Paper 4915-2020

Take Control of Your Data!

Diane Hatcher, Core Compete

ABSTRACT

As organizations embrace analytics, data is increasingly taking center stage in the modern enterprise. Users are finding innovative ways to build analytics into their daily processes, using data from warehouses, data lakes, and external sources to feed into their models. Often, the data sets and insights they fuel are stored in sandboxes and personal workspaces, inaccessible to others that might benefit from them. The proliferation of these data islands inhibits collaboration and creates storage and governance concerns.

In this emerging Age of Analytics, organizations need to:

- manage the proliferation of data sets to control storage costs
- enable users to find the most relevant data assets in a timely manner
- monitor access across data marts and sandboxes
- maintain good stewardship and governance across the enterprise

Powered by SAS® Visual Investigator, an Analytics-Code and Data Registry (ADR), is a solution to help businesses understand, share, and gain control over their data and analytics ecosystem. The solution provides the framework for a knowledgebase of analytics and data assets supported by integrated analytics project governance, compute environment provisioning, and application-aware workflow and alerting tools.

The ADR solution is a collaboration between SAS and Core Compete that enables businesses to build a data-driven, knowledge-sharing culture of governance and collaboration to empower better analytics and drive innovation.

INTRODUCTION

For organizations invested in SAS solutions, the usage of SAS spans the gamut from ETL, analytical data preparation, reporting, and analytics. The largest volumes of SAS workloads tend to focus on the former – ETL and analytical data preparation. SAS capabilities for data management have always been one of the strongest features of Foundation SAS. Users can take advantage of this power, combining SAS data step and PROC SQL to solve virtually any type of data challenge.

A side effect of this flexibility can be the creation of code and data "sprawl". SAS assets (code and data) are spread across the organization as team and individual user content silos.

For IT, however, this creates challenges for managing the SAS environment in two ways. First, the growth "sprawl" has ramifications for managing costs of the environment. IT must handle constant requests for additional computational and storage resources. Without insights into the details of how SAS is being used (and what is no longer relevant), it is virtually impossible to make rational decisions on managing existing infrastructure. As a result, IT can only solve user needs by purchasing more hardware – servers and storage - increasing the total cost of ownership (TCO) of the SAS environment.

Second, as regulations around data privacy and auditability continue to grow organizations are realizing it is critical to understand the scope and nature of the content in the sprawl. They need a deeper level of understand how SAS is being used in order to provide better governance and manage TCO. The U.S. Federal Government also passed a law requiring an inventory of data assets.

Gartner[1] confirms this– organizations are increasingly looking to deploy some type of information governance model to address user-generated data sprawl. The goal is to have an inventory of information assets to make them more accessible, understandable, and governable. A solution must be able to support information governance and manage data sprawl and pipelines – i.e., to take control of your data.

ANALYTICS CODE AND DATA REGISTRY (ADR)

Core Compete is working with SAS to create an Analytics Code and Data Registry (ADR) solution to catalog and provide insights into your SAS workloads and data assets to help organizations take control of SAS-enabled "sprawl". The ADR solution builds upon SAS Visual Investigator (SAS VI) to provide a powerful user interface to allow organizations manage their SAS assets in a robust and actionable manner.

SAS VI is an interface originally built as a generic application framework to support case management, network analysis, and search to work (Figure 1). SAS Fraud and AML solutions take advantage of this framework to provide a way to understand connections between different entities that drive fraudulent activity. The ADR solution also has similar business requirements, so using SAS VI as the foundation makes a lot of sense.

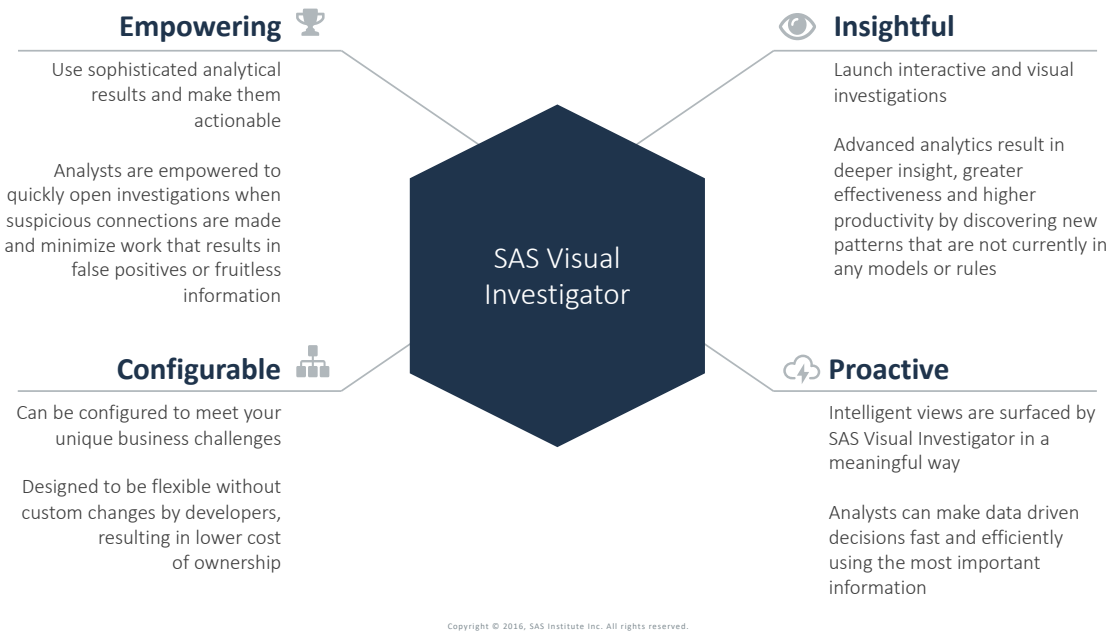


Figure 1. SAS Visual Investigator Benefits

ADR ARCHITECTURE

The underlying architecture for ADR combines SAS VI with SAS® Viya® and Elasticsearch, a distributed open source search engine. Elasticsearch's ability to index content provides the ability to query with speed and scalability (Figure 2).

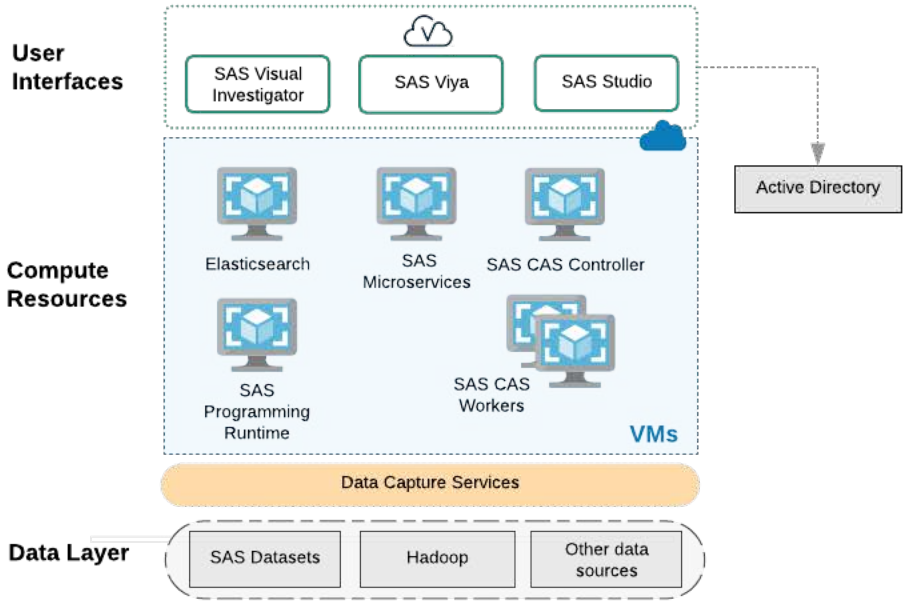


Figure 2. ADR Architecture

Core Compete's extensions add additional capabilities to the ADR, including:

- Data capture services to automatically gather metadata about the data assets
- SAS workload analysis for understanding usage patterns
- Knowledgebase data model and entity definitions
- Customer-specific customizations for workflow, alerts, and reporting

These specific extensions are not detailed in this paper but feel free to contact the author to find out more.

ADR FUNCTIONALITY

The purpose of the registry is to:

- create a searchable knowledgebase of your SAS workloads and data assets
- support governance around SAS usage by tracking SAS usage at a granular level
- provide alerts when anomalies are detected in SAS workloads and/or data pipelines

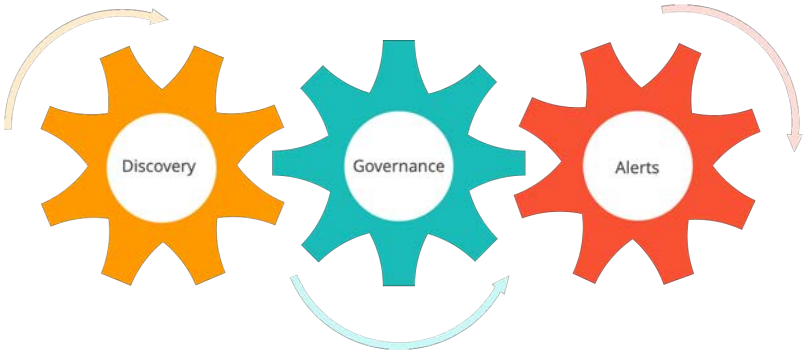


Figure 3. ADR Functionality

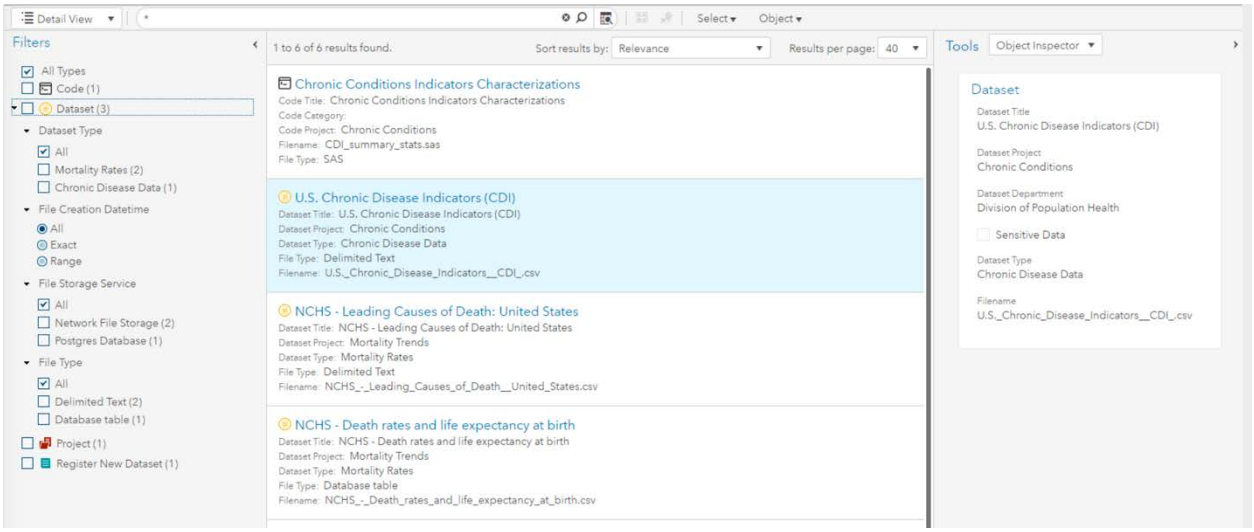
ADR KNOWLEDGEBASE

At the core of the ADR solution is a searchable knowledgebase that contains data about your SAS data assets and workloads. Entities define high-level objects that can be searched against. For ADR, we are focused primarily on entities like Dataset and Workload. The key attributes we are tracking include:

- Creation of new instances of each entity type
- Characteristics of each instance compared to expected policies
- Relationships between the entities

ADR DATASET ENTITIES

In addition to standard attributes one would expect around datasets, such as name, location, and data size, we also capture attributes that are specific to SAS and help with understanding and managing the overall storage footprint for the environment.



Display 1. ADR Dataset Attributes

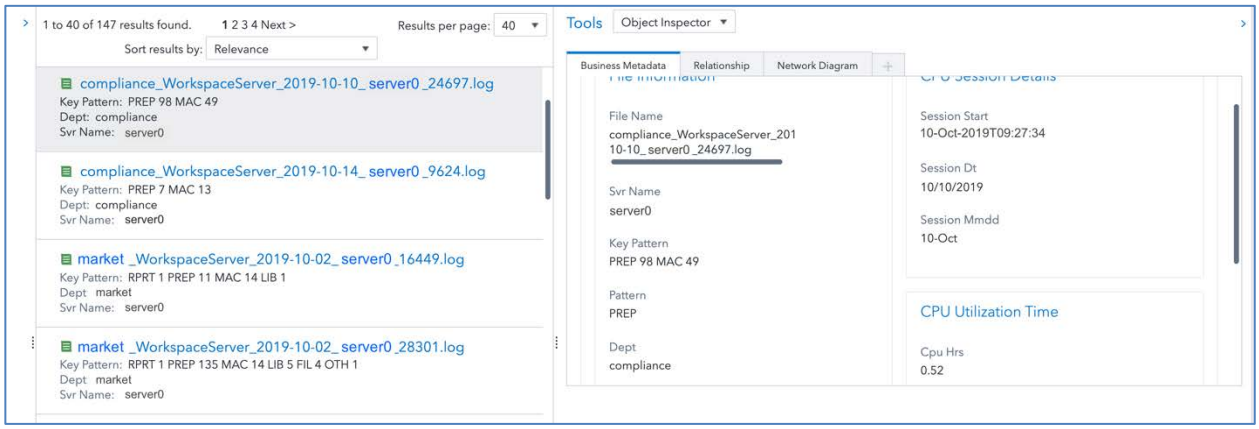
Additional attributes include:

- Length of a dataset record, number of observations
- Variable formats and sizes
- Table compression

Furthermore, we apply ML techniques to score the dataset metadata to determine whether there are any red flags for potential policy violations, such as PII data exposure or duplicate data. In these cases, we can also add the ability to capture deeper characteristics based on data profiling algorithms.

ADR WORKLOAD ENTITIES

For workload entities, we can analyze SAS log files to gather and store information about the SAS workloads that create and read these dataset entities. SAS logs provide rich details around the actual execution steps, performance, and data access beyond what is available in SAS code files. Moreover, it is possible to resolve macro variable references, allowing us to get a more complete picture.



Display 2. ADR Workload Attributes

Workload entities are critical for understanding SAS usage at the granular level; details needed to fully manage the SAS environment. Core Compete can breakdown SAS workloads to precisely understand:

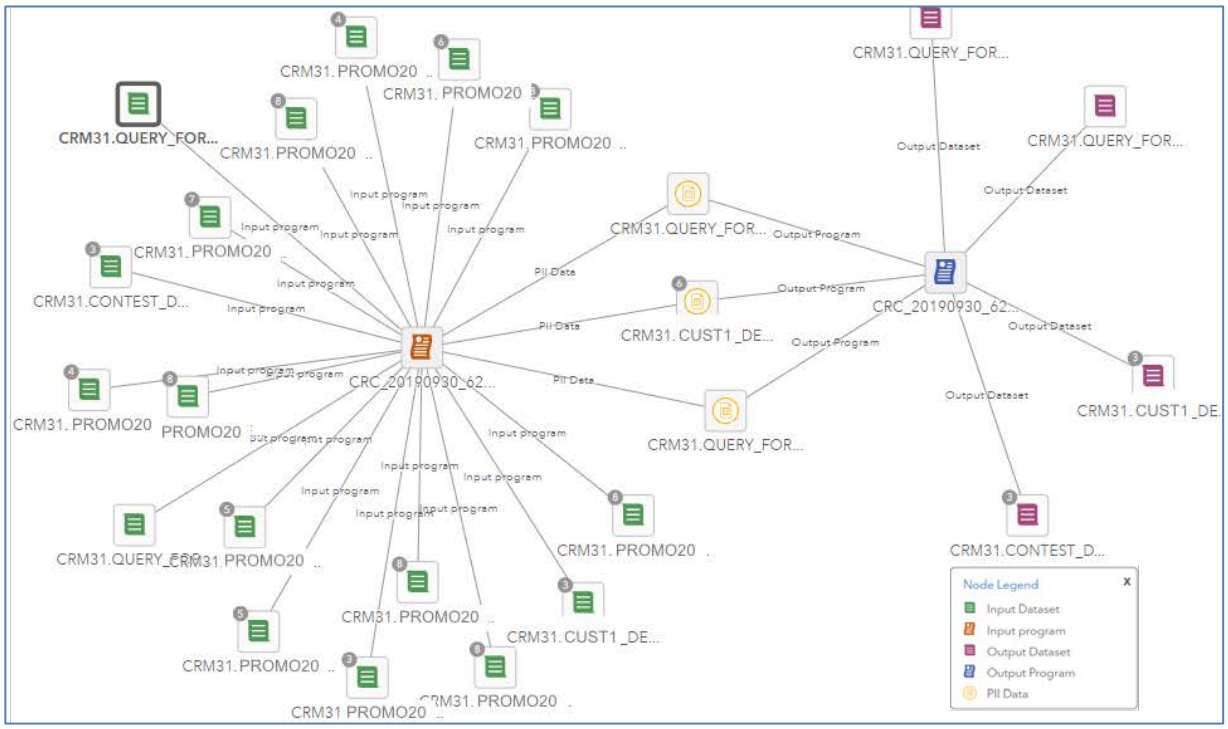
- Which specific SAS workloads are consuming the most system resources?
- What are the characteristics of the workload – steps taken, data sources accessed, etc.?
- What data sources are being read and written within a workload?

ADR GOVERNANCE

The ADR can leverage native SAS VI capabilities to provide a governance framework for SAS assets. The ADR is an important tool in support of regulatory compliance and data- driven decision-making.

The ADR knowledgebase creates a common catalog for your SAS data, which helps the business to better understand and manage the data being stored and shared in SAS datasets. And, importantly, it provides a collaboration environment to enable business and IT colleagues to have a common frame of reference around the organization's SAS assets.

In addition, the ADR provides data usage transparency with lineage and impact analysis. Data pipeline traceability can be displayed to show how the SAS data is getting generated and which workloads are creating and updating the data.



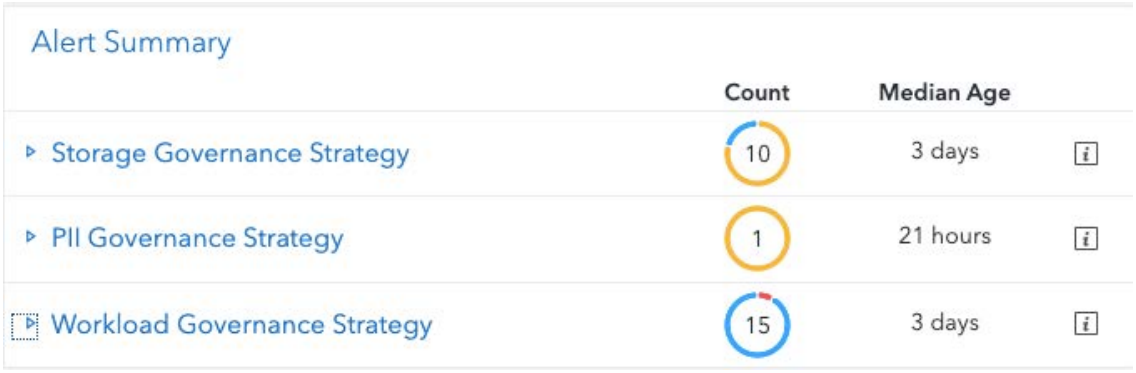
Display 3. ADR Data Lineage

Data lineage is especially important if ADR detects that PII data may be present in the SAS data. The ADR can score the table and column metadata to determine if PII data is present and flag that data for review by IT. Linked PII data (data that could directly identify an individual) should be, at a minimum, encrypted and access limited to specific users. Linkable PII data (multiple data fields when combined can be used to identify an individual) should be detected and managed, as well. Having the data lineage information helps IT and business stakeholders understand where the data is coming from, so appropriate actions can be taken at the source, if necessary.

ADR ALERTS

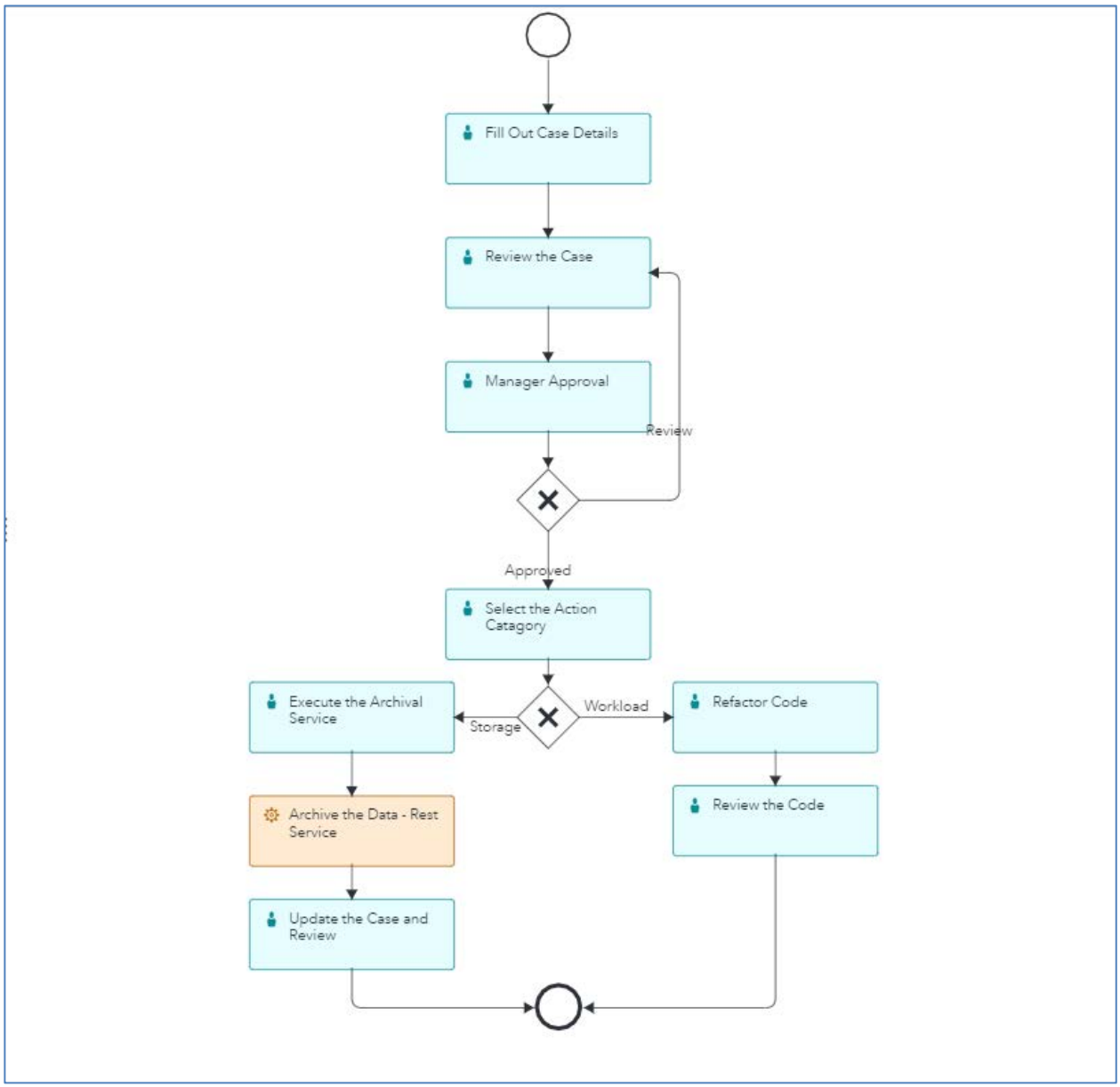
One of the key strengths of the SAS VI solution is the ability to create workflows and alerts that make ADR's output actionable to the organization. Exceptional behavior and anomalous SAS data can be automatically alerted to stakeholders. SAS VI supports a rigorous case management process to help resolve these issues with full transparency.

Alerts can be grouped under different subjects and scored with different priorities. An alerts dashboard highlights open governance cases, so you can easily find and investigate new, unexpected behaviors or issues. Alerts can be defined for data anomalies, workload performance issues, or for data integrity, data security, data quality, etc. Alerts are routed to the appropriate teams for resolution and resolved through a best practice workflow to guide teams through their resolution.



Display 4. ADR Alerts Dashboard

Workflows can be defined to support multi-step processes that involve multiple stakeholders. So, for example, if new SAS datasets are being created in the production environment, a workflow can be triggered to ask the user to provide additional metadata about the data – such as business purpose, update policy, etc. – then submit the information to IT for approval. IT can then run data validation processes to ensure that the data is not duplicate nor contain sensitive information. Workflows can also be created to request access to data, or have users sign up as consumers of data to be notified of changes to data.



Display 5. ADR Workflow Example

CONCLUSION

A 2017 Gartner survey on data and analytics trends identified "data risk and information governance" and "deriving value from data" as two of the top three most noted challenges. Data “sprawl” issues have continued to grow as organizations grapple with how to manage and govern user-generated data without impacting their ability to complete analytics and reporting projects.

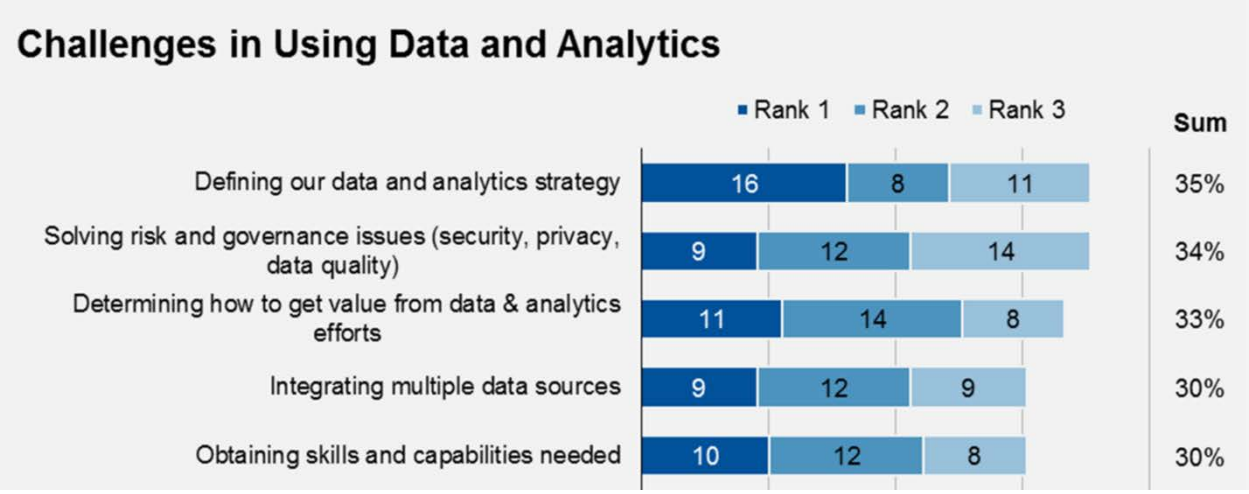


Figure 4. Source: Gartner (December 2017)

The joint solution from SAS and Core Compete of an Analytics Code and Data Registry, leveraging SAS Visual Investigator, provides a compelling solution for cataloging and governing SAS data and workload assets. ADR provides:

- The ability to have data stewards define and endorse governance policies, encouraging the democratization of data usage through sharing and collaboration
- Governance processes for SAS-driven assets to manage the growth of permanent data storage and system resources
- Automated tracking of data usage to surface insights on what data is most useful

The ADR solution provides the framework for a knowledgebase of analytics and data assets supported by integrated analytics project governance, an application-aware workflow and alerting tools. The ADR solution allows you to take control of your data!

REFERENCES

[1]Zaidi,E., De Simoni, G., Edjlali, R., and Duncan, A. <December 13, 2017>. “Data Catalogs Are the New Black in Data Management and Analytics” *Gartner*, ID: G00338777.

115th Congress (2017-2018). “H.R.4174 - Foundations for Evidence-Based Policymaking Act of 2018”. January 14, 2019. https://www.congress.gov/bill/115th-congress/house-bill/4174.

ACKNOWLEDGMENTS

Special Thank You! to Bryan Goodliffe from SAS and Kumar Majety from Core Compete for their work and collaboration on the ADR solution and this paper. Teamwork makes great work!

Thanks also to Rick Thompson of Core Compete for his keen eye for details.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Diane Hatcher
CORECOMPETE
+1-919-410-6203
Diane.Hatcher@corecompete.com
https://corecompete.com