Thirsting for Insight? Quench It With 5
Data Management for Analytics Best Practices.

§sas

THE POWER TO KNOW®

Contents

Data Management: Why It's So Essential

We all know that good, clean water is core to life. Without it, we can only survive for around three days. So what happens if your water source is polluted? Well, unless you filter the water sufficiently, there will definitely be some negative consequences. To get better results, you could enrich the water with fluoride, filter out the arsenic, and deliver it at the right pressure and temperature.

In a similar way, "unpolluted" data is core to a successful business – particularly one that relies on analytics to survive. But preparing data for analytics brings with it different requirements than storing data in a warehouse. How difficult is it to manage unfiltered data and get it ready for analytics?

Most data scientists spend 50 to 80 percent of their model development time on data preparation. That cuts sharply into the time they could have spent generating insights.

For data scientists and business analysts who prepare data for analytics, data management technology from SAS acts like a data filter – providing a single platform that lets them access, cleanse, transform and structure data for any analytic purpose. As it removes the drudgery of routine data preparation, it reveals sparkling clean data and adds value along the way. And that can lead to higher productivity, better decisions and greater agility.

SAS adheres to five data management best practices that support advanced analytics and deeper insights:

- Simplify access to traditional and emerging data.
- Strengthen the data scientist's arsenal with advanced analytic techniques.
- Scrub data to build quality into existing processes.
- Shape data using flexible manipulation techniques.
- Share metadata across data management and analytics domains.

The Basics of Data Preparation

Data scientists and business analysts often know ahead of time what data they want to profile or visualize prior to preparing and modeling it. But what they don't know is which variables are best suited – with the highest predictive value – for the type of model being implemented and the variable being modeled. Identifying and accessing the right data are crucial first steps. Because the intensely iterative data preparation process is what separates a good model from an exceptional one.

The process starts when a business problem is identified. Let's take customer churn as an example. The first step is to gather data that represents what you want to model (the churned customers) along with all other data related to those churned customers. The goal is to increase the explanatory value of the churn model so you can predict which customers might churn in the future. You can do this by analyzing the predictive value of individual data elements and/or combinations of data elements.

Once you know that a data element does not have any effect on the behavior, you drop it – then you start to analyze different data elements or combinations of data elements. The process repeats until you reach a certain confidence level in terms of how well the model explains the behavior. Only then can the model accurately predict who will churn next, and when.

But before you can build an effective model, you'll need consistent, reliable data that's ready for analytics. That's where our five data management for analytics best practices can help.
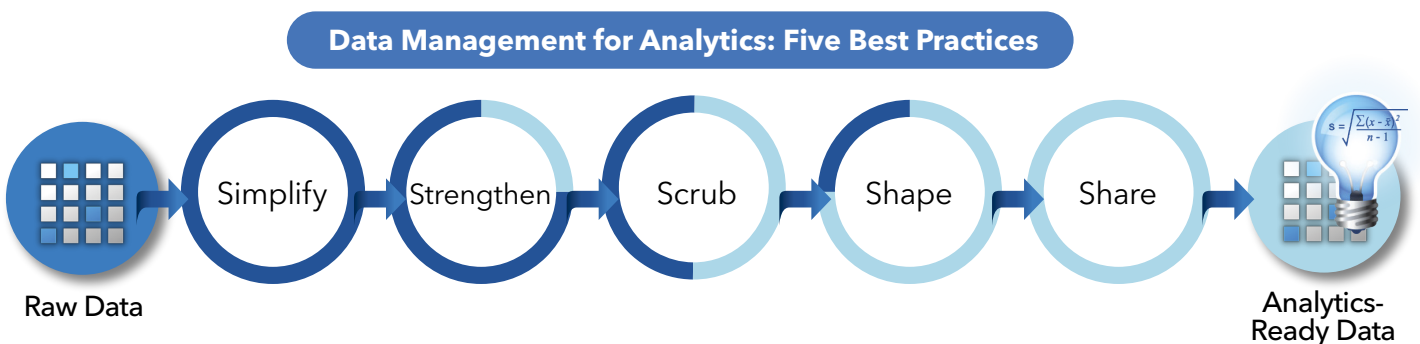

Figure 1: In preparing data for analytics, SAS recommends five data management best practices that transform raw data into analytics-ready data.

1: Simplify Access to Traditional and Emerging Data

Business analysts and data scientists want as much data as they can get their hands on, whether it's from emerging big data sources like Hadoop, SAS® data sets or their own Oracle or Teradata warehouses. The data preparation process involves determining what data can best predict an outcome. And because more data generally means better predictors, bigger really is better in this case.

But accessing all that data is challenging. One reason is that different data sources, formats and structures make it hard to bring the data together. And statistical analysis essentially only cares about two data types – character and numeric. Yet some data sources, like relational databases, have 10 to 20 different numeric data types.

SAS has a plethora of native data access capabilities that make working with a wide variety of data sources easy. Here's why:

- **SAS simplifies access to multiple data sources.** From an Excel spreadsheet to a relational database table to Hadoop, SAS removes the complexity of reconciling data types by automatically converting them.

- **SAS minimizes data movement and improves governance** – and drastically boosts performance – by pushing data processing down to the data source via SQL pass-through and the SAS Embedded Process – a portable, lightweight

> Data management for analytics is not the same thing as data management for an enterprise data warehouse. Analytical data management adds value along the way by completing summarizations and adding metadata to variables before putting them into memory.[1]

[1] blogs.sas.com/content/corneroffice/2015/02/18/5-questions-to-prepare-you-for-the-internet-of-things/

SAS execution engine that runs anywhere: in-memory, on Hadoop, on Oracle or with other database technologies.

- **SAS provides self-service data preparation capabilities** with intuitive user interfaces that make data accessible to more users, with less training. This, in turn, frees IT personnel from iterative data provisioning tasks so they can be more productive.

- **SAS enables agile, secure techniques for managing data.** For example, SAS uses data virtualization to quickly create virtual business views of data without moving it. And it can enforce security policies through dynamic data masking so sensitive data is not compromised.

2: Strengthen the Data Scientist's Arsenal With Advanced Analytic Techniques

With advanced analytic techniques from SAS, you can make better decisions from your data. That's because traditional ETL and data integration technologies do not have the more sophisticated statistical analysis capabilities that SAS provides inside of the ETL flow. Here are some examples:

- **Frequency analysis** goes beyond simple counts to help identify outliers and missing values that can skew other measures like mean, average and median (measure of central tendency), as well as effects analyses like forecasting. Frequency analysis is also used to determine percentiles and binning limits for creating data segments that require different treatment during the modeling and forecasting phases.

- **Summary statistics** describes the data by providing several measures, including central tendency, variability, percentiles and cardinality. Cardinality shows how many unique values exist for a given variable. Business analysts use this technique to understand the distribution and variance of the data – because data may not be normally distributed, as many statistical methods assume.

- **Correlation** is used during the analytic model building process, when business analysts try to understand the data to determine which variables or combination of variables will be most useful based on predictive capability strength. Business analysts have to understand which variables potentially influence one another, and to what degree (as in Figure 1). They need to understand all of this for standalone variables, as well as for variables in combination with others.
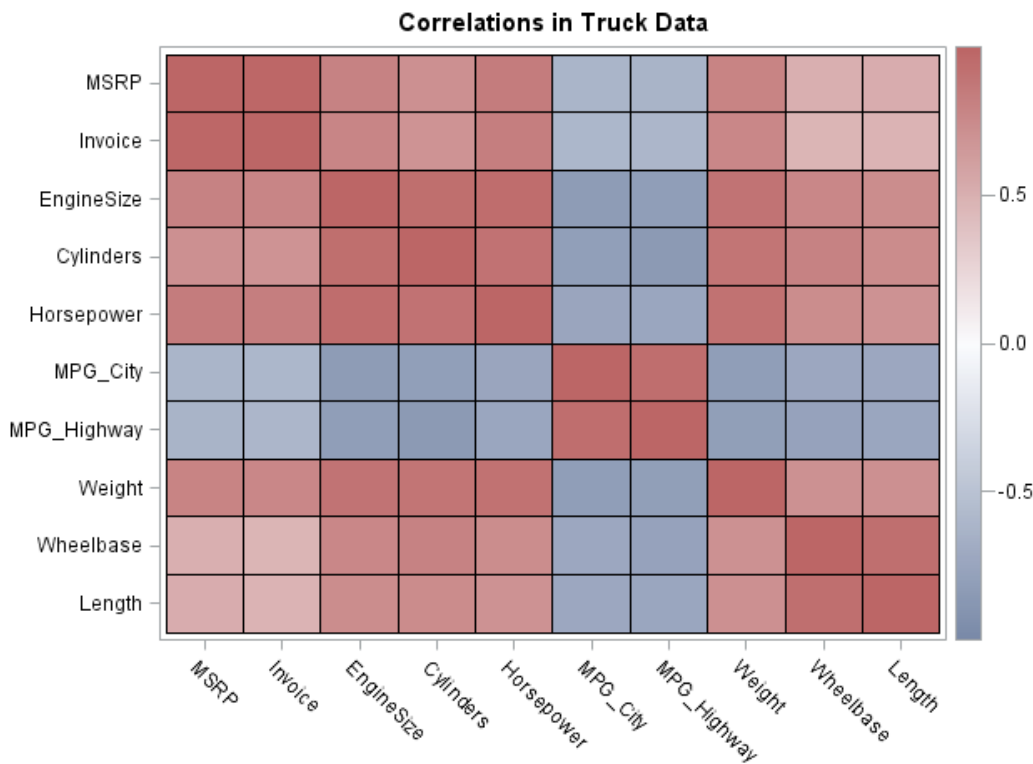
Correlations in Truck Data

Figure 2: A correlation heat map shows measures of association – that is, the degree of the relationship – between variables.

3: Scrub Data to Build Quality Into Existing Processes

Did you know that up to 40 percent of all strategic processes fail because of poor data? An analyst-recognized leader in data quality, SAS recognizes the importance of having data you can trust.

Data cleansing begins with understanding the data through profiling, correcting data values (like typos and misspellings), adding missing data values (like ZIP code), finding and dealing with duplicate data or customer records, and standardizing data formats (dates, monetary values, units of measure). Cleaning data can also include automated selection of best records and cleaning data in multiple languages.

SAS has an industry-leading data quality platform that:

- Incorporates the cleansing capability into your data integration flow to make IT resources more productive.

- Puts data quality in database – that is, pushes this processing down to the database to improve performance.

- Removes invalid data from the data set based on the analytic method you're using – such as outliers, missing data, redundant data or irrelevant data.

- Enriches data via a process called binning – which simply means grouping together data that was originally in smaller intervals. For example, the individual value of age alone may not have much relevance, but age groups could, such as "between 35 and 45." Or "high cholesterol" could be used to group (bin) cholesterol values that are greater than 190. Binning is also a way to reduce cardinality, because it creates relevant data where there was none before.

4: Shape Data Using Flexible Manipulation Techniques

Without flexible methods of manipulating data, it can be difficult to structure the final data set. Here's why: Typical analytic methods expect a "flattened" data set, often called "one-row-per-subject." A flattened data set is a single table that contains all the data from all the sources, organized so that each record holds all the relevant data for the subject being analyzed.
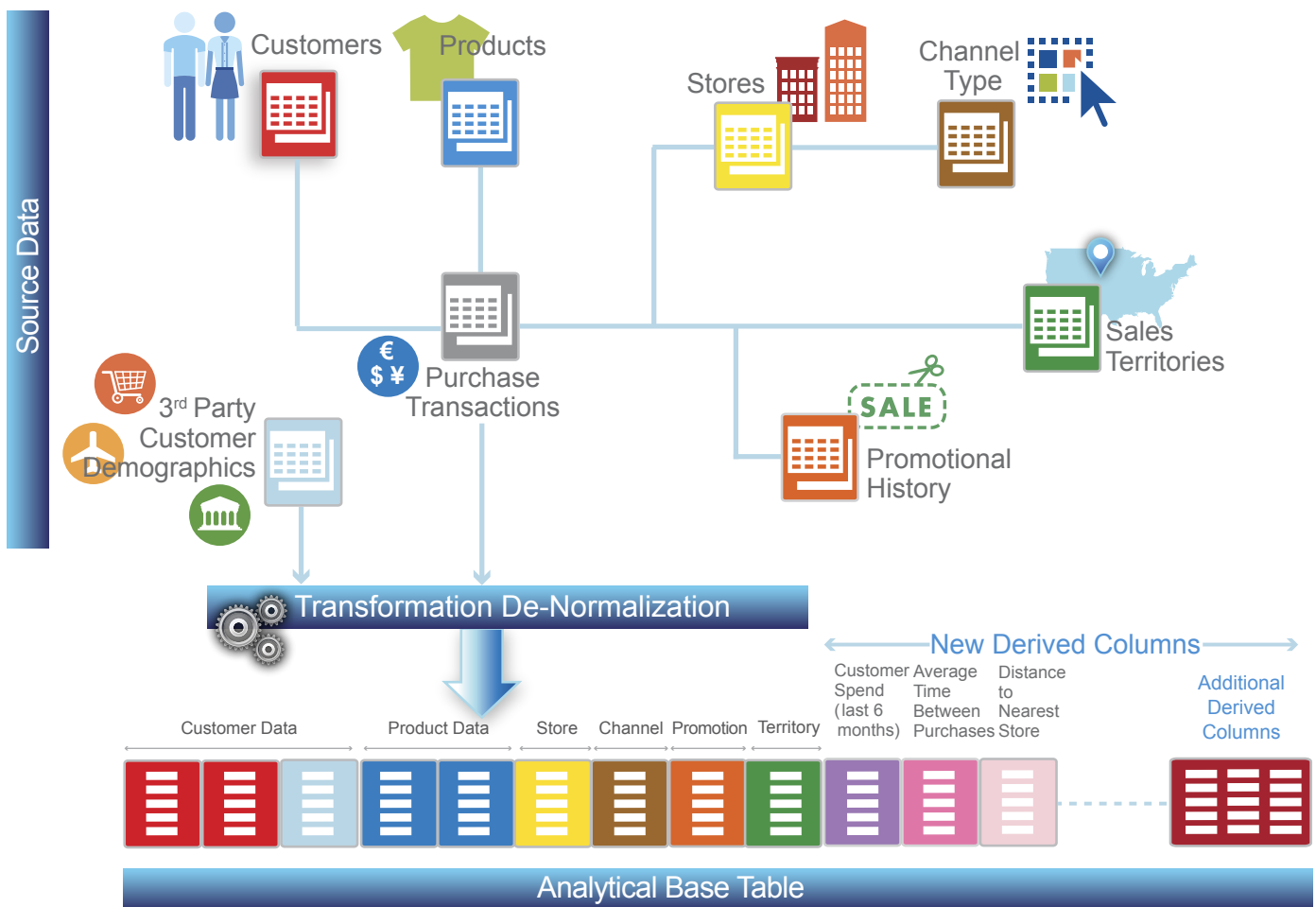
Figure 3: Preparing data for analytics requires merging, transforming, de-normalizing and sometimes aggregating your source data from multiple tables into one very wide table, often called an analytical base table (ABT), with potentially thousands of columns.

Getting data into a flattened data format can be problematic. If you're trying to predict when your customers may churn, for example, the customer data you use has to be collected from multiple sources (data warehouse, transactional and third-party data sources). Then it has to be organized into a single row for each customer. The resulting data set can become very wide, with potentially thousands of columns.

Database systems are not designed with a single-row-per-customer data structure in mind. As a result, many database systems limit the number of columns a single table can have. Transaction systems record every transaction as it happens, resulting in a high volume of records for each customer. These transaction records need to be consolidated and transposed to be joined with the customer records pulled from the data warehouse.

Such data transposition is an example of what can be an arduous data manipulation task. If done programmatically, it can involve hundreds of lines of code. That's because transaction data needs to be sorted, groups of transactions for a particular customer need to be identified, and transactions need to be summarized, perhaps per product. The resulting "new" data structure needs to be part of the one-row-per-subject table that will be used for analysis.

SAS simplifies data transposition with intuitive, graphical interfaces for transformations. Plus, you can use other reshaping transformations. Those include frequency analysis to reduce the number of categories of variables; appending data; partitioning and combining data; and a variety of summarization techniques.

Data preparation is the key to a successful analytic model. One company SAS worked with did not reuse metadata or data transformation logic from its analytical model development area when it deployed the model into operational systems. Because of this, the company spent $1 million and engaged numerous IT personnel – for six months – just to recode, retest and deploy the model.

5: Share Metadata Across Data Management and Analytics Domains

SAS has a common metadata layer that allows data preparation processes to be consistently repeated. This promotes more efficient collaboration between those who initially prepare data and the business analysts and data scientists who ultimately complete the data preparation process and analytic model development.

Common metadata also provides lineage information on the data preparation process, so you can answer questions like: Where did the data come from? What was its quality? What data was used, and where else has it been used? How was the data transformed? What additional reports or information products are developed using this data?

Because of a common metadata layer between the data preparation and model development environments, SAS makes it easier to deploy models. As each model is registered in metadata and made available along with its data requirements, it becomes less of a challenge to adopt.

Applying metadata across the analytics life cycle delivers savings on multiple levels. When a common metadata layer serves as the foundation for the model development process, it eases the intensely iterative nature of data preparation, the burden of the model creation process and the challenge of deployment. Advantages include:

- **Faster testing and increased productivity** due to automated model development and scoring.
- **Creation of more models with greater accuracy** because of automated model management.
- **Faster cycle times** that increase profitability and result in more relevant and timely models.
- **Less time spent on mundane data work**, and more focus on model development and evaluation.
- **Knowledge that can be reused** across the enterprise after it's obtained during the data preparation process.
- **Increased flexibility** to accommodate changes because of better manageability and governance over the analytics life cycle.
- **Auditable, transparent data** that meets regulatory requirements – you can trace data from start to finish.

Trusted Data, Proven Analytics

Following these data management best practices from SAS can position your organization to get the full value from advanced analytics. Our industry-leading approach gives you access to all types of raw data and lets you cleanse, transform and shape it for any analytic purpose. As you glean continually deeper insights from your data, you can embed that knowledge into your analytical models, share your new discoveries and automate decision-making processes across the business.

Find out more about how SAS, an analyst-recognized leader in data quality, data integration and advanced analytics, supports data management best practices and helps your business build competitive advantage: sas.com/data

sas

THE POWER TO KNOW®