Paper SAS5193-2020

Expedite Analytics Delivery and Empowerment with a Common Analytics Base Table

Jennifer L. Nenadic, SAS Institute Inc.

ABSTRACT

One of the most time consuming and challenging aspects of an analytics project is the data preparation, and with each new project, that data preparation effort often starts at ground zero once again. It doesn't always have to be that way! Learn from how the data and analytics organization at SAS has taken the concept of an Analytical Base Table (ABT) to the next level by creating what we call a Common ABT. This Common ABT can be used to jump-start many different analytics projects, significantly reducing the time to delivery while also creating the foundation for analytics empowerment. This technique will help you shift out of the reactive approach to your data preparation and get more proactive and strategic. During this session, we talk about both the technical aspects of the Common ABT approach as well as process-related and people-related recommendations to set your organization up to be successful.

INTRODUCTION

From this paper, you will gain a better understanding of how to architect and implement a technique called the Common ABT, that can be used to streamline and centralize the standard data preparation tasks for analytics models. This paper contains high-level technique guidelines, recommendations, and tips for how analytics empowerment can be started or enhanced using this technique.

Although this paper is primarily focused on data preparation techniques for analytics models that require an ABT, similar concepts could also be applied to data preparation efforts for analytics visualizations and reporting.

This paper is intended for data and analytics leadership (Chief Data Officer, Chief Analytics Officers, and related centers of excellence leadership) who are looking for frameworks and techniques to expedite analytics deliveries and equip analytics organizations. This paper is also relevant for data engineers, data management professionals, data-savvy analysts, and data scientists who regularly prepare data for analytics models.

This paper is not intended to address the following:

- This paper is not a silver bullet nor perfect solution. Productionalizing models and reducing time to delivery for data preparation activities is a well-known and well-documented industry challenge. There is no perfect solution. Although the techniques discussed in this paper address some of the situations encountered, it is not intended to address all your needs.
- This paper is not intended to be a full representation on how to establish an analytics empowerment program
- This paper is not intended to be an in-depth guide for how to create an analytics model
- Because not all analytics models leverage an ABT design, this paper is not intended to support analytics modeling efforts which to not use an ABT structure.

ANALYTICS BASE TABLE APPROACH AND BACKGROUND

WHAT IS AN ANALYTICS BASE TABLE (ABT)?

An ABT is a common industry term for the flat and wide table structure that is often used to support various analytics techniques such as predictive modeling or forecasting. If you have a data warehousing or dimensional modeling background, the simplest expression of an ABT is a fully denormalized star schema as shown below in Figure 1.

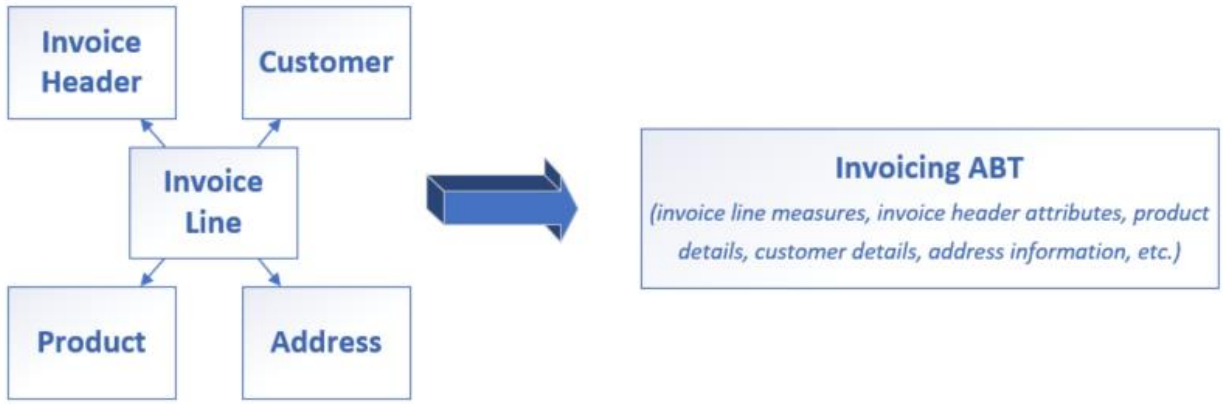

Figure 1 Simplified representation of an Analytics Base Table (ABT)

With that said, ABTs are typically more complicated. They bring together information on various subjects to provide a holistic view. For example, in predicting customer churn, an ABT may include (but not be limited to) the customer purchase history, customer satisfaction metrics, associated customer marketing campaigns, sales representative details and attributes, customer sales pipeline, and many other areas of data. These data elements typically live in a multitude of systems that you must join together and integrate into the ABT structure.

ABTS AND THE HOLISTIC ANALYTICS DATA FLOW

During the analytics lifecycle, data from as many data sources as possible are brought together to create a very wide ABT. Because data scientists and analysts rarely know at the onset which attributes are impactful and which are not, the ABT used to develop an analytics model can often range on the scale of hundreds if not thousands of columns. In the creation of the ABT, attributes provided in the data sources are transformed and exploded out into many different possible metrics. The data is summarized and arranged to support the specific requirements of the model to be developed. This could include summarizing information to a specific time increment, removing unneeded data elements and rows, etc. The result is a highly specialized ABT, as shown at a high level in Figure 2.

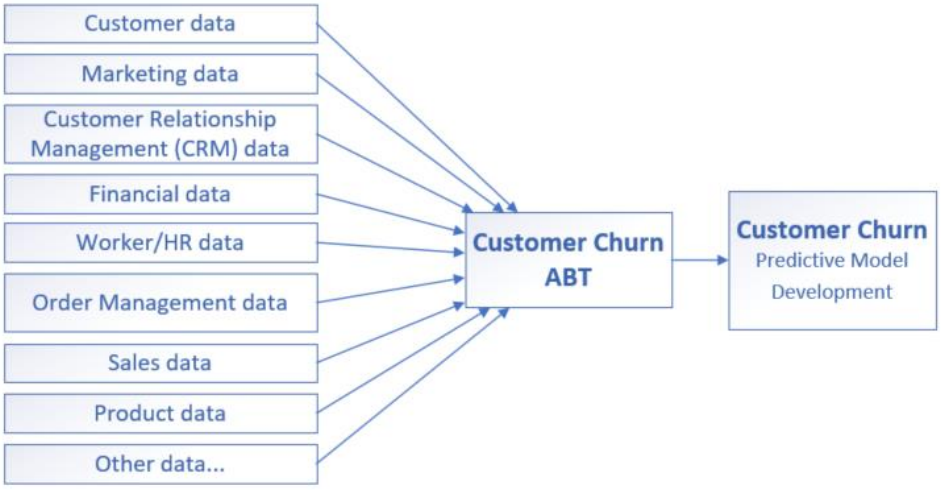

Figure 2 Sample ABT for predictive model development

As the modeling development concludes and the champion model is selected, it is common that most of the columns (and sometimes complete data sources) are found to be inconsequential and not used by the analytics model. For data scientist, these final columns are defined as statistically significant. When operationalizing a model, also called productionalizing, the original ABT created for the analytics model development is reduced in scope to only the columns needed by the published, champion model. In the sample shown below in Figure 3, the data sources are reduced to only those data sources required to produce the actual attributes used by the predictive model. The benefit of this step is that it avoids unnecessary load on the compute tier for columns that are not being used.

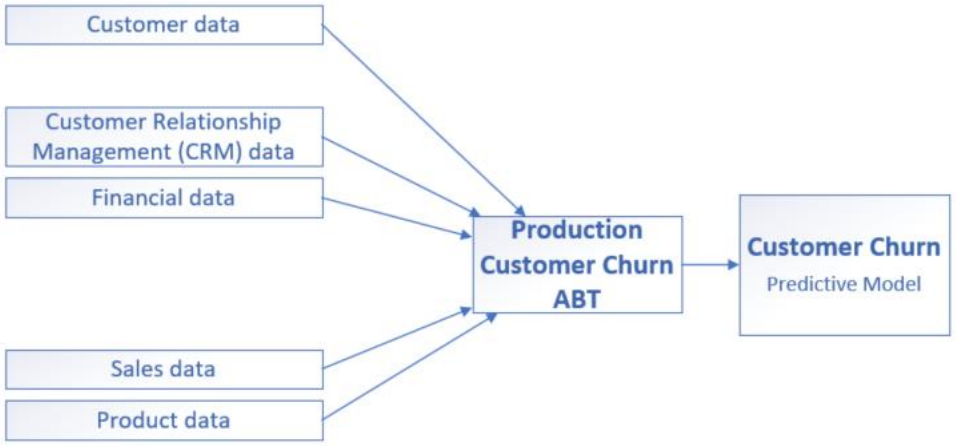

Figure 3 Sample ABT once put into production

LIMITATIONS OF MODEL-SPECIFIC ABT

Because ABTs are developed with the exact attributes, transformed metrics, and summarizations required for each specific analytics model, you can find yourself in a position where each new model must join and integrate data from the same general sources over and over. When this happens, scenarios like those diagramed in Figure 4 emerge.

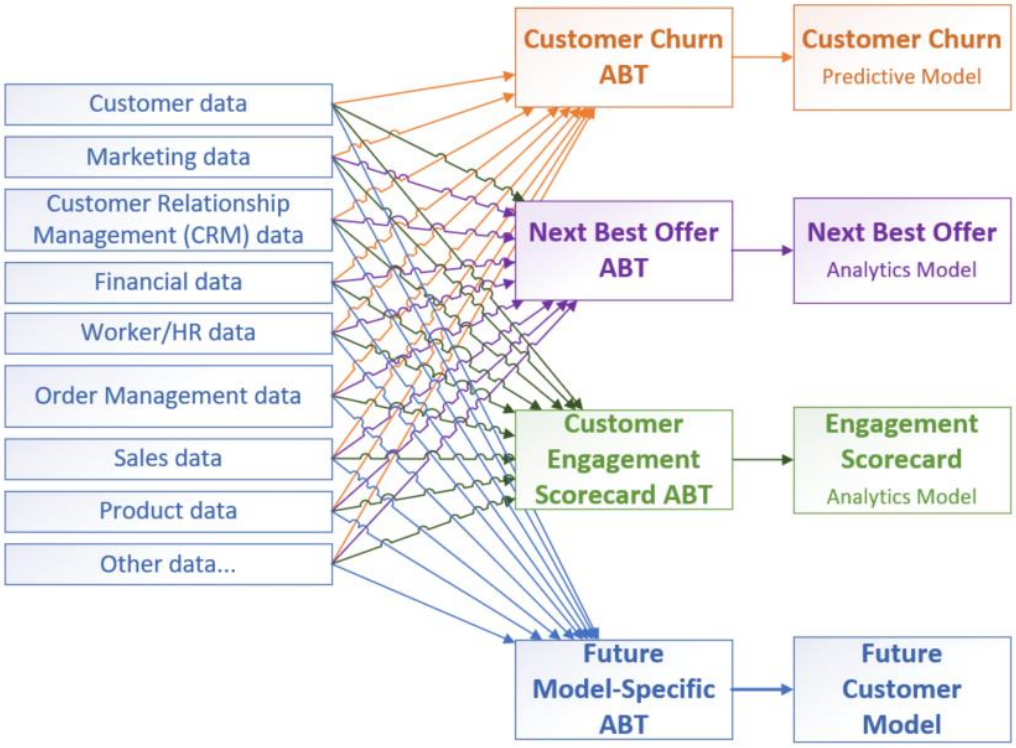

Figure 4 Situations when models rely on similar data sources

This can impact your team as well as other teams if your organization leverages an analytics empowerment model. Specifically, when each analytics effort pulls from similar data sources and repeats similar data integration tasks and logic. Depending on your organization's approach to data access, you may find that some groups are getting the same data from different data locations (whether from the Enterprise Data Warehouse or the source systems directly or other copies of this data). Having every analytics effort implement similar logic means that every analytics effort must start over from the beginning. You can't expedite the data preparation process, as the data fields are put together in a way exclusively for use by the model for which it was built. This is where the Common ABT approach can be a lifesaver!

THE COMMON ABT TECHNIQUE

This section outlines the conceptual approach to the Common ABT and the many benefits it can provide to your organization. The goal and intention of a Common ABT is to perform the joins, integrations, and basic transformations that are commonly found in the ABTs created for specific models. A Common ABT is not intended to replace model-specific ABTs, but to address the redundant code and logic that must be implemented with each new analytics effort. Please see Figure 5 below as an example.

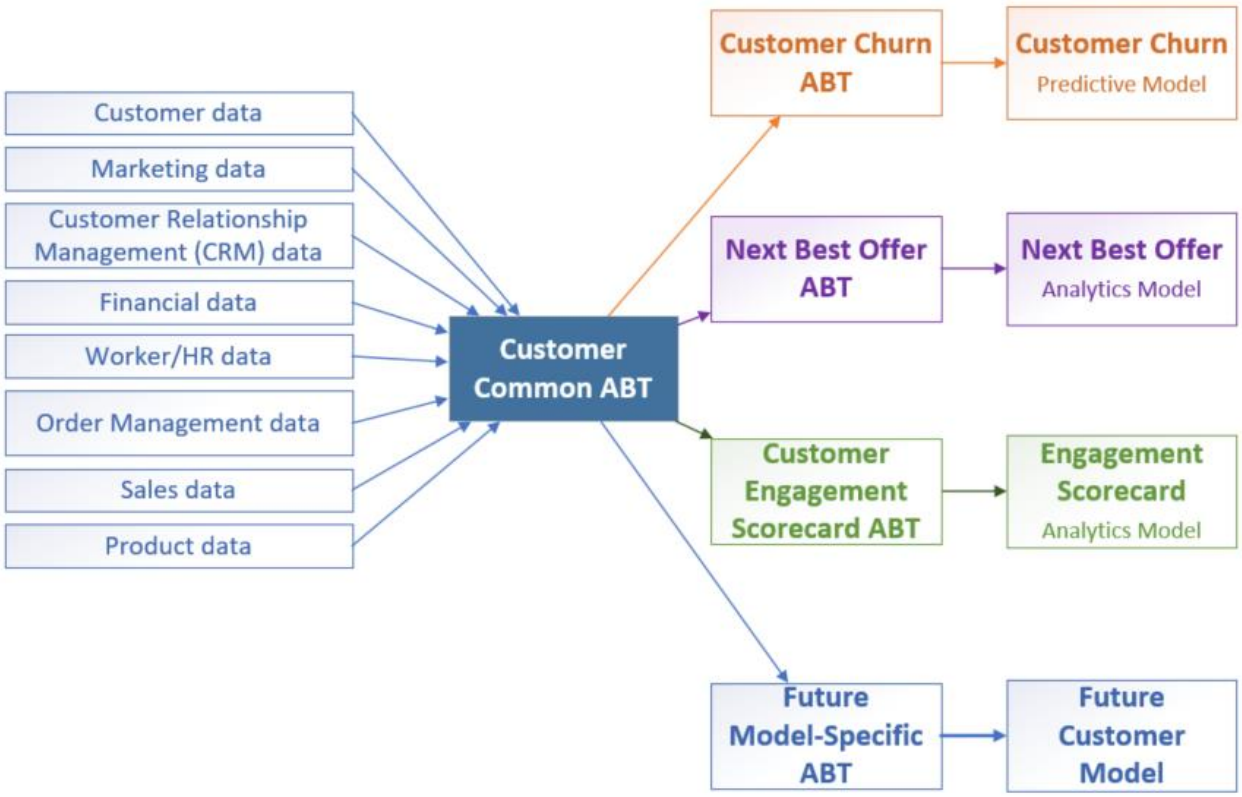

Figure 5 Sample Common ABT used for customer-related analytics efforts

In this scenario, a Common Customer ABT has been created that integrates the data from all the data sources used by the models listed into one centralized ABT structure. This eliminates a significant portion of the time each analytics project must spend identifying data sources, assessing possible metrics, designing how to integrate and join all the data sources, etc.

COMMON ABT OBJECTIVES

Below are some key objectives and expectations of the Common ABT:

Common ABTs become an invaluable data asset in the analytics model development process. As described before, a Common ABT should think on a larger scale than one individual analytics effort. They should include all the standard data sources that analytics projects in this domain (such as customer, worker, product, etc.) may require and then create a fully functioning ABT at the lowest level of granularity expected for analytics needs. See Figure 6 below as an example of a possible Customer Common ABT data structure design where the baseline metrics are all provided at a monthly level. This structure is designed under the premise that all analytics efforts will operate at a monthly level or above.

Customer Key	Month	Marketing metrics	Sales metrics	Financial metrics	Other monthly metrics
ABC	Jan2019	…	…	…	…
ABC	Feb2010	…	…	…	…
ABC	Mar2019	…	…	…	…
ABC	…	…	…	…	…
DEF	DEC2019	…	…	…	…

Figure 6 Sample Customer Common ABT at monthly level

There is a balance to this approach. If you discover that 90% of analytics efforts require data at a monthly level, but 10% require it at a daily level, you must decide if it's more helpful to create multiple Customer Common ABTs. Remember, if the goal is to expedite analytics delivery, you'll have to make some judgment calls on when and where to apply this technique. Building a Common ABT for one use case where it is not believed to have other applications is probably just adding extra overhead that would not be a good use of time; however, if it is envisioned that many other analytics models could take a similar approach, then structuring a Common ABT around those needs allows those future efforts to deliver in expedited timelines.

Further, as you are building your Common ABTs, the goal is to build in some initial intelligence in addition to just combining data sources. Below are some questions to consider when designing a Common ABT (with additional recommendations included in the "Recommendations For Getting Started" section below):

- What are the common data sources?
- What are the underlying baseline metrics and attributes?
  Do not limit yourself to just the data fields. Consider adding intelligence by creating true metrics and exploding out the source system column(s) into other metrics that could be useful for various analytics projects.
- What is the timeline for the data? You may have data going back decades, but that doesn't imply that the data should always be used in analytics efforts.
- What is the typical time element being used for your analytics models?

Lastly, a Common ABT, is executed and refreshed on a scheduled cadence. Unlike model-specific ABTs, whose scope is reduced during operationalization, Common ABTs are a core data asset that is refreshed and available regularly. The columns and attributes created and added to the Common ABT are always available (as long as they are relevant for analytics projects) and the production version of the Common ABT mirrors the version used for model development.

MODEL-SPECIFIC ABT OBJECTIVES

As shown in Figure 5 above, the goal of the Common ABT technique is not to replace a model-specific ABT, but to rather reduce its scope and allow the data prep team to focus primarily on the model-specific metrics and summarizations instead of identifying and joining various data sources.

The primary source of data for model-specific ABTs should be the associated Common ABT. Because every analytics model has different primary requirements, it is expected that the model-specific ABT needs to summarize, column subset, row subset or even transform the data provided in the Common ABT to meet the specific needs of the analytics project.

Further, as your business evolves, there may be situations where the analytics model needs data that falls outside the current scope of the associated Common ABT. This is okay and will happen. As part of the balance in creating a Common ABT that meets standard needs, it is possible that a scenario arises where a need is only partial met. Those projects should use the Common ABT as their baseline, and augment where needed. You do not want to stifle creativity during analytics model development. You may find that the new data sources are relevant in the final champion model.

Just as described in section, "Error! Reference source not found.," the model-specific ABTs are usually reduced in scope when operationalized to only include those columns needed by the champion model. This part of the analytics development and operationalization does not change with the Common ABT approach.

USING COMMON ABT FOR ANALYTICS EMPOWERMENT

There are many possible facets and objectives when considering an analytics empowerment model. The great news is that the Common ABT technique has direct benefits to some of the aspects to consider with an empowerment model. This section will outline some of those considerations.

Below are some of the possible goals and objectives of an empowerment model and how the Common ABT technique can support them.

- Self Service and Enablement – One of the big focuses in establishing an empowerment model is how to effectively provide the necessary capabilities to the groups wanting to perform analytics efforts. Groups that do not know where to go or where to start obtaining the data they need to perform the analytics they require can now have a baseline table already available. No longer do they have to figure out all the data sources, data structures, data joins, etc. Rather, they are equipped with a pre-joined Common ABT with some initial intelligence already incorporated. This is a huge win for analytics groups who may or may not have access to data engineers with the skills to do complicated data integration. Now you've enabled these analytics groups to get started with a big piece of the data preparation aspects already complete.
- Governance and Controls – Another key focus for a successful empowerment framework is establishing the appropriate governance and controls. For a center of excellence, this is paramount. By owning the Common ABT, you make it easy for people to conduct analysis because they must work through your organization to get access. This allows you to set up the appropriate data access governance and controls and ensure that data security and data privacy requirements are being met for any analytics initiative. With GDPR and other legislations arising, you benefit with the Common ABT because you are in control of the data being used.
- Increased Visibility and Awareness – One objective in empowerment, especially for centers of excellence, is to create a knowledge base of all the analytics initiatives (past and present). If your company leverages a dispersed or hybrid organizational model where analytics projects are not centralized in one group, it can be more challenging to keep track of what everyone is doing. The Common ABT technique makes analytics easier and incentivizes groups to partner with you for their data needs instead of trying to accomplish it themselves. By working through your governance and access process (mentioned above), you increase visibility into the analytics efforts everyone is undertaking and can more easily begin to build this treasured knowledge base!

BENEFITS OF A COMMON ABT

The Common ABT technique has many benefits. Below are some you may experience in your organization.

- Time to create a model-specific ABT is significantly reduced

- Models leverage a common single source of truth (SSOT) for their analysis efforts. Analytics teams can focus on what they do best and no longer have to become data source experts.
- As the business evolves and the Common ABT is enhanced with new data sources and new attributes, all analytics teams can immediately benefit.
- If done well, a Common ABT can introduce attributes and metrics the analytics project team may not have considered that could have statistical significance on the model being developed.
- By having a primary source for analytics efforts via the Common ABT, higher visibility and stronger controls on data governance programs benefit for this standardization.
- For analytics centers of excellence, this technique can foster a culture of empowerment for teams wanting to do analytics without having to devote extended cycles on data preparation activities.

RECOMMENDATIONS FOR GETTING STARTED

As you are getting started with the Common ABT technique, we would like to share some tips to help ensure you are set up to be successful.

DURING COMMON ABT DEVELOPMENT

The amount of value your organization experiences with the Common ABT largely depends on the investments made during its development. It is strongly recommended that you create this table structure design in partnership with data engineers, data scientists, business subject matter experts (SME), and data subject matter experts (SME). At SAS, this approach has been used internally and was based on data available in SAS' internal Enterprise Data Warehouse. Analytics in its very nature is a balance of science and art. To get the maximum value out of your Common ABT, you must incorporate a bit of both. This is where the partnerships can be so valuable. Data engineers may not always be able to think of possible metrics, but can share the data that is available. Business SME's and data scientists can help transform the attributes and measures from your data warehouse or data sources into intelligent metrics. For example, in your order management system or data mart, you may have all the products a customer may have purchased. For an analytics effort, this could be exploded out in many ways. This is where the intelligence and ultimate added value can be quickly realized. Please see the list below of examples of columns you might choose to create in your common ABT based off the order management transactional history:

- Column with a separated list of all specific products newly ordered by the customer in this time period
- Column with a separated list of all specific products the customer now has in their possession as of this time period (so new plus existing).
- Individual columns for each product category (and/or product segment, etc.) to include the number of products within this category that the customer ordered in this month
- Individual columns for each product category (and/or product segment, etc.) to include the number of products within this category that the customer has in their possession as of this month

Cardinality can be an important factor in developing an analytics model. Specifically, this means the number of possible distinct values for a column. Having varying levels of cardinality gives data scientists more options and can influence the model being developed. This is one of the reasons it is so helpful to have a data scientist included in the effort to help bring in these perspectives to the process. If these different columns weren't available, model-specific ABTs would need to do heavier transformations to create them. This is a place where upfront investment pays off and expedites model development and delivery.

Further, it is essential to take the time to create a strong data dictionary to accompany your Common ABT. The Common ABT will likely contain hundreds if not thousands of columns, and even those data engineers that help implement it can quickly lose track of what column definitions, column data sources, etc. In your data model, include a business definition as well as a data source (table and column), so that the consumers have clarity on where the data originated. If you are building your Common ABT off your Enterprise Data Warehouse or off a specific data mart, you must notate which table and column in the EDW or data mart it originated. The more information you include, such as transformations, key calculations, etc. are heavily valuable for your team during production support as well as the consumers and analytics project teams you hope to enable.

ON-GOING SUPPORT AND ENABLEMENT

When you are embarking on a Common ABT journey, it's important to recognize and plan for the on-going support and enablement cycles that are required. This is probably the easiest to overlook and the most influential factor for your long-term success.

A Common ABT is rarely a, 'build it and walk away,' approach. Below are some consideration points for the on-going support and management of the Common ABT:

- Analytics project teams need support and cycles from data SMEs. They must understand the information represented and how it can aid their analytics models and typically are not data SMEs themselves. Depending on the size, scale, and number of data subjects represented, it may likely require cycles from the business SME's or data SMEs engaged during the Common ABT development and design to provide this support to the analytics project team. Having data SMEs available to support analytics project teams is essential to ensure that analytics project teams are using the data correctly for their analysis purposes. Analytics groups that do not get the needed data SME support end up either making false assumptions about the data or stop using the Common ABT because they don't understand the data and how to use it properly.
- As you build the Common ABT, it is important to establish the necessary data security and data access processes. Consider storing the Common ABT in a database that allows you to create views on top of it. Views provide flexibility to make only the data subjects that are required for the specific analytics project available. Especially with new legislation around the General Data Protection Regulation (GDPR), you must ensure only the data needed for an analytics effort is made available. Views are a great way to help manage that security and access without a lot of overhead.
- It is recommended that you create a feedback loop for scenarios where analytics project teams begin to require data outside the scope of the currently available Common ABTs. This provides you with insight on how you can continue to enhance the Common ABT so that it stays relevant and valuable for current analytics efforts.
- As your company adds more data sources that may be relevant to this Common ABT, you can proactively add these sources to the Common ABT as a mechanism to provide on-going value that lasts.

CONCLUSION

Analytics efforts can often be time consuming with the largest allocations being devoted to data preparation tasks and the building of an Analytics Base Table (ABT). When each analytics effort is forced to build an ABT from scratch, there is limited tangible value gained from analytics efforts of the past. By using the Common ABT technique, you can create an analytics-centric foundation that centralizes and addresses many of the arduous data preparation tasks that consume so many cycles during the analytics project lifecycle. Further, these techniques create the opportunity to further enable analytics empowerment,

governance, and management that is based on a single source of truth. Although the Common ABT does not address every situation and every opportunity, it can provide tangible value in your time to delivery and ability to support multiple analytics efforts in a centralized and successful way.

CONTACT INFORMATION

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

Jennifer L. Nenadic
Director, Data & Analytics Center of Excellence, SAS®
Jennifer.Nenadic@sas.com
http://www.linkedin.com/in/jennifer-nenadic

BASIC INSTRUCTIONS

WRITING GUIDELINES

Trademarks and product names

To find correct SAS product names (including use of trademark symbols), if you are a SAS employee, see the. Otherwise, see SAS Trademarks.

- Use superscripted trademark symbols in the first use in title, first use in abstract, and in graphics, charts, figures, and slides.
- Do not abbreviate product names. For example, you cannot use "EM" for SAS® Enterprise Miner™. After having introduced a SAS product name, you can occasionally omit "SAS" for certain products, provided that your editor agrees. For example, after you have introduced SAS® Simulation Studio, you can occasionally use "Simulation Studio."

Writing style

- Use active voice. (Use passive voice only if the recipient of the action needs to be emphasized.) For example:

  The product creates reports. (active)
  Reports are created by the product. (passive)

- Use second person and present tense as much as possible. For example:

  You get accurate results from this product. (second person, present tense)
  The user will get accurate results from this product. (future tense)

- Run spellcheck, and fix errors in grammar and punctuation.

Citing references

All published work that is cited in your paper must be listed in the REFERENCES section.

If you include text or visuals that were written or developed by someone other than yourself, you must use the following guidelines to cite the sources:

- If you use material that is copyrighted, you must mention that you have permission from the copyright holder or the publisher, who might also require you to include a copyright notice. For example: "Reprinted with permission of SAS Institute Inc. from *SAS® Risk Dimensions®: Examples and Exercises*. Copyright 2004. SAS Institute Inc."
- If you use information from a previously printed source from which you haven't requested copyright permission, you must cite the source in parentheses after the paraphrased text. For example: "The minimum variance defines the distance between cluster (Ward 1984, p. 23)

1.