Paper SAS1824-2015

Bust Open That ETL Black Box and Apply Proven Techniques to Successfully Modernize Data Integration

Brandon Kirk, SAS Institute Inc., Cary NC

ABSTRACT

So you are still writing SAS® DATA steps and SAS macros and running them through a command-line scheduler. When work comes in, there is only one person who knows that code, and they are out—what to do? This paper shows how SAS applies extract, transform, and load (ETL) modernization techniques with SAS® Data Integration Studio to gain resource efficiencies and to break down the ETL black box. We are going to share the fundamentals (metadata folder organization and naming standards) that ensure success, along with steps to ease into the pool while iteratively gaining benefits. Benefits include self-documenting code visualization, impact analysis on jobs and tables to identify objects impacted by change, and being supportable by interchangeable bench resources. We conclude with demonstrating how SAS® Visual Analytics is being used to monitor service-level agreements and provide actionable insights into job-flow performance and scheduling.

INTRODUCTION

Every great success has a story that will get you on the edge of your seat. Since this is a story of data management modernization practices, I cannot promise that you will be on the edge of your seat, but I will share techniques to make your story of modernization a success.

Business and IT strategies are aligned more than ever as they pertain to a digital information process that enhances business effectiveness and customer satisfaction. This emphasis and the SLAs around executing the bi-modal strategy are becoming highly regarded. Real-time information insights from solutions that deliver on the digital business strategy require detailed planning, integration, and operational steps to ensure that internal and external collaboration are robust, agile, and meet fast changing business demands on a global scale. The challenge of big data with the velocity of information and real-time decisions are constituting modern data management practices.

In this paper we discuss the challenges of implementation, the move to more automated methodologies for integration, a collaborative approach to keep the business and IT service experience real and meaningful, and some dynamic examples of how SAS uses this approach for better business practices. The techniques shared are from SAS IT experience working with SAS® Data Management solution and specifically the SAS® Data Integration Studio interface for interactive data integration development. The primary technology features that enabled our modernization approach are SAS® code import to GUI based jobs, metadata management, and integrated change management with code promotion.

THE BLACK BOX

Whether we are talking about data management to support warehousing, business intelligence query and reporting, or data management for analytics, there is always that one person who builds and runs code that produces numbers that no one else can reproduce. If that person is on leave, it is a mad scramble to figure out where the code lives, what it is doing, and how did it come up with that answer? We all know there is a better way, so we force documentation that lasts less than a month and is out of date. Why do we put up with this black box coding mentality? This mentality requires a deep bench of specialized critical experts. If you can't have dedicated subject matter experts, then your entire business support model struggles.

Is it because that is the way we have always done it? Possibly it is the perception that if we add too much process or a "tool," then our senior programmers will be less efficient. These are folks who can transform data or write complex procedures before having their first cup of coffee. We do not dare place any barriers to their creativity and code-generating techniques. All these points seem like the magic black box is the only way. Well, it's not. Not even close. What if with a few steps you can create an ecosystem

with SAS® Data Integration Studio that opens up visibility to data flow, impact analysis on change to objects, secure content and data, all supportable by interchangeable bench resources.

DATA MANAGEMENT MODERNIZATION GOALS

The modernization buzz for data management is having an auto documented enterprise data management environment. This modern approach provides the following critical benefits in today's climate of data-driven decision culture. Know the impact of change. Support agile needs with bench mentality. Repeatable data transforms for code reuse and productivity enhancements. Produce more with less. More efficient. More reliable. Visibility into DM schedule and ability to dashboard against SLA along with optimizing resources.

It does take a mentality shift, but it does ***not*** take a huge leap. You can take small steps and show incremental benefits as you move into a modern approach to data management. When implementing change you always want to embrace one area where someone can retain a comfort zone. So for our senior SAS coders we do not ask them to ditch their advanced certified programming syntax from memory on day one. We just want them to surround their code with data inputs and data results as targets. As we grow to see the benefits of impact analysis, then team members start embracing the ability to see lineage and visual comprehension to the data flow. Naturally, job processes start to show multiple nodes to visualize steps and capitalize on concurrent processing capabilities. As a peer, business colleague, or manager we can avoid the days of the mad scramble because there is a visual roadmap to what is happening in the code flow. The deeper we embrace the more we start to see additional benefits. From the manager's perspective, additional individuals can contribute interchangeably to code support, so you can move to a bench resource mentality. As a peer, covering vacations or working in a new subject area is not a scary thought. You can find your way around easily knowing the information you are seeing and reading is current.

WHAT ARE THE BENEFITS?

As a result of our data modernization efforts our group has realized the following direct benefits:

- Reduces number of support calls, especially outside of office hours
- Reduces number of missed SLA data delivery time
- Simplifies upgrades for new SAS versions
- Provides production data integration process monitoring/alerting
- Provides documented data integration standards with auto-documented data flow
- Promotes interchangeable knowledge experts
- Provides change management
- Provides revision control
- Supports bench resource mentality
- Provides security with Metadata authorization along with auditability
- Provides impact analysis
- Enables integration and governance for big data

HOW DO YOU GET THERE?

Making the move to modernize your data management is ***not*** a big bang approach. It is an incremental progression with realized benefits for each iteration. The characteristics of each iteration are shown in the figure below. These characteristics best define the mental facets that you will encounter as you progress through each phase; each is just the mindset that you apply to a particular data process or a collection of processes. It is important to note that these are not necessarily sequential and, it is natural that some areas of processing move more quickly than others to leverage all capabilities and benefits.

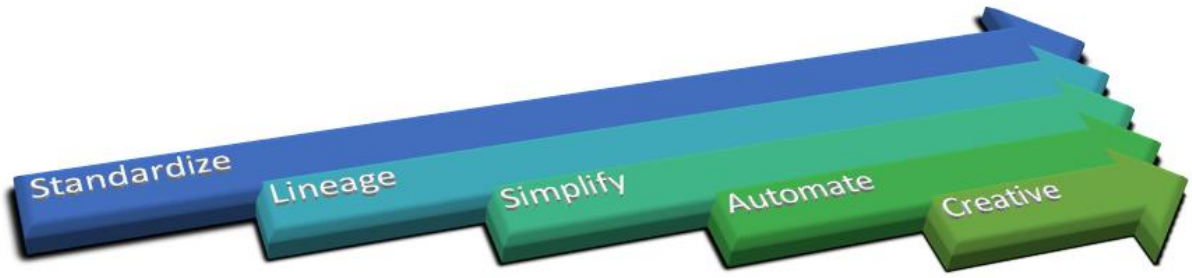

Figure 1. Modernization Characteristics

There is one characteristic, Standardize, that is the foundation of modernizing data management and one that will continue to grow and evolve throughout the entire modernization effort. The Standardize characteristic ensures that everyone is on common ground in the environment when it comes to location and naming of content. I recently had this same mental shift and evolution of standards with my family's media files. We had pictures, videos, and music saved randomly in folders. It was a nightmare to find anything and costly to backup. A simple exercise of agreeing to a folder organization and naming and methodology has reduced the time we spend saving, locating, and backing up our media.

Characteristic	Description
Standardize	A standardization template is the key to being proficient traversing metadata and aligning definitions of terms. Each organization should develop a metadata organization and naming standard. Plus, common naming definitions provide quick interpretation of an object. This is living documentation that will continue to evolve the further you proceed. For an example of our internal foldering standards please refer to **Appendix A**.
Lineage	The main goal in the lineage phase of the methodology is to ensure that you can run the processes you run today within a modern data management tool. Start small by placing code in user-written transforms. Next, surround the processes (transforms) with all the data inputs and targets. The goal is to keep the code that is used today but augment it with visual data flow, process monitoring, and impact analysis.
Simplify	Break down larger processes into smaller, specific deliverable processes. Leverage transforms, which show data flow through the process. The goal is to gain performance increases through concurrent and threaded processing.
Automate	Continue to build out process flows. Build custom transforms that enable your data management goals. There might be elements to your data flow or typical data requests that could be developed as a common tool for multiple processes to leverage. Custom transforms provide that foundation of building plug-and-play code that is customized for solving your business requests.
Creative	The creative mode is when you are now reaping the benefits of being in a proactive environment. You are innovating and revolving data delivery because you have more insight into data and analytic initiatives.

WHAT IS NEEDED TO SUCCEED?

There are a few things that will help along the way to ensure adoption and provide the skills to be successful. The most critical piece is to have the people behind the goals. There can be some reluctance to change. Often, reluctance is motivated by a lack of understanding of why there is a need to evolve what is perceived as working. Without those change agents that help with adoption, momentum can be slow. Here are some suggestions to facilitate understanding and boost momentum.

Embrace the concept of metadata. Make sure there is understanding of what that means and that metadata is a wealth of information and protection. There are huge benefits to having the metadata layer, but without understanding and adoption, using the metadata for insight is lost and can appear burdensome.

Since the metadata layer is so critical, define and organize your metadata content to promote readability, adoption, and understanding. There are huge efficiency gains with blueprints on how metadata is organized and named. Then use that same framework across all areas for ease of use and consumption. There is no wasted effort to reinvent or redesign as processes are developed and implemented because a flexible template is already available. We started with a simple ETL naming standard that has grown into a set of comprehensive, living development standards that cover all the metadata objects, such as tables, jobs, flows, and reports used in our environment. The goal is for everyone to have the same methodology and naming when performing like tasks. Pulling data from any order entry system and pulling data from a problem ticketing system are very different, but if each of those processes are named and stored in metadata similarly, then I know both exactly where to find those jobs and the purpose of the object. There are many flavors of metadata organization, so you really have to think through what is best for your organization. You can find a number of publications that cover different naming schemes but the most important thing is to come up with standards that work for your environment.

Figure 2 shows a brief sample of a few of our naming conventions. Every programmer knows that jobs and tables that begin with the PRE_ prefix represent jobs that extract data and target tables that are the results of that extraction, respectively. Likewise, jobs and tables prefixed by STG_ represent transformation jobs and targets from transformation. Similarly, LOAD_ prefixed job and table objects are the jobs that load a finalized target table. Of course, they could be named with prefixes such as EXTRACT_ or TRANSFORM_. There is no magic in the actual name used, but in the fact that the purpose and location is known and understood automatically by the programmers who use them.

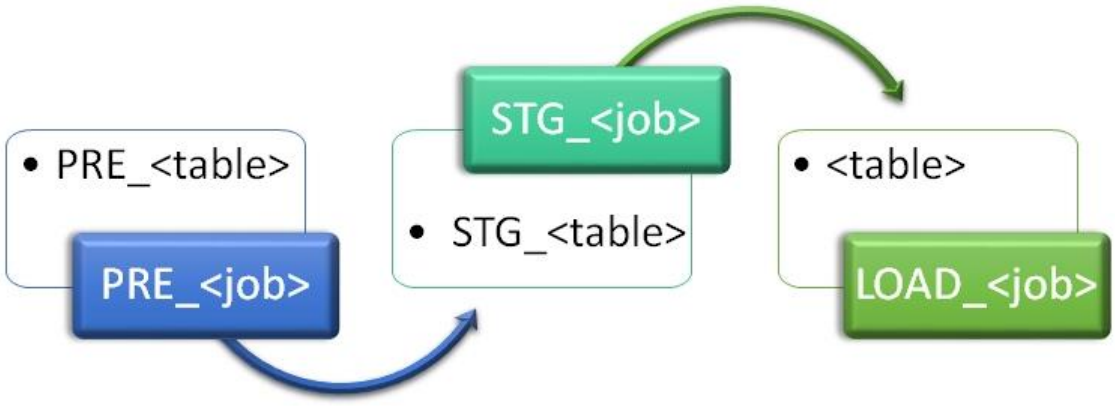

Figure 2. First Step in Our Standardization Process

MANAGE BIG DATA

The arrival of distributed computing with open-source technologies is leading to an overabundance of distributed processing and data flow programming interfaces. Competitive business is requiring data driven decisions and machine learning through advanced analytic techniques and modeling. Data delivery and analytics strategies are designed with an assumption of data abundance with the capabilities of distributed computation and in-memory analytics. These areas need the same management methodology as the traditional data warehouse. The immense volume, velocity, and variety of big data

makes it even more imperative that we maintain a focus on the characteristics that provide an environment consistent with data modernization techniques.

Being a group that embraced data management we had a solid foundation. As mentioned in the standardize characteristic, your data management standards will be living and evolving. We introduced new objects and definitions to our standards but not a new way of processing. This enabled us to focus on integrating data to unlock the intelligence potential of relating traditional warehouse data with unstructured and semi-structured information. We have continued to embrace the capabilities to manage, integrate, and govern metadata on distributed data through SAS Data Integration studio. Now we are able to leverage comma delimited files stored on Hadoop and integrated with traditional RDBMS data. The ability to enrich and iterate through the data management process is a large portion of an analytic lifecycle.

DATA MANAGEMENT FOR ANALYTICS

The cornerstone of self-service BI is a solid understanding of how to balance the analytical environment and data flow with the governance process. Our data management foundation has given SAS IT the ability to accelerate our entire business into data visualization with SAS® Visual Analytics while retaining a high level of governance. This is where solid automation methods come into play for IT environments, and where those methods that are serviced by IT to line-of-business professionals have been mutually agreed upon via open dialog, requirements processes definition meetings, and questionnaires. At SAS, in particular SAS IT, we have established an analytic on-boarding process with five general areas.

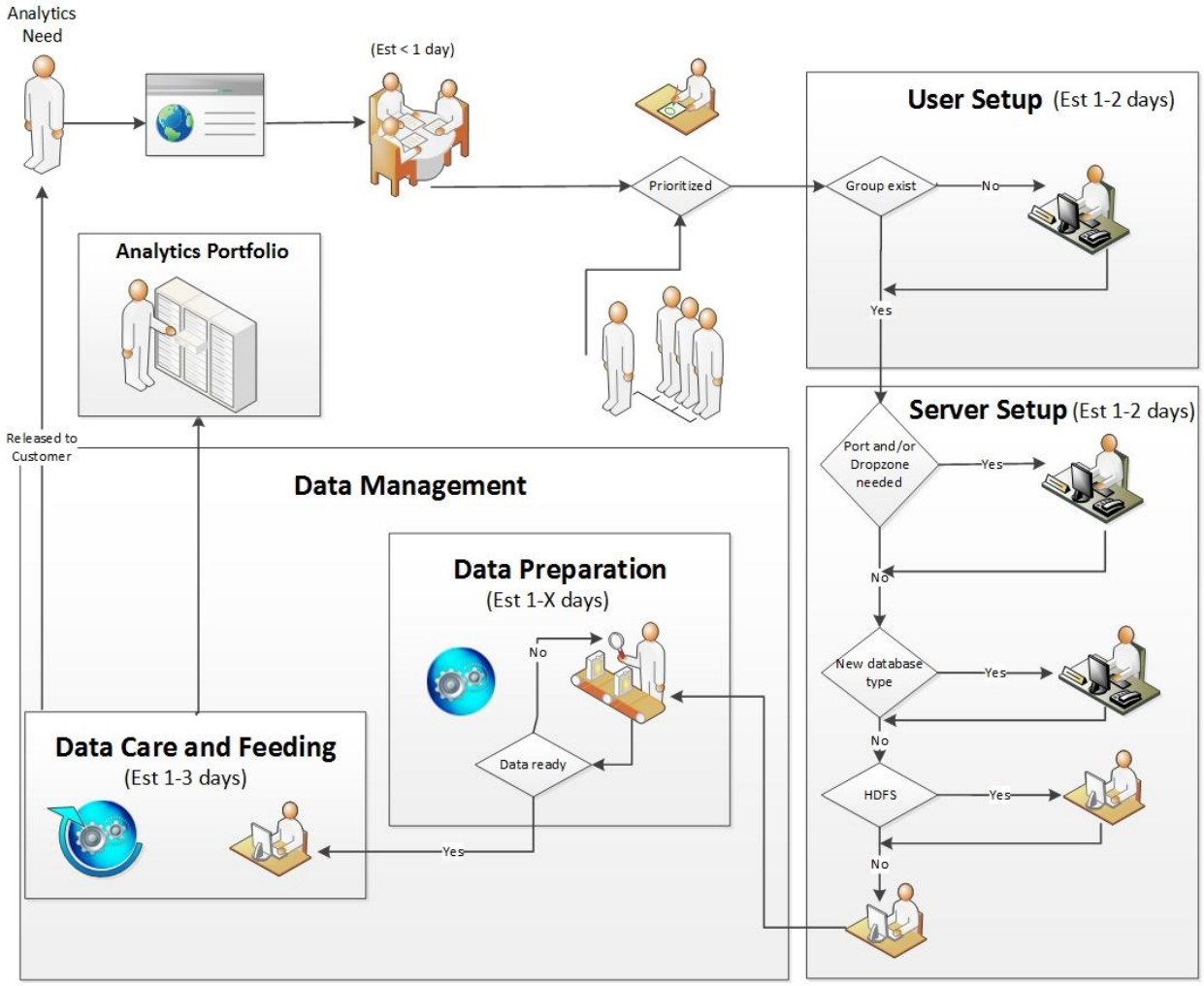

Figure 3. Analytic On-Boarding Process. Since we are putting structure around the data management

process, this enables us to ensure that the business is aware of the lifecycle and the timing to expect as we work through these phases.

Step	Day Estimate	Description
User Setup	1-2	Define the security requirements and setup permission groups based on the business desires on report/data access and administration.
Server Setup	1-2	The business data subject drives the need for data access and LASR needs. At this point, the information about source data location is determined (Warehouse, Business Data Mart, Spreadmarts, or operational system) then decisions are made to have a push/pull mentality into the Analytics server.
Data Preparation	1-x	Now that the source and target locations have been defined, a data integration expert can prepare the data for analysis. Business and IT can explore the data and iterate through the additional data needs (quality, metrics, format, additions).
Data Care and Feeding	1-3	Once solidified on the data needs, update the data on a business-defined schedule with enterprise-monitored execution.
Analytics Portfolio	1	Ensure that the request is highlighted as a completed analytic subject area in the portfolio for the enterprise. Add analytic ROI to show the value of analytics to the organization. Time might be needed to show tangible ROI but always make it a point to come back and quantify the analytics value.

THE BALANCE OF ANALYTICS AND GOVERNANCE

The majority of analytic initiatives have had data flow freedom. The goal is to retain business continuity while leveraging digital resources efficiently. To establish clarity in and around the digital balancing act it is important to know that the management and governance is more of an iterative process that happens on a need basis versus a sequential process that happens at a pre-defined cadence. When designing data for analytics there will inevitably be additional metrics to enhance analysis. Consider the following examples:

- When bringing data together for analyzing SAS Café data we pulled in information about the purchase and purchaser. Next, we added geographical coordinates for the Café and office locations to see travel patterns. We then added climate data to see the effects of weather on buying patterns.
- When bringing data together for analyzing weblog data we pulled as much information about the visitor as we could leverage. We added geographical coordinates and language preferences. We added profile data and a touch point history. There were data quality enhancements to parse URLs for behavioral patterning.

Each analysis leads to additional data management efforts. As we have referenced above, with a modern approach to data management, the iterative enhancement just relies on the next available data integration expert who can easily augment with additional data requirements. Having this common approach also enables the data scientist to easily find and enhance data for analytic modeling efficiently.

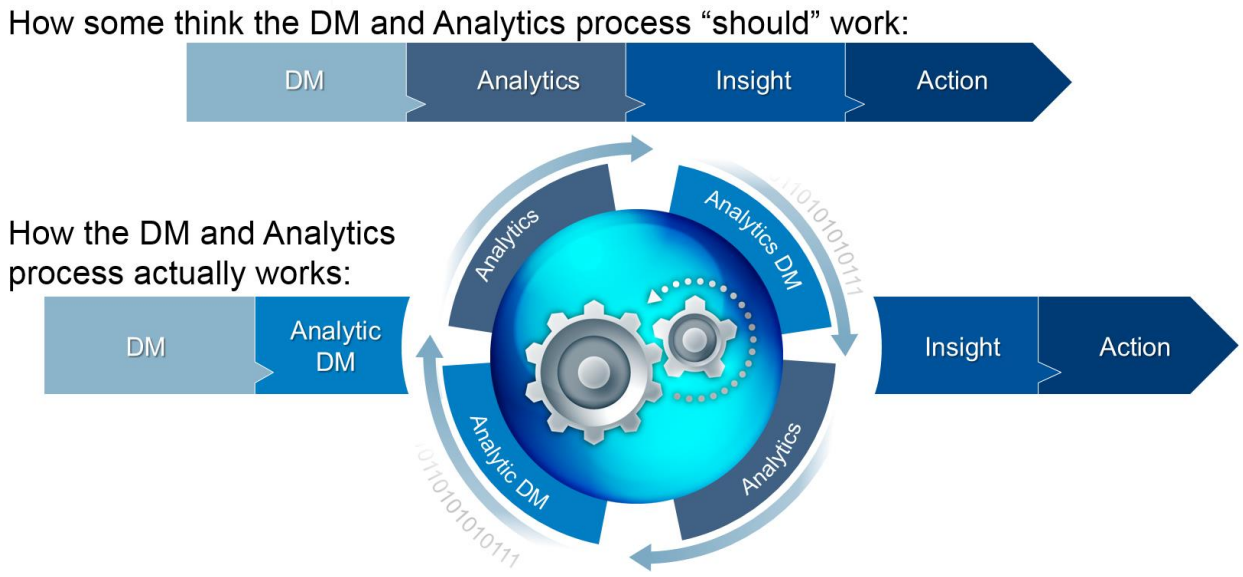

Figure 4. Iterative DM for Analytics

BUSINESS VISIBILITY

When the business submits a request for analytics where does it go? Does the business have insight into what is going on or is their request lost in the IT black hole? SAS IT has automated the request to fulfillment workflow to ensure that the business is aware of the process and the progress on analytic data management requests. If 80% of an analytics lifecycle is in the data management, we want to streamline that as much as possible. Now that 80% of time is visible from a high-level dashboard, it can be drilled into deeper for details. This provides a collaborative view into the analytics portfolio and what is in-progress and next on the list.

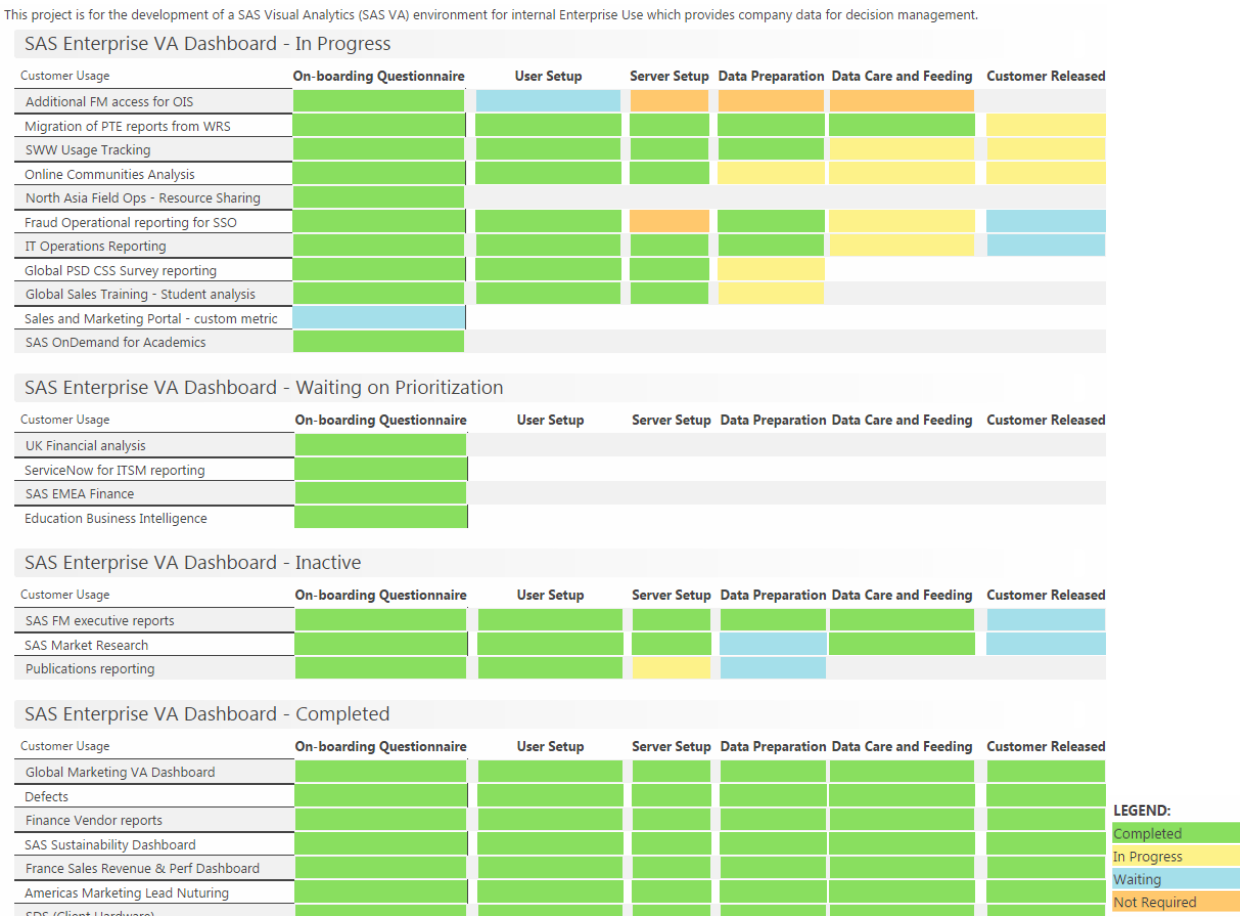

Figure 5. Business and IT remain aligned when visibility into the progress of analytic initiatives is readily visible. This visibility also becomes an input into our analytics portfolio. We reduce duplicate efforts and build upon completed initiatives for greater insights.

OPTIMIZE YOUR DATA MANAGEMENT

For a long time I struggled with seeing the full picture of the data management activities that fell under my responsibility. I spent a lot of time memorizing and orchestrating data dependencies purely off a timing mentality. There were Gantt chart spreadsheets to show typical scheduling and resource heavy periods, which were inaccurate and always out of date and based only on a typical run.

Leveraging your move to modernization will allow you great insights into daily flow, out-of-band processes, and predictive analytics on data preparation service-level agreements. Today I leverage SAS Visual Analytics to provide the following insights from our scheduling logs.

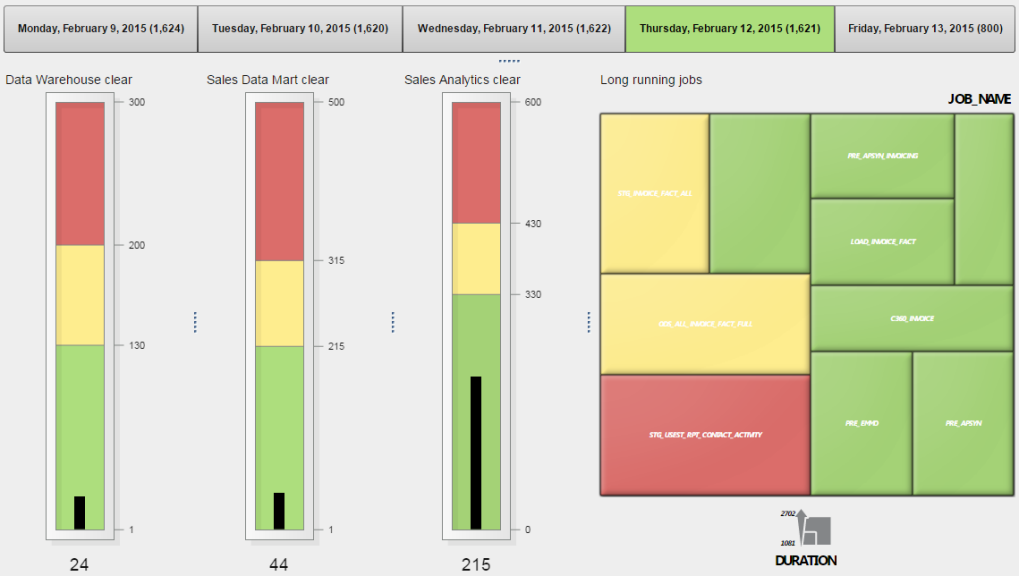

Figure 6. A dashboard provides performance against critical delivery times along with highlighting any long running processes on the day selected.

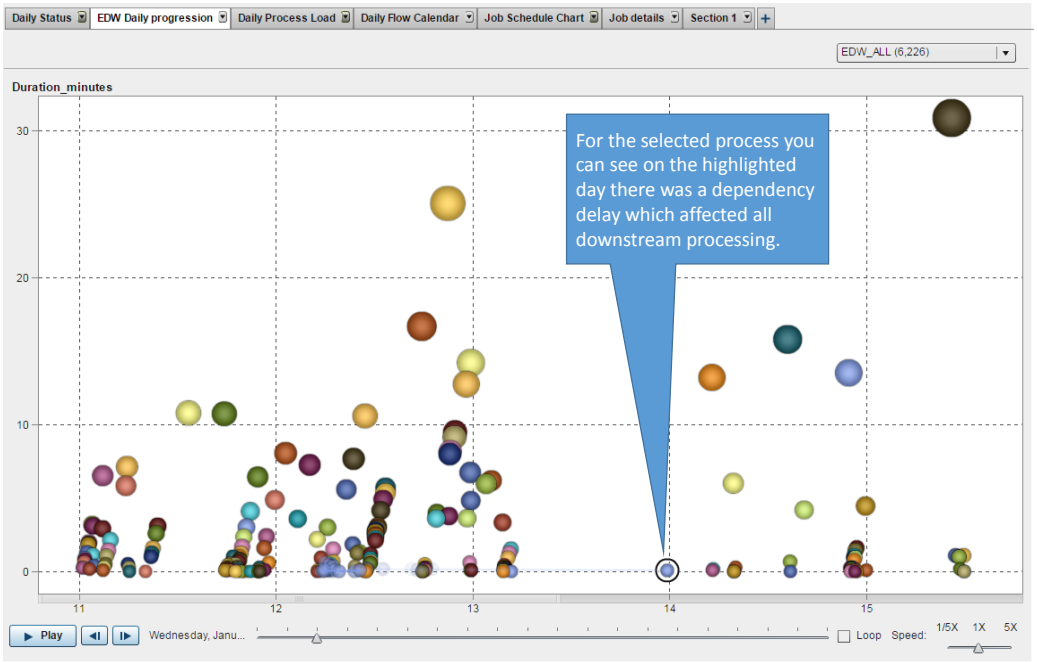

Figure 7. We have a time series bubble plot to put job flows in motion. This visualization enables you to compare process performance and their impacts from day to day.

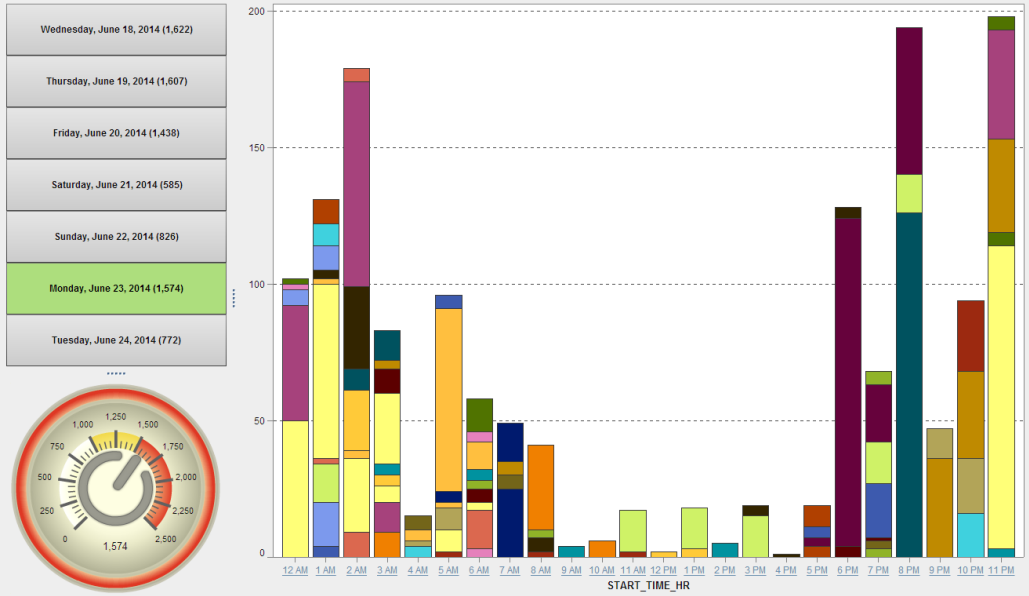

Figure 8. We have the ability to monitor system load and peak hours of data management activity on the selected day. The hours are drillable to see the job flows (the different colors) and the number of processes running by the minute. For an added flare I added the speedometer to reflect the total job processes completed on the environment for the date selected.

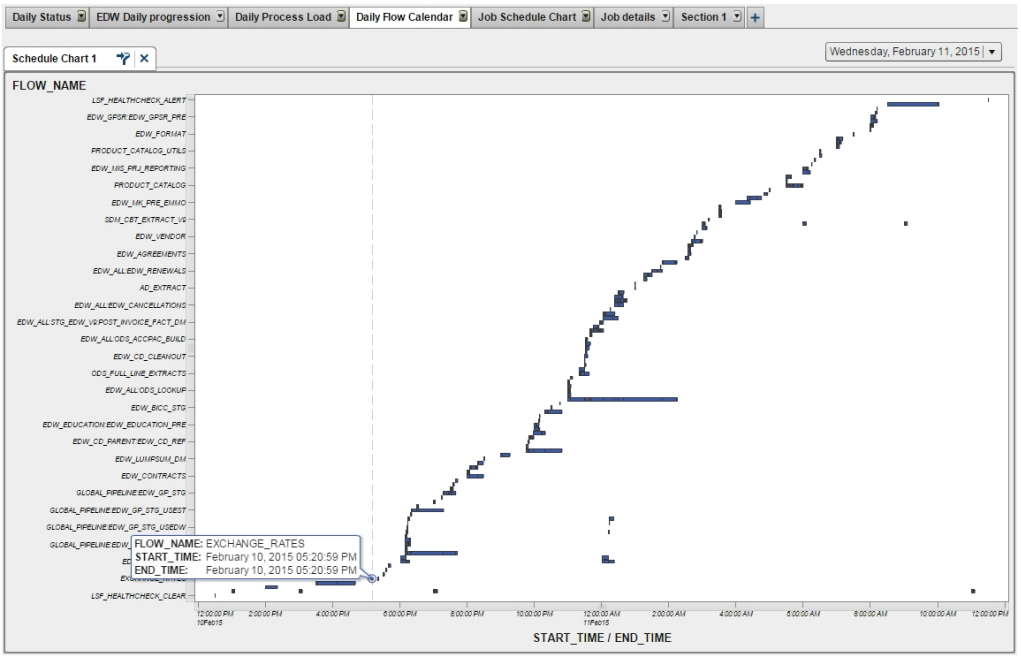

Figure 9. Overall data flow calendar of all job flows and jobs within a 24hr period. The duration and cadence of repetitive flows are evident at first glance.

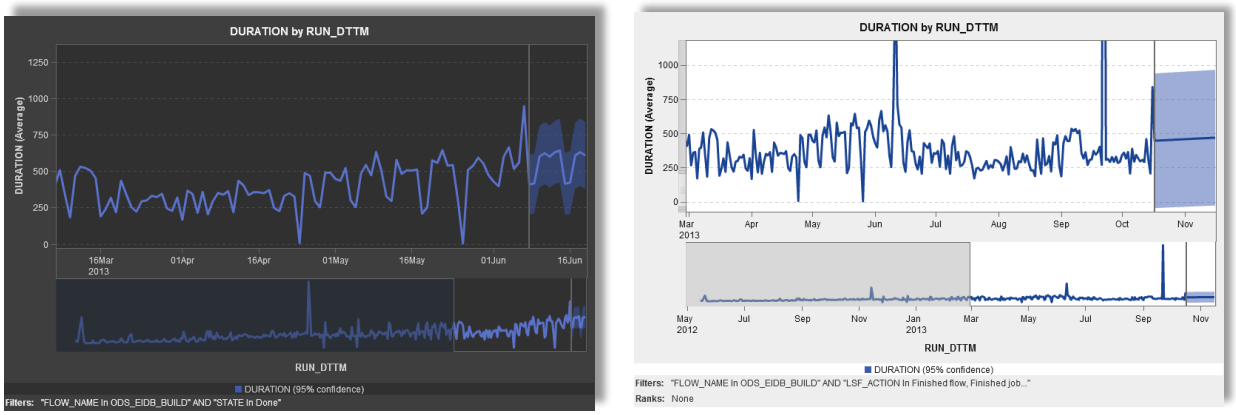

Figure 10. Predictive analysis on job or flow performance. In the figure on the left, we were able to run predictive analysis on a job's performance to predict duration in the future. The job duration was creeping to our high water mark so I was able to proactively assign a resource to address performance.

CONCLUSION

At SAS we implement industry best practices and leverage SAS technologies to drive our business forward. Our experience of modernizing our data management methodologies has successfully improved business effectiveness while gaining resource efficiencies. The modernization characteristics described in this paper provide the foundation needed to move at your own pace during the adoption process while gaining benefits along the way.

REFERENCES

Cearley, D.W. and C. Claunch. 2013. "The Top 10 Strategic Technology Trends for 2013." *The Top 10 Strategic Technology Trends.* The Gartner Group. 12.

Dyche, Jill. 2015. *The New IT: How Technology Leaders are Enabling Business Strategy in the Digital Age*. New York. McGraw Hill.

Dyché, J. and E. Levy. 2006. *Customer Data Integration: Reaching a Single Version of the Truth*. Hoboken, NJ: John Wiley & Sons. 294.

English, L. 2009. *Information Quality Applied: Best Practices for Improving Business Information, Processes, and Systems*. 1st ed. Indianapolis, IN: Wiley Publishing, Inc.

Gidley, S. and N. Rausch. 2013. "Best Practices in Enterprise Data Governance." *Proceedings of the SAS Global Forum 2013 Conference*. Cary, NC. SAS Institute Inc. Available: http://support.sas.com/resources/papers/proceedings13/084-2013.pdf

Grasse, D. and G. Nelson. 2006. "Base SAS® vs. SAS® Data Integration Studio: Understanding ETL and the SAS Tools Used to Support It. *Proceedings of the Thirty-First Annual SAS Users Group International Conference*. Cary, NC. SAS Institute Inc. Available: http://www2.sas.com/proceedings/sugi31/toc.html

Inmon, W.H. 1992. *Building the Data Warehouse*. Boston: QED Technical Pub. Group. 272.

Kimball, R. 1996. *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses.* New York: John Wiley & Sons. 388.

Nelson, G. 2012. "Best Practices for Managing and Monitoring SAS® Data Management Solutions." *Proceedings of the SAS Global Forum 2012 Conference*. Cary, NC. SAS Institute Inc.

Newcomb, C. 2013. "A data governance primer, part 1: finding the root cause." *The Data Roundtable* Blog. Cary, NC. SAS Institute Inc.

Power, E. and G. Nelson. 2008. “ETL and Data Quality: Which Comes First?” *Proceedings of the SAS Global Forum 2008 Conference*. Cary, NC. SAS Institute Inc. Available: http://www2.sas.com/proceedings/forum2008/TOC.html

Redman, T. C. 2008. *Data Driven: Profiting from Your Most Important Business Asset*. Boston, MA: Harvard Business Review Press. 257.

CONTACT INFORMATION

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

Brandon Kirk
100 SAS Campus Drive
Cary, NC 27513
919-531-3825
Brandon.Kirk@sas.com
http://www.sas.com
www.linkedin.com/in/BrandonKirkSAS