Paper SAS3142-2019

Data Preparation in the Analytical Life Cycle

Ivor G. Moan, SAS Institute Inc.

ABSTRACT

The term data preparation summarizes all tasks that are done when collecting, processing, and cleansing data for use in analytics and business intelligence. These tasks include accessing, loading, structuring, purging, unifying (joining), adjusting data types, verifying that only valid values are present, and checking for duplicates and uniform data (for example, two birthdates for one person). Data preparation can be costly and complex because of increasing data quantity and the number of data sources. Data preparation is a new paradigm that is now shaping the market. Previously, data management concentrated on designing and running extract, transfer, load (ETL) and data quality processes in order to feed analytic processes. This situation was all very well when data volumes were smaller and the velocity of new data was slower. We're now seeing a trend toward more dynamic data management, with data preparation playing the role of self-service data management. Traditional data management processes can produce data up to a point, but dynamic fine-tuning and last-minute work is being done in a self-service way, using data preparation tools. It's more and more important to shape the data and get it right for analytics. Data preparation has become important because many more companies are data-driven. Businesses make decisions based on data, and it is enormously important that you can access data quickly and prepare it for analysis and business intelligence.

INTRODUCTION

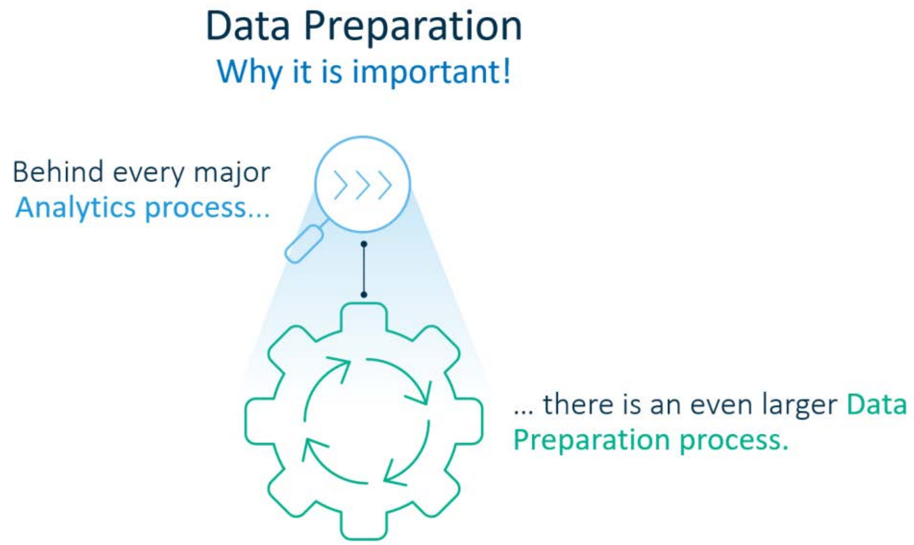

Data Preparation
Why it is important!
Behind every major Analytics process...
... there is an even larger Data Preparation process.

Figure 1. What Is Data Preparation and Why Does It Matter?

In this session, we discuss what Data Preparation (Figure 1) is and how it fits-in with the Analytical Life Cycle. We'll see that many more people are involved in Data Preparation – driven by the need for self-service Data Management and by the demands of getting data into shape for analytical processes. In many cases Data Preparation can be seen as the last mile in the journey of the data. Up until this point data will have traveled from its sources and will have been integrated, collated and cleansed. But analysts often see data differently than the Data Engineers. They see data that's missing which are necessary for particular analytical models. They see opportunities to derive data to more effectively use analytical models. And they see data that, while of a high quality, can often be improved through last minute data quality activities.

What to do? Should the analysis of the data be postponed while data processes are improved, even if much of these analysis' are experimental in nature rendering the data process updates immediately obsolete! Or should the Data Scientist grab a hold of the data that's available and render it in the way that they see fit to speed the Analytical Life Cycle that's driving businesses today?

UNDERSTANDING DATA PREPARATION IN THE ANALYTICAL LIFE CYCLE

There are two main phases in the analytical life cycle (Figure 2): discovery and deployment. First, we ask a question. The discovery process is driven by asking business questions and defining what the business needs to know. The business question must then be translated into a representation of the problem, which can be solved using predictive analytics.

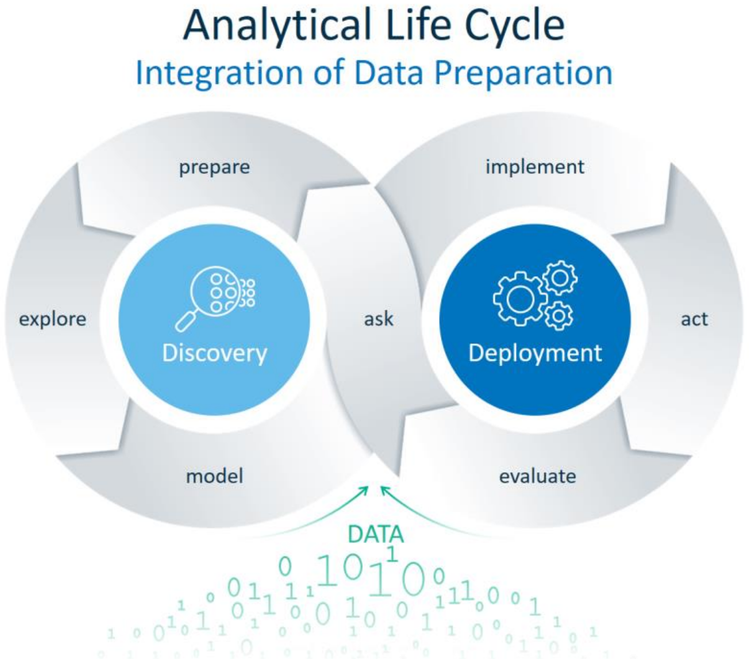


Figure 2. Integration of Data Preparation in the Analytical Life Cycle

Predictive analytics requires suitable data that is prepared appropriately. Technologies like Hadoop, and faster, cheaper computers have made it possible to store and use more data, and more types of data, than ever before. However, this has only amplified the need to join data in different formats from different sources and to transform raw data so that it can be used as input for predictive modeling. The data preparation stage has become even more challenging because of new data types from connected devices (for example, machine sensor data and web logs from online interactions). Many organizations still report that they spend an inordinate amount of time, sometimes up to 80 percent, dealing with data preparation tasks.

DATA PREPARATION IS AN ONGOING PROCESS

Exploring data involves using interactive, self-service visualization tools. The tools need to serve a wide range of users, from the business analyst with no statistical knowledge, to the analytically savvy data scientist. The tools must enable these users to search for relationships, trends, and patterns to gain deeper understanding of the data. Therefore, the data exploration step refines the business question and the approach formed in the initial "ask" phase of the project. The data exploration step also develops and tests ideas about how to address the business problem. However, it might be necessary to add, delete, or combine variables to create more focused models, which involves more data preparation.

In the modeling stage, analytical and machine-learning modeling algorithms are used to determine relationships in the data and to answer the business question. Analytical tools search for a combination of data and modeling techniques that reliably predict a desired outcome. There is no single algorithm that always performs best. The "best" algorithm for solving the business problem depends on the data. Experimentation is key to finding the most reliable answer, and automated model building can help minimize the time to results and boost the productivity of analytical teams. Data preparation continues to be important in the modeling stage as more data can be developed and added.

Once you have built your models, you then need to implement or deploy them. But even then, data preparation does not stop. A model is only as good as the data that it uses, so models (and data) must be kept up-to-date as much as possible. Data preparation and management is a necessary and continuous step in all stages of the analytics life cycle.

CUSTOMER DEMAND

The current situation of data preparation is largely driven by expanding quantities of data and data sources. The advent of big data, along with a wider range of data formats and new data sources (for example, social media and machine sensor data), have meant that it is harder to store and use data. At the same time, organizations recognize that it is becoming more and more essential to use data effectively to support decision making.

Users want more and more data. They want to be able to include their own data and external data in their analyses. Self-service data preparation is popular because it is more flexible, independent, lower-cost, faster, and easier to control (Figure 3). It also creates less work for other departments.

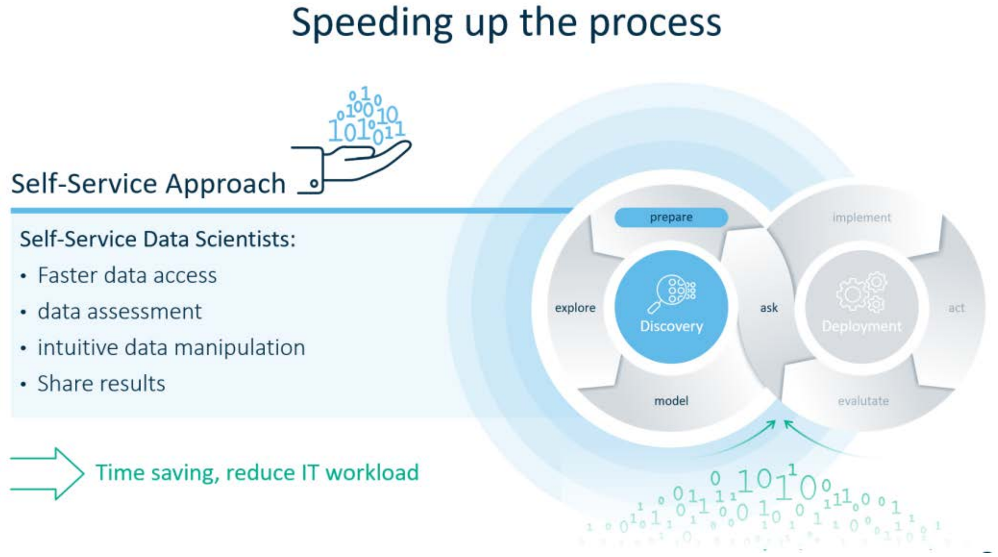
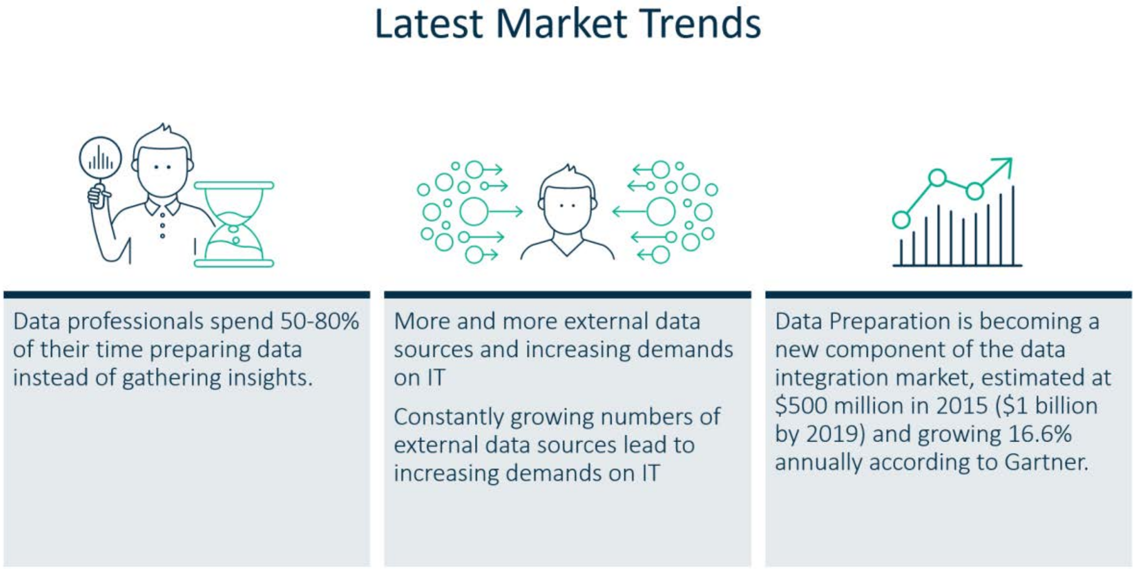


Figure 3. Self-Service Data Preparation Speeds Up the Analytical Life Cycle

Gartner has commented that “The self-service data preparation software market is expected to reach $1 billion by 2019, with a 16.6% annual growth. Adoption is currently 5% of potential target users and is expected to grow to more than 10% by 2020. Vendors must understand the market opportunities when planning their business strategy.” However, this expansion in self-service creates a headache for data scientists. Self-service requires high-quality data preparation, and unfortunately, that is time-consuming and there are few shortcuts.



Latest Market Trends

Data professionals spend 50-80% of their time preparing data instead of gathering insights.

More and more external data sources and increasing demands on IT

Constantly growing numbers of external data sources lead to increasing demands on IT

Data Preparation is becoming a new component of the data integration market, estimated at $500 million in 2015 ($1 billion by 2019) and growing 16.6% annually according to Gartner.

Figure 4. Latest Market Trends

The transition between data preparation and analytics is critical. It needs strong analytics and visualization, but it also needs strong data management so that the data can quickly reveal the required information. Fast markets need agile companies!

A new role has therefore emerged in many companies: the data engineer, who is responsible for data preparation or software engineering. Data engineers do the work before the data is handed over to data scientists for analytical modeling.

THE IMPORTANCE OF DATA QUALITY

The new and emerging data engineer role is a recognition of the fact that data quality is essential. Data management is not just about collecting and formatting data. Data management is also ensuring that its quality is suitable. Data quality is therefore becoming an essential theme in the data preparation world (Figure 5).

SAS has been ahead of the curve in data quality for some time. We recognized quite a long time ago that data preparation is more than just reading data. Data preparation also needs to include questions of data quality. Analytical models need to be fed with high-value data. If the data is not clean and high quality, then the output will be correspondingly bad.

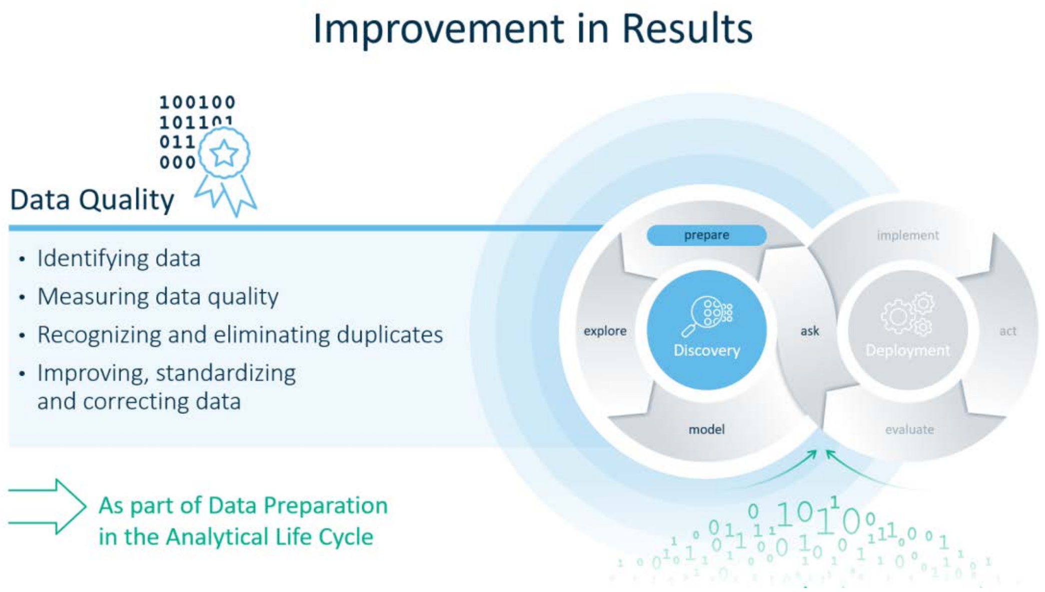


Figure 5. Data Quality Improves Model Results

DATA INSPECTION

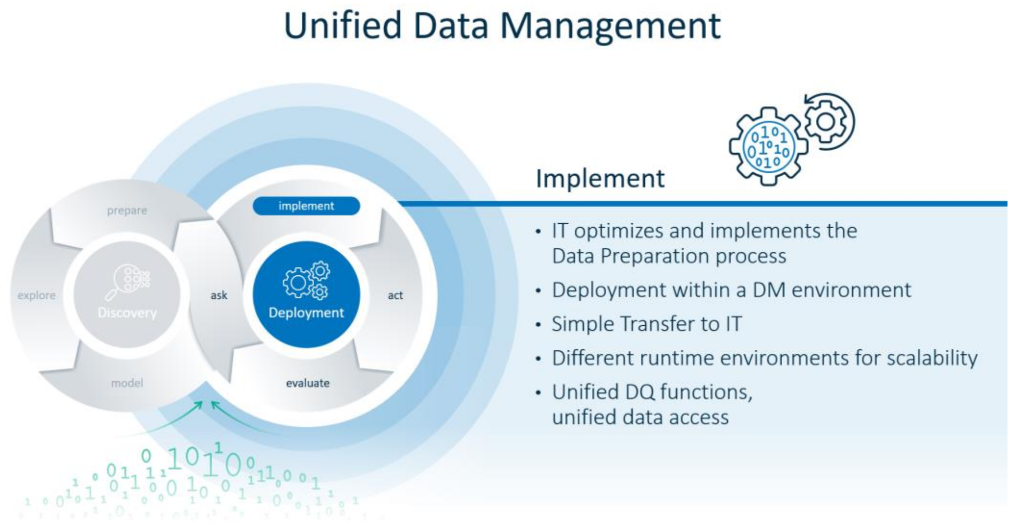
In SAS® Data Preparation, for example, you can see data that has already been imported and is available for use. You can view a sample of the data to get a feel for it, and at first glance everything might look fine. However, you can also look at the data profile in a little more detail. This might show that data has been stored in a variety of ways (for example, some full-form and some with abbreviations). This can create serious problems for analytical models and needs to be resolved before the different sources of data are brought together.

The data now needs to be corrected and standardized. There are several possibilities available for this. For example, for time series analysis, we can filter data and remove any items with missing data. Inconsistencies in how the data is written needs to be corrected and cleansed to remove anomalies and duplications. All this is a key part of data preparation and is becoming both more recognized and essential.

These two areas have very much driven recent developments in the data preparation and data management world, and in the tools that are available. Self-service tools are more ubiquitous, and are coupled with elements that ensure data quality, giving the best of all worlds.

DATA MUST BE GOVERNED – AND THAT COVERS DATA PREPARATION

Data preparation might be a relatively new area for many businesses, particularly as a separate entity. However, data preparation processes and practices must still comply with the organization’s data governance processes and rules. This is also the case for data integration and data management solutions. All data-related processes must fit within the organization’s overall data governance solution (Figure 6).



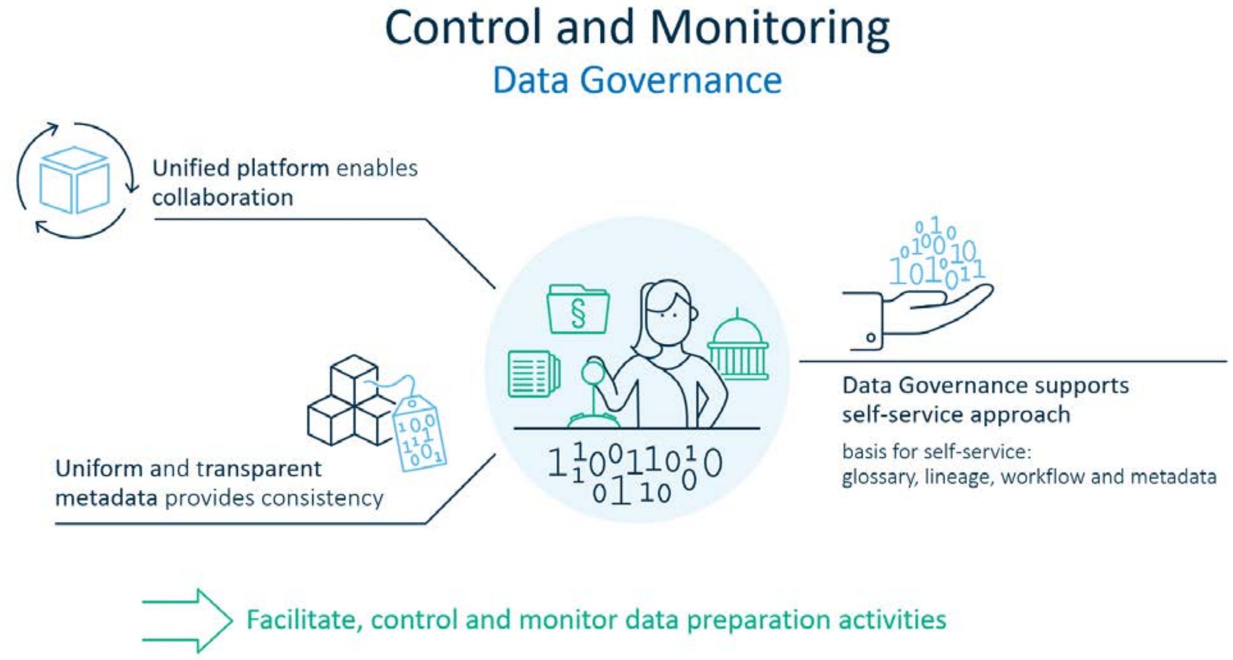
Implement

- IT optimizes and implements the Data Preparation process
- Deployment within a DM environment
- Simple Transfer to IT
- Different runtime environments for scalability
- Unified DQ functions, unified data access

Figure 6. Unified Data Management

Why does this matter? First, a big part of analytics is that different user groups work together across the analytics life cycle, including IT, data scientists, and business analysts. They need to be operating with the same data and using the same principles, or the outcome of the analytical modeling is likely to be at best ambiguous and at worst downright wrong.

This collaboration must follow and be driven by data governance principles (Figure 7). In other words, data governance is a key part of the process and should be used to enable better cooperation and joint working. It certainly should not be seen as a hindrance to be overcome or circumvented by any means possible.



Unified platform enables **collaboration**

Uniform and **transparent metadata** provides consistency

Data Governance supports self-service approach

basis for self-service:
glossary, lineage, workflow and metadata

Facilitate, control and monitor data preparation activities

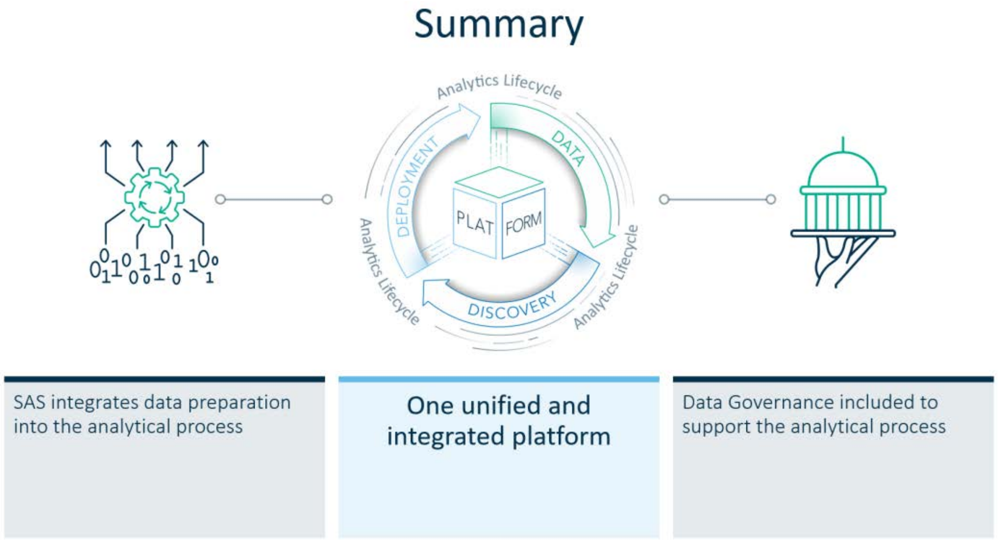
Figure 7. The Role of Data Governance in the Analytical Life Cycle

Governance can facilitate glossaries and deliver more transparency. In practice, this means that nontechnical business users who do not work with the data every day can still be selfsufficient and get all the information they need by serving themselves without worrying about data quality. The organization can also be confident that users are all drawing on high-quality data, and that the data is being used appropriately and in line with legal or ethical requirements.

Figure 8. Bring the Appropriate Tools for Data Preparation

SELF-SERVICE AND DATA PREPARATION

Modern data preparation tools must work closely with data governance functions to accelerate self-service. Self-service analytics can only function alongside self-service data preparation. It is unfortunate but true that business users given access to self-service analytics, but without good quality data, will simply pull in the data they need, from whatever source they can – and then assume that the outcome will still be good. The analytics life cycle will only really work when we have self-service with good quality everywhere.



SAS integrates data preparation into the analytical process

One unified and integrated platform

Data Governance included to support the analytical process

Figure 9. Summary of the Analytical Life Cycle

CONCLUSION

There are two key messages about data preparation in the analytical life cycle.

Perhaps the most important thing is to understand that the analytical life cycle is an integrated process. There are various user groups that are active in this process, and various tools that operate in different phases of the life cycle. Harmonious collaboration and ease of transition from one phase to another is very important.

I think an integrated analytics platform, covering both analytics and data preparation – and here I mean processes that ensure data quality, data integration, and data governance – facilitates the entire analytical life cycle. This is a very important point. Customers who want to accelerate the analytics process are particularly well-served with an integrated platform.

The second important point is the central role of data governance. In my experience, governance is a vital support for self-service in the analytical life cycle. It is very important that users are self-reliant and can obtain the knowledge they need (for example, by using a glossary, or via metadata management) about the data that they need to use and in the right context. Therefore, governance is an essential part of the analytical life cycle.

CONTACT INFORMATION

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

Ivor G. Moan
SAS Institute Inc.
In der Neckarhelle 162
69118 Heidelberg, Germany
Email: Ivor.Moan@sas.com
Web: www.sas.com