of data needs to be governed, and that's where data governance comes into play. Big data governance is part of a broader data governance program that formulates, monitors and enforces policies relating to big data. Because its early focus has been on cost-effective analytics, some big data platforms – such as Hadoop, a low-cost open source parallel processing platform – have limited support for big data governance disciplines (e.g., metadata, data quality and information security).

SAS® Spans Data Management and Analytics

Even before their profession became in vogue, data scientists had been using SAS for decades to solve complex statistical problems. Beyond this core capability, SAS offers a comprehensive approach to supporting big data management, analytics and governance – whether that data is in stream, in Spark, in database or in Hadoop. SAS provides a common metadata backbone across SAS Data Management, SAS Analytics and third-party platforms to support data governance with embedded data quality, data integration, master data management and data preparation capabilities. This backbone is critical, because once data lands in a big data environment like Hadoop, much of its descriptive information is lost - making access, management and governance more difficult. Let's review some of the areas where big data governance comes into play.

Big Data Governance Considerations

There are five broad categories of big data that need to be analyzed and governed:

1. **Web and social media data**. This includes clickstream and social media data such as Facebook, Twitter, LinkedIn and blogs. Most companies now try to integrate this data with master data and core business processes such as customer loyalty programs. The data governance team needs to establish policies regarding the acceptable use of social media data, especially since regulations and precedents are continually evolving. Metadata is also critical to web and social media. For example, two sites might measure the term "unique visitors" differently for clickstream analytics. One site might measure unique visitors within a month, while the other might measure unique visitors within a week.

2. **Internet of Things (IoT) data**. IoT technologies allow both wireless and wired systems to communicate with other devices. The IoT uses a device such as a sensor or meter to capture an event (e.g., speed, temperature, pressure, flow or salinity). This event is relayed through a wireless, wired or hybrid network to an application that translates the captured event into meaningful information. It's the data governance team's responsibility to set policies around IoT data. For example, data stewards need to draw up guidelines around the acceptable use of geolocation data that can be used to build a profile of individuals without potentially violating their privacy. Data stewards need to establish retention policies around the massive volumes of IoT data that can easily overwhelm IT budgets if not properly controlled. The data governance program needs to address any data quality concerns, such as sensor readings in environments with high moisture content and lots of congestion. Finally, the information security team needs to secure the supervisory control and data acquisition (SCADA) infrastructure from vulnerability to cyberattacks.

3. **Big transaction data.** This category includes health care claims, telecommunications call detail records and utility billing records. Big transaction data is increasingly available in semistructured and unstructured formats – and data governance challenges apply to this data. Consider that telecommunications companies need to set acceptable policies regarding the use of GPS data from their wireless customers. They would need to ask, for example, whether to send an SMS containing a discount coupon based on the proximity of a subscriber to a specific retail outlet.

4. **Biometrics data**. Biometric information includes fingerprints, retinal scans, facial recognition and DNA. Advances in technology have vastly increased the availability of biometric data. Law enforcement, the legal system and intelligence agencies have been using this information for a long time. However, biometric data is now available in the commercial arena, where it can be combined with other types of data such as social media. This presents new business opportunities as well as

[1] IDC. Data Age 2025: The Evolution of Data to Life-Critical: Don't Focus on Big Data; Focus on the Data That's Big.

governance issues relating to privacy and data retention. For example, should retailers be able to marry facial recognition of in-store customers with their online profiles to build a detailed view of customer behavior?

5. **Human-generated data**. Human beings generate vast quantities of data such as call center agents' notes, voice recordings, emails, paper documents, surveys and electronic medical records. This data may contain sensitive information that needs to be masked. Companies need to establish policies regarding the retention period for this data to adhere to regulations and manage storage costs.

Big Data Reference Architecture

The big data landscape is littered with acronyms, technologies and tools that are confusing, even to sophisticated data management practitioners. So how can companies start implementing big data technologies? A reference architecture provides a common framework that describes all the components needed for the required functionality. Enterprise data architects can use a big data reference architecture to plot their big data road maps and discover gaps in their technology implementations. Data scientists and business users can take advantage of this approach to make sense out of a complex landscape. This architecture reflects three main sections (Ingestion; Cleansing, Integration and Governance; and Analytics, Security and Life Cycle) that encompass 16 regions. In this paper, we'll discuss how several SAS offerings map to the sections and regions of a big data reference architecture (see Figure 1).

Providing a Smooth Transition for Military Veterans

The Institute for Veterans and Military Families offers two free SAS programming courses and certification exams as part of its curriculum that helps service members, veterans and family members transition into new careers. It also uses SAS to drive its programs and operations, enabling it to get greater insights on performance and impact on those it serves. Data that previously lived in silos is now combined into a single data warehouse to better serve the organization's analytics needs. Decision makers have immediate access to accurate, interactive reports – which increases both precision and quality of actions, and delivers fast results.

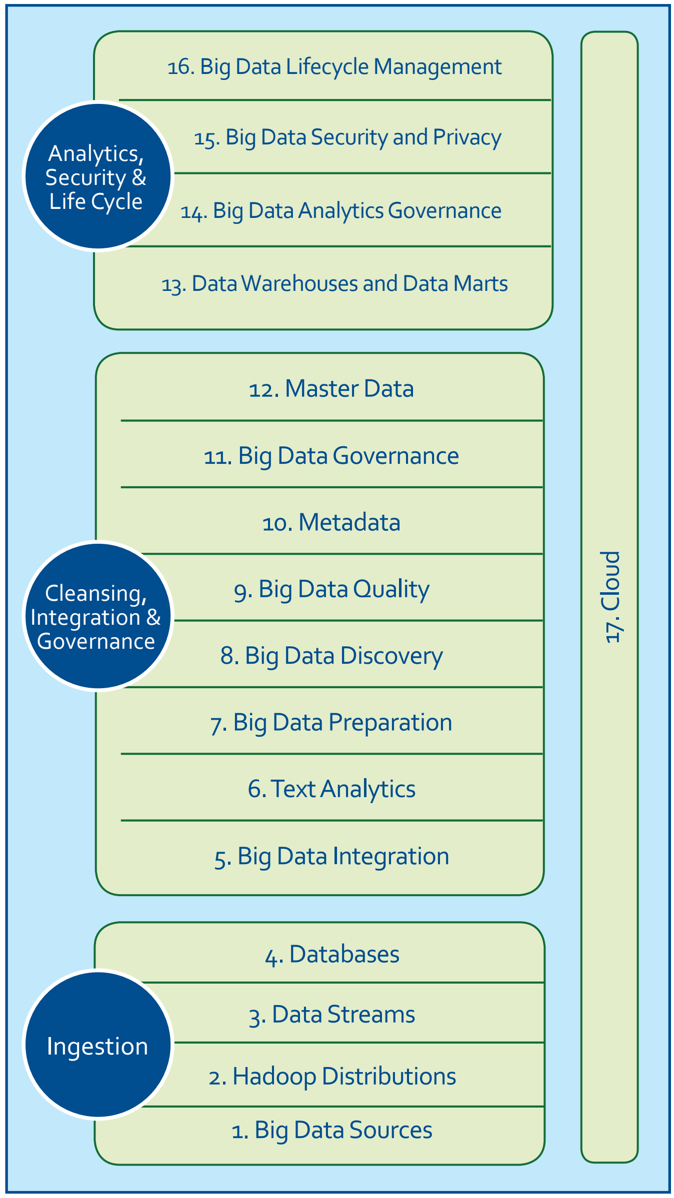

Figure 1: A comprehensive big data reference architecture should encompass three overarching methods and technologies (shown in Sections 1, 2 and 3). Section 1: Ingestion; Section 2: Cleansing, Integration and Governance; and Section 3: Analytics, Security and Life Cycle.

Section 1: Ingestion

Before doing anything with the data, organizations must first employ various methods of pulling data from a number of big data and traditional sources.

Region 1: Big Data Sources

SAS provides access to numerous traditional and big data sources through SAS/ACCESS® software, SAS Federation Server and other solutions like SAS Event Stream Processing. SAS Federation Server provides a layer to virtually blend data from multiple sources without physically moving the data. SAS/ACCESS software and SAS Federation Server support multiple big data sources, such as Apache Hive, Greenplum, SAP HANA, Teradata, HAWQ, Impala, Pivotal, Amazon Redshift and Spark.

Region 2: Hadoop Distributions

SAS provides access and processing on the most prevalent Hadoop distributions, including Cloudera, Hortonworks, MapR, IBM BigInsights and Pivotal. In addition to the access SAS provides to Hadoop and Impala, another SAS offering – SAS Data Loader for Hadoop – supports parallel profiling, data ingestion and data quality for big data. And it pushes the processing into Hadoop for improved performance.

Region 3: Data Streams

SAS Event Stream Processing examines, filters and analyzes large volumes of real-time data – like smart meter readings, health devices and stock prices – and it processes millions of events per second covering multiple data points. For example, a hedge fund may use this capability to process hundreds of feeds relating to weather, commodity prices, share prices and news feeds to make split-second decisions to buy or sell stocks. This real-time data needs to be governed like any other information. Figure 2 shows a data feed in SAS Event Stream Processing and how streaming analytics can be applied to make real-time decisions. Business users and data scientists can govern term definitions and store allowable values using other components of the SAS Data Management solution. In this manner, SAS is uniquely positioned to support both the analytics and governance of streaming data.

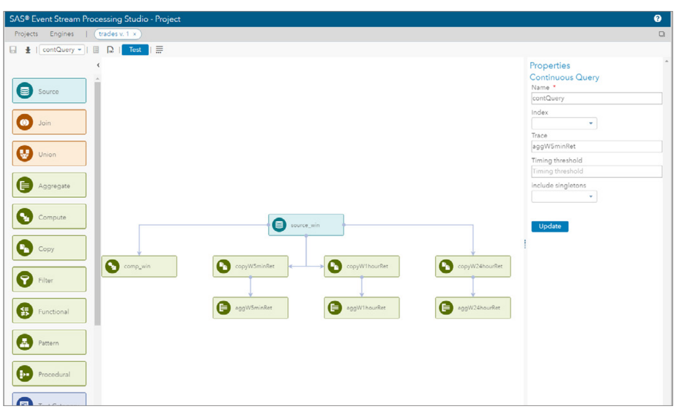

Figure 2: SAS Event Stream Processing provides real-time embedded analytics, data quality and pattern matching across different input streams at speeds reaching millions of events per second.

Region 4: Databases

SAS can read and write from more than 60 data sources, including relational and nonrelational databases, PC files, Hadoop and data warehouse appliances. SAS can also store data in SAS data sets, which are files stored in a library that SAS creates and processes. A SAS data set contains data values that are organized as a table of observations (rows) and variables (columns) of differing types and lengths that can be processed by SAS software. SAS Data Management can govern the information in both SAS data sets and other external data sets, while SAS Business Data Network can manage the business glossary for attributes in the data stores. In addition, SAS Data Quality or SAS Data Loader for Hadoop can profile and standardize the data.

Section 2: Cleansing, Integration and Governance

Once data has been collected from all these different sources, it's important to use data discovery, blending and cleansing, then incorporate policies and metadata management.

Region 5: Big Data Integration

SAS Data Integration Studio (a component of SAS Data Management) provides a visual interface to design data integration jobs for SAS users and DI developers looking to accomplish enterprise extract, transform and load (ETL) tasks. SAS Data Integration Studio supports file movement to and from Hadoop (and other traditional and big data sources), as well as a data transformation library for writing Hadoop programs in Pig, Hive and MapReduce – just to name a few of the supported transformations. SAS Data Loader for Hadoop (see Figure 3) provides a guided user interface built for business analysts and data scientists to read and write to and from Hadoop. It also lets users summarize, aggregate, merge, transpose or join data in Hadoop. Before ingesting data into Hadoop, data stewards may use the SAS workflow to gain approvals from business, IT and legal departments around the acceptable use of data.

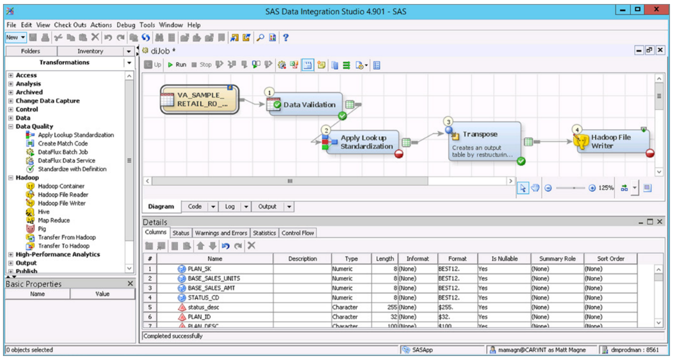

Figure 3: SAS Data Loader for Hadoop provides a web-based, guided user interface to query, transpose, transform, cleanse and profile data in Hadoop.

Region 6: Text Analytics

Organizations have huge volumes of unstructured information, including agent notes, call logs, claims adjuster notes, social media and warranty claims. SAS Text Analytics helps to derive value from this unstructured data by answering business questions such as, "Can we review our agent notes to understand why our customers are calling customer service?" SAS Sentiment Analysis helps organizations quickly understand opinions from across multiple sources, including websites, blogs, communication centers, emails, forms, surveys, internal files and reports. The software evaluates text for positive and negative connotations, including subtle emotional content, and provides a detailed breakdown to easily communicate exactly what comments mean in relation to overall sentiment and changes. Finally, SAS Data Management offerings use directed and automated identification analysis to determine the type of data in a column – as well as field extraction, which is used to extract tokens from unstructured text in fields or PDF files.

Smart Data Exploration Advances K-12 Public Education Programs

To ensure that its 86 school districts were funded accurately, the South Carolina Department of Education needed to move away from inconsistent, error-prone spreadsheets and convert data into easily consumable reports for different stakeholders. With a solution that combined SAS Data Management, SAS Business Intelligence and SAS Visual Analytics, the school system can now validate information from all its school districts and provide interactive, customizable reports that help people make smart decisions, faster. Now students are uniquely identified so that each school district has the proper funds and services.

Regions 7, 8 and 9: Big Data Preparation, Big Data Discovery and Big Data Quality

For users who need to access, profile, cleanse and transform data to prepare it for reporting or analytics, SAS Data Preparation provides an interactive, self-service environment. The solution promotes collaboration and efficiency, and reduces dependency on IT and data specialists. This means companies can act quickly, make decisions confidently and support teams in working together. The solution includes prebuilt transformations, and data cleansing functions run in memory to increase processing speed. Advanced analytics, data visualization and data preparation capabilities are seamlessly combined. Automatically generated code can be shared with IT and scheduled to run during every source data update. Data preparation plans can be reused and shared. Data lineage graphs allows users to visualize relationships between data assets, so it's easier to understand the origin of data and trace its processing.

SAS Data Loader for Hadoop enables business users and data scientists to prepare, cleanse and integrate data in Hadoop without writing code. The same process can be run in different runtime environments, including in stream, in memory, in database or in Hadoop. SAS Data Loader for Hadoop can profile data natively, in cluster and in parallel without moving it out of Hadoop (see Figure 4). SAS Data Loader for Hadoop also natively supports standardization, parsing, matching and de-duplication inside Hadoop. Additional data quality directives include casing, gender analysis, pattern analysis and field extraction. This allows users to apply case changes to the data, guess the gender based on values to improve customer segmentation, and guess acceptable data patterns based on field values. Field extraction pulls useful tokens from unstructured or free-form text within a field – such as name, organization, address, email and phone number. For example, NJ, New Jersey and N Jersey within a column's data would be categorized as "state" to aid in data exploration.

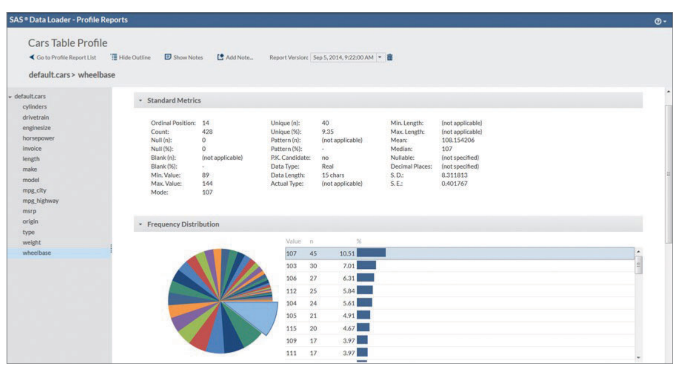

Figure 4: SAS Data Loader for Hadoop supports in-cluster, parallel data profiling in addition to standardization, matching, gender analysis and other data quality functions.

Region 10: Metadata

SAS provides a common metadata repository that spans data management, analytics and third-party data sources and tools. SAS Business Data Network supports the creation of a workflow-driven business glossary as well as relating and tracking the associated business rules, reference data, technical owners, data stewards and other associated roles. Data stewards should use SAS Business Data Network to identify a small subset of business terms as critical data elements (CDEs). CDEs are business terms that have a significant impact on financial reporting, operating performance, regulatory compliance or brand reputation. SAS Metadata Server is a centralized resource for storing, managing and delivering metadata from SAS tools and third-party data repositories. SAS Lineage provides the visualization of relationships and impact analysis to see how changes to one data element might affect information in other systems (see Figure 5).

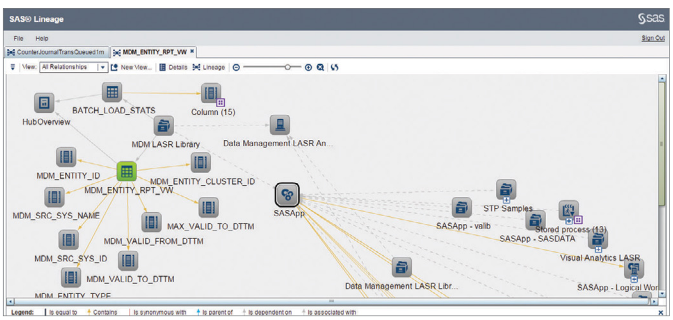

Figure 5: SAS Lineage provides a powerful lineage relationship visualization tool that spans data management, analytics and third-party metadata.

Region 11: Big Data Governance

Data scientists and the business community regularly push the boundaries in terms of finding new insights from new types of data. But data governance teams need to establish acceptable use standards for emerging data types. For example, vehicle telematics data may be gathered by insurance companies for policy underwriting purposes. However, this data contains a lot of sensitive information such as driver location and speed. The data governance team needs to set standards around who can view what data and for what purpose. These acceptable use standards should be documented in SAS Business Data Network. Entities defined in SAS Business Data Network (see Figure 6) can be approved through a rigorous governance process and then incorporated into SAS Data Management software to apply the appropriate governed standards to the data. For data that falls outside of the standards, users can route, track and fix data quality issues through a remediation interface.

Region 12: Master Data

SAS supports multiple data domains, including party, organization, site, supplier, product and custom entities. Data can be accessed from Apache Hive, Cloudera Impala and other traditional relational sources. Master data capabilities encompassing market-leading SAS Data Quality enable profiling, address verification, standardization, data enrichment, transformation, cross-field entity matching, survivorship capabilities and business rule monitoring.

SAS Business Data Network and SAS Lineage can be used to view lineage and semantics for attributes in data repositories. Data stewards should ensure that CDEs for the master data hub are well-governed, via SAS Business Data Network. For example, the attribute PHONE_NUM in the systems of record should be linked to the business term called "phone number" in SAS Business Data Network. This business term should be flagged as a CDE in SAS Business Data Network and linked to a business rule, such as: "Primary phone number should not be null." In addition, SAS Data Loader for Hadoop includes a guided user interface with master data match, merge and survive functionality that executes in memory in Spark and in the various supported big data distributions. And SAS Data Quality includes comprehensive methods of obtaining a single view of a single domain.

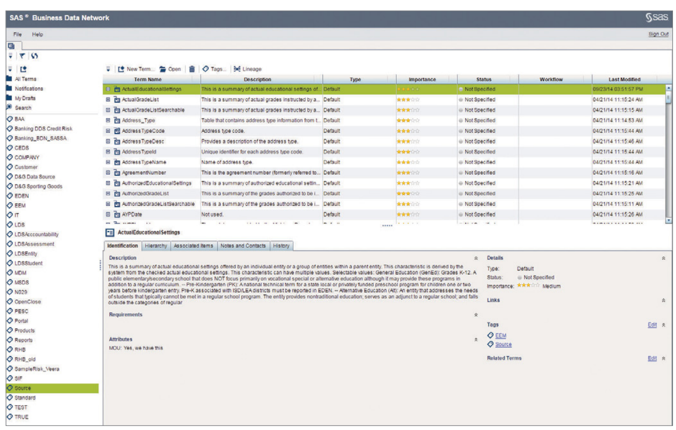

Figure 6: SAS Business Data Network provides a business glossary to relate business concepts to technical metadata like fields, tables, analytical models or reports.

Protecting Fragile Species Through Improved Conservation Funding

World Wildlife Fund (WWF) wanted to find the best way to market to donors, based on their preferences for how and when to be contacted. At the same time, it needed to keep costs under control – because money wasted on ineffective marketing efforts could affect the organization's mission. By using SAS Analytics and SAS Data Management to maximize marketing efforts, the US office of WWF significantly increased revenue from donations while reducing acquisition costs. That means more funding to protect the planet.

Section 3: Analytics, Security and Life Cycle

With high-quality, well-governed data in hand, organizations are ready to make use of an architecture that's dedicated to analytics, visualization and reporting. The best approach is to do this in a way that allows all data to be managed comprehensively, throughout its life cycle.

The data governance program should outline the policies and usability of sensitive and personal data. In addition, data governance should clearly describe what the company considers sensitive data, and identify where this data resides in the enterprise, while also specifying which roles can use and see this data. The governance council should work with an information security officer and/or data protection officer to knit data governance into these roles, ensuring that sensitive data – as well as personal data that's protected under laws like the General Data Protection Regulation (GDPR) – are fully traceable and secure across all big data processing.

Region 13: Data Warehouses and Data Marts

Traditional data warehouses and data marts, such as Oracle or IBM Db2, are being used in tandem with emerging Hadoop repositories to provision large volumes of data for big data analytics. One can access that data using SAS/ACCESS and employ SAS data sets and SAS Scalable Performance Data Server to store data for analytical processing. For improved performance and governance, SAS can run model scoring inside Hadoop (Cloudera and Hortonworks distributions), IBM Db2, IBM Netezza, Oracle, Pivotal (previously Greenplum), SAP HANA, Teradata Aster, Teradata Enterprise Data Warehouse and SAS Scalable Performance Data Server.

Region 14: Big Data Analytics Governance

Governance is crucial for all types of analytics efforts. This is particularly true in an age where big data models are used with advanced analytics techniques like artificial intelligence (AI), machine learning and natural language processing. Without analytics governance, organizations face a range of risks, ranging from competitive sabotage of models to the problem of data models that are built but never used. To make full use of the time and effort spent preparing data and building models, companies must learn to trust their data and build repeatable analytics processes.

As the complexity and numbers of analytical models multiply, analytics, IT and compliance teams need to be on the same page about the topic of analytics governance to ensure legitimacy and integrity of models. Important questions to consider here are:

- What is the process to accept new model requests?
- What is the process to sign off on a new use case for an existing model?
- Has the legal department signed off on the acceptable use of a new attribute in a model?
- What is the process to change an input variable in an existing model?
- What is the process for business sign-off on a new model?

SAS Model Manager is a web-based product that streamlines the process of inventorying, managing, administering, monitoring and publishing analytical models. SAS Workflow Studio is fully integrated with SAS Model Manager to manage and track workflow tasks within the same user interface. SAS Decision Manager augments SAS Model Manager by providing decision-building capabilities that combine analytical models and business rules into decision flows as well as providing automatic discovery capabilities to suggest business rules based on the data. These data-driven business rules can be used to improve decisions previously based only on analytical models. SAS Decision Manager and SAS Model Manager support the critical capability to put analytics and decision into action by publishing both for different execution targets, including real-time, streaming, in-database and in-Hadoop environments.

Bolstering Insurance Data Protection While Building Customer Trust

Interamerican, the largest private insurance company in Greece, has a vast amount of data to handle and protect. Interamerican used the opportunity presented by GDPR compliance requirements to bolster data protection, build customer trust and retain its loyal customers. Under its new data governance model, employees operate in a more secure, efficient way. New tools, expertise and knowledge are used not only to comply with regulations and protect the data, but also to advance data management practices by providing enhanced capabilities for data analysis, data quality and data handling.

Region 15: Big Data Security and Privacy

As privacy takes center stage around the world, many organizations are trying to understand how they can effectively identify, manage and secure all the personal data they process. SAS for Personal Data Protection includes five capabilities that help organizations protect data, regardless of format or location:

- **Access**, blend and assess data from many different file types, relational data sources like Oracle and emerging big data technologies like Apache Hadoop.
- **Identify** and extract personal data from structured and unstructured data sources using data filters, sampling techniques and sophisticated algorithms.
- **Govern** data while enforcing policies, monitoring data quality and managing business terms across the organization –assigning owners to terms and linking them to policies or technical assets like reports or data sources.
- **Protect** data using role-based data masking and encryption techniques to secure sensitive information – and dynamically blend the data without moving it – to minimize sensitive data exposure.
- **Audit** information to pinpoint problems and identify potential issues quickly, then give authorities interactive reports to prove compliance. Identify users, files, data sources and types of personal data, show who has accessed personal data, and provide details of how it's being protected.

Region 16: Big Data Lifecycle Management

Data profiling and data quality analysis provide the ability to understand and categorize the usefulness of the data and how it relates to other data sets. SAS Business Data Network can be used to store or point to policies related to data source onboarding, how and when to archive data, what stage the data is in its life cycle, and who owns the decision rights around that data. Additionally, SAS can relate these business concepts to the specific technical metadata (for example, a data job) that could be used to archive data from one source to another.

Region 17: Cloud

SAS Data Management can access cloud-based data from many cloud environments, including Amazon Aurora, Amazon Redshift, Amazon Elastic MapReduce, Azure SQL Database and the cloud versions of traditional databases. This list continues to grow. SAS software can be deployed in cloud environments such as Amazon Web Services, Microsoft Azure, Google Cloud Platform, OpenStack, VMWare and other infrastructure-as-a-service providers. SAS also provides its own cloud hosting environment. SAS Data Management products can be bundled with other SAS solutions and hosted on the SAS Cloud. The SAS Cloud handles the details of installing, configuring and administering SAS environments so organizations can focus on using insights from analytics to run the business.

How SAS® Data Stewards Can Protect Data: An Example

SAS offers several ways to help data stewards address requirements for information security. For example, SAS Business Data Network allows data stewards to classify a data element as "Public," "Internal," "Confidential" or "Highly Confidential." SAS Federation Server supports heightened security with dynamic data masking and encryption, and ensures that the right users have access to the right data. This is useful when working with data, such as claims adjuster notes, that could contain protected health information subject to privacy guidelines, like the US Health Insurance Portability and Accountability Act.

Summary

With the amount of digital data likely soaring to a trillion gigabytes by 2025, there's little doubt that data scientists and other data professionals will continue to rely on the power of SAS to discover new insights. Consider streaming data – which continually expands in size and scope as big data gushes from smart devices like automobiles, electric grids, factory equipment and wearables. Through its comprehensive approach to governing, managing and analyzing all types of data, SAS is well-positioned to accommodate this data evolution. Learn why the marketplace recognizes SAS as a leader in uncovering and amplifying value from the unprecedented data deluge by visiting sas.com/dm.

107968_G76308.0718