SAS®: A Comprehensive Platform for Big Data Governance, Data Management and Analytics

Sunil Soares
January 2016
## Contents

SAS® Spans Data Management and Analytics ................................................................. 1
Big Data Governance Considerations .......................................................................... 1
Big Data Reference Architecture .................................................................................. 2
Section 1: Ingestion ....................................................................................................... 3
Section 2: Cleansing, Integration and Governance ...................................................... 4
Section 3: Analytics, Security and Life Cycle ............................................................... 7
Summary ..................................................................................................................... 9
Today we are drowning in data from sources such as social media, call transcripts, smart meters and digital pictures. In its 2011 Digital Universe study, *Extracting Value From Chaos*, IDC estimates that the amount of information in the digital universe exceeded 1.8 zettabytes (1.8 trillion gigabytes) in 2011 and is doubling every two years.\(^1\)

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 (such as insurance actuaries) had been using SAS for decades to solve complex statistical problems. Beyond this core capability, SAS offers a comprehensive platform to support the data management, analytics and governance of big data – whether that data is in-stream, in-Spark, in-database or in-Hadoop. Recently, SAS and Hortonworks, a provider of Apache Hadoop, have partnered to create the Data Governance Initiative\(^2\) to help organizations meet enterprise big data governance requirements. SAS is also partnering with Cloudera to combine and integrate with SAS Data Governance technologies for an integrated view of metadata.\(^3\)

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.** This includes clickstream and social media data such as Facebook, Twitter, LinkedIn and blogs. Companies will increasingly integrate this data with master data and with 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. The data governance team needs to establish a number of 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 also need to establish retention policies around the massive volumes of IoT data that can be 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. Data governance challenges also apply to this data. For example, telecommunications companies need to set acceptable use policies regarding the use of GPS data from their wireless customers. For example, should the company send an SMS containing a discount coupon based on the proximity of a subscriber to a specific retail outlet?

4. **Biometrics.** Biometric information includes fingerprints, retinal scans, facial recognition and DNA. Advances in technology have vastly increased the available biometric data. Law enforcement, the legal system and intelligence agencies have been using this information for a long time. However, biometric data is increasingly available in the commercial arena, where it can be combined with other types of data such as social media. This opens up new business opportunities as well as several 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, email, paper documents, surveys and electronic medical records. This data may contain sensitive information that needs to be masked. Companies also 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).

![Figure 1: A comprehensive big data reference architecture should allow for three main sections: Section 1: Ingestion; Section 2: Cleansing, Integration and Governance; and Section 3: Analytics, Security and Life Cycle.](image-url)
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, including Impala, EMC Greenplum, Hadoop, HP Vertica, IBM PureData System for Analytics, Pivotal HAWQ, SAP HANA, Teradata and Teradata Aster.

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 down to the database 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 and stock prices, and 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. However, this real-time data also needs to be governed like any other information. Figure 2 shows a currency feed in SAS Event Stream Processing. Business users and data scientists can govern term definitions and store allowable values using other components of the SAS Data Management platform. 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 hundreds of thousands of events per second.
**Region 4: Databases**

While SAS can read and write from more than 60 relational and big data sources, 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 sets contain information that needs to be governed by SAS Data Management. SAS Business Data Network should manage the business glossary for attributes in these data stores. In addition, SAS Data Quality or SAS Data Loader for Hadoop should be used to 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. 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.

**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

![Figure 3: SAS Data Loader for Hadoop provides a web-based, guided user interface to query, transpose, transform, cleanse and profile data in Hadoop.](image)
Zenith Insurance Uses SAS® to Centralize Data and Prepare for Solvency II

Working with SAS Data Management, Zenith brought all the company’s data into a SAS enterprise data warehouse. High-quality data helped streamline reporting and offered management a clearer view of the business. “SAS is instrumental in helping us achieve a more accurate view of our data, allowing us to move faster and be more competitive,” said Jason Cabral, Group Head of Pricing and Management Information at Zenith.

Regions 7 and 8: Big Data Discovery and Big Data Quality

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.

Region 9: 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

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.
the visualization of relationships and impact analysis to see how changes to one data element might affect information in other systems (see Figure 5).

**Region 10: Information Policy Management**

Data scientists 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 imported into SAS Master Data Management. Users can route and track data quality issues through a remediation interface.

**Region 11: Master Data Management**

SAS Master Data Management supports multiple master 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. SAS Master Data Management was organically built on top of the market-leading SAS Data Quality – and includes profiling, address verification, standardization, data enrichment, transformation, cross-field entity matching and business rule monitoring.

SAS Master Data Management includes SAS Business Data Network and SAS Lineage to view lineage and
Texas Parks and Wildlife Department (TPWD) Uses SAS® to Improve Customer Experience and Grow Revenues

Challenge
With nearly 100 state parks and historical sites, 51 wildlife management areas, eight fish hatcheries and numerous field offices statewide, the Texas Parks and Wildlife Department (TPWD) is charged with managing and promoting the use of millions of acres of park lands and waterways. However, 85 percent of Texas’ population is concentrated in nine urban centers and was spending as much as 90 percent of its time indoors. As a result, department revenue was largely flat.

Solution
TPWD partnered with SAS to develop the License Utilization and Revenue Enhancement System (LURES), a comprehensive data management and analytics application that provides a 360-degree view of the customer – parks, hunting, fishing, boating, magazine subscriptions and more. For instance, using GIS and SAS Analytics, analysts uncovered opportunities to develop the popularity of fishing among urban demographic groups, specifically middle-class Hispanic families in the Houston metropolitan area. This allowed TPWD to expand its popular Neighborhood Fishin’ program (which regularly stocks 17 strategically selected lakes near urban populations with catfish and rainbow trout) by stocking a Houston lake that was closer to neighborhoods with higher concentrations of Hispanic families.

Results
TPWD used SAS Data Management to integrate and improve the quality of customer data. For example, SAS data quality routines helped TPWD identify a flawed business process that resulted in nearly a quarter of the addresses in its license database to have an incorrect county designation, among other typical data problems (such as typographical errors made during data entry at retail locations). TPWD used SAS Data Quality for data profiling, spelling corrections, address standardization, matching and geocoding to fix incorrect location attributes. The department also used SAS Metadata Server to improve and enforce auditing and policy compliance.

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 them to manage all of the data comprehensively, throughout its life cycle.
Region 12: Data Warehouses and Data Marts

Traditional data warehouses and data marts such as Oracle or 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. Finally, SAS can run model scoring inside DB2, EMC Greenplum, Oracle and Teradata for improved performance and governance.

Region 13: Big Data Analytics and Reporting

SAS Visual Analytics, SAS Visual Statistics and SAS High-Performance Analytics products provide advanced capabilities for big data analytics, visualization and reporting. However, data scientists, data management, risk management, legal, compliance and other business and IT users are often disconnected from various aspects of the governance of analytical 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 creating, managing, administering and monitoring analytical models. SAS Workflow Administrator 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 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.

Region 14: Big Data Security and Privacy

SAS Business Data Network allows data stewards to classify data from an information security perspective. For example, a data steward may classify a data element as “Public,” “Internal,” “Confidential” or “Highly Confidential.” SAS Federation Server also supports heightened security with dynamic data masking and encryption by ensuring that the right users have access to the right data. For example, claims adjuster notes may contain protected health information that is subject to privacy guidelines based on the US Health Insurance Portability and Accountability Act.
**Region 15: 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, we 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 16: Cloud**

SAS Data Management not only manages cloud-based data, but its web-based services architecture also lends itself well to cloud-based executions. SAS solutions can be deployed to Amazon EC2, Rackspace or other platform-as-a-service providers. SAS Data Management products are also bundled with various SAS solutions hosted on SAS Cloud. Customers have the following options for cloud deployments: bring your own license (installed on Amazon Web Services, for example); enterprise hosted – a SAS solution that is hosted in a SAS environment and supported by SAS; and remote managed services for those who want to run SAS within their own data center (or their private cloud), but still want SAS to remotely manage the environment.

**Summary**

With the amount of digital data doubling each year, and the number of connected things expected to rise to 25 billion by 2020⁴, 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, or data in motion. Continually growing in size and scope, this big data will eventually gush from countless smart devices – stoplights, soda machines, dishwashers and exercise bikes, to name a few. SAS offers a robust platform and comprehensive approach to big data governance, data management and analytics that can accommodate the data evolution. Learn why the marketplace recognizes SAS as a leader in uncovering and amplifying value from this unprecedented data deluge by visiting sas.com/data.

---

**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.

© 2015 Copyright Information Asset, LLC. All rights reserved.

THIS MATERIAL MAY NOT BE REPRODUCED, DISPLAYED, MODIFIED, OR DISTRIBUTED WITHOUT THE EXPRESS PRIOR WRITTEN PERMISSION OF INFORMATION ASSET, LLC.

Product or company names mentioned herein may be the trademarks of their respective owners.

This report is for informational purposes only and is provided “as is” with no warranties whatsoever, including any warranty of merchantability, fitness for any particular purpose, or any warranty otherwise arising out of any proposal, specification, or sample.

---

⁴ gartner.com/newsroom/id/2905717
SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.