› Conclusions Paper

Big Data, Analytics and Hadoop

How the marriage of SAS® and Hadoop delivers better answers to business questions – faster

Featuring:

Georgia Mariani, Principal Product Marketing Manager for Statistics, SAS

Wayne Thompson, Manager of Data Science Technologies, SAS

Contents

It's the perfect arranged marriage: a low-cost, distributed data storage and processing platform, coupled with strong analytics to make sense of it all.

- **Hadoop** is an open-source software framework for storing and processing huge data sets on a large cluster of commodity hardware. Hadoop delivers distributed processing power at a remarkably low cost, making it an effective complement to a traditional enterprise data infrastructure.
- **SAS** brings data discovery and advanced analytics to the relationship. Both are faster with in-memory processing and more accessible with either an interactive programming or graphical user interface – your choice, depending on your role, skills and preference.

Many types of applications in all vertical markets can benefit from this close relationship. According to TDWI research, organizations are using Hadoop to better understand website behavior via clickstreams (23 percent), sentiment analysis and trending (22 percent), sales and marketing opportunities (17 percent), fraud detection (17 percent), churn and other customer behaviors (12 percent), and customer base segmentation (11 percent).[1]

Hadoop does not replace enterprise data warehouses, data marts and other conventional data stores. It supplements those enterprise data architectures by providing an efficient and cost-effective way to store, process and analyze the daily flood of structured and unstructured data.

Hadoop Made Simpler and More Powerful

"Many organizations have been like the proverbial deer in the headlights, frozen by the newness and enormity of big data," said Philip Russom in a TDWI Best Practices Report on Hadoop. [2] "The right combination of Hadoop products can thaw 'analysis paralysis' by enabling the management and processing of big data, for which traditional data warehouses and business intelligence tools were not designed."

In a TDWI survey of 263 respondents, the vast majority (88 percent) said they consider Hadoop an opportunity because it enables new ways to extract value from big data. However, 12 percent see it as a problem, largely because of a shortage of Hadoop expertise. "The challenge with HDFS [Hadoop Distributed File System] and Hadoop tools is that, in their current state, they demand a fair amount of hand coding in languages that the average BI professional does not know well, namely Java, R and Hive," said Russom.

SAS both simplifies and augments Hadoop. SAS treats Hadoop as just another persistent data source, and brings the power of SAS In-Memory Analytics and its well-established community to Hadoop implementations.

- SAS enables users to access and manage Hadoop data and processes from within the familiar SAS environment for data exploration and analytics. This is critical, given the skills shortage and the complexity involved with Hadoop.
- SAS augments Hadoop with world-class data management and analytics, which helps ensure that Hadoop will be ready for enterprise expectations.

"Hadoop is very important to our customers," said Wayne Thompson, Manager of Data Science Technologies at SAS. "It is a very efficient way to store data in a very parallel way to manage not just big data but also complex data. The SAS Analytics environment, collocating on the Hadoop cluster, enables you to run very advanced, distributed, statistical and machine learning algorithms."

> SAS and Hadoop are natural complements. SAS treats Hadoop as just another data source and technology that can be brought to bear for appropriate use cases. SAS brings world-class analytics to the merits of Hadoop.

"The analytic computations are done inside of the Hadoop cluster without having to drop intermediate data down to disk," said Thompson. "Hadoop is used to manage the data, to load the data into memory and distribute it across the cluster. But SAS is also collocated – installed in the Hadoop cluster. It

[1] TDWI Best Practices Report, *Integrating Hadoop Into Business Intelligence and Data Warehousing,* Philip Russom, 2Q2013

[2] TDWI Best Practices Report, *Integrating Hadoop Into Business Intelligence and Data Warehousing*, Philip Russom, 2Q2013

doesn't matter if it's Cloudera or HortonWorks, etc. Once the data is lifted into memory, SAS takes over to multitask the calculations – to do the explorations, the predictive modeling and also some machine learning. In this case, we don't use [the native Hadoop computational approach] MapReduce; we use our own threaded kernel instructions inside the database – and manage that across the cluster to get answers back almost instantaneously."

"Since in-memory processing is so fast, the time to process advanced analytics on big data is reduced," wrote Fern Halper, Research Director for Advanced Analytics at TDWI. "This frees up more time to actually think differently, experiment with different approaches, fine-tune your champion model, and eventually increase predictive power. For example, a training set for a predictive model which might have taken hours to run through one iteration now takes minutes utilizing in-memory techniques. This means that more/better models can be built, which helps to derive previously unknown insights from big data.

"Once data is in-memory it can be accessed quickly and interacted with more effectively. For example, if someone builds a model that now is able to run faster, they can share intermediate results with others and interact with the model more quickly. It can be changed on the fly, if needed, as others look at it and make suggestions."[3]

Here's the kicker: It's point-and-click, drag-and-drop easy, if you want it to be. Or you can have programming-based flexibility, if that's your preference.

In a SAS-hosted webinar held before the Strata 2014 big data conference in Santa Clara, CA, Thompson demonstrated both approaches: using SAS Visual Statistics and SAS In-Memory Statistics for Hadoop.

A Graphical User Interface for Exploring Big Data in Hadoop

Thompson demonstrated how easy it is to develop models – in this case, to better understand the contributors to a charitable cause – so as to understand how to maximize donations. The interface is intuitive – and fast. Drag and drop a variable into the desktop and see what effect it has. Grab other variables to see how they might be correlated with donation amount. Drag and drop to do autocharting. Zoom to see details in a pop-up window.

"This is very fast and furious," said Thompson. "Working on the fly, we can drag and drop data onto the desktop, perhaps first a histogram, then a correlation matrix to identify strongly or weakly correlated variables. We can do lots more exploratory analysis, very quickly."

Thompson goes on to express the correlation matrix with the click of the mouse as a multiple linear regression, showing donation amount as a function of the other selected variables. He highlights the row and selects a predictive model from a pop-up menu. The system automatically develops a regression model. He then grabs from the left pane a few more variables that might be of interest, drops them onto the desktop, and once again a regression model is automatically set up and developed.

"It's very easy, very interactive, and just about anybody could do it," said Thompson. "Very quickly the data is loaded into memory, it is only read from disk one time, then the computations are done across the grid, and I can work very interactively in an exploratory manner."

> "I'm developing the model interactively. The data is only loaded once into memory, and then we repetitively analyze the data in memory, without reading back to disk – using SAS, not MapReduce, running inside Hadoop, collocated with the data."
>
> Wayne Thompson, Manager of Data Science Technologies, SAS

Since the sample regression model showed most of the variables to be influential on the target variable (donation amount), Thompson displays the statistics detail and modifies the thresholds. An autogenerated line chart compares predicted to actual donations across various bins by decile – then investigates facets of each decile. Looking at residual diagnostics, we can see how well the model is performing for each decile. Click, select and voila: A heat map shows where the model needs tuning. Interactively add model effects, or exclude model effects to refit the model. And so on.

[3] TDWI Checklist Report, *Eight Considerations for Utilizing Big Data Analytics with Hadoop*, Fern Halper, March 2014

And that's just for one algorithm. There are logistic regressions, generalized linear models, decision trees, random forests, integrated model comparisons and clustering, to name a few. "Working interactively, we have the ability to slice and dice the data in so many different ways without ever dropping the data back down to disk," Thompson said. "And without having to learn to program. It's very easy to develop these models." Warning: It can also be addictive.

> "Data visualization tools are becoming de rigueur with Hadoop and big data in general. Over one-third of respondents report using data visualization tools with Hadoop today (38 percent), and another 42 percent anticipate doing so within three years."
>
> Philip Russom, TDWI
> *Integrating Hadoop Into Business Intelligence and Data Warehousing*

Where does SAS Visual Statistics fit in with SAS Visual Analytics and SAS® Enterprise Miner? "SAS Visual Statistics is a new product for advanced analytics that seamlessly uses SAS Visual Analytics for preliminary data exploration and evaluating ad hoc models," said Thompson. "SAS Visual Statistics adds new capabilities, such as more tuning parameters for model development and additional methods such as generalized linear models, interactive decision trees and clustering. More important is that these products are very tightly coupled, both from a licensing and functional perspective."

As far as SAS Enterprise Miner, the products are optimized for different purposes, Thompson explained. "SAS Visual Statistics is a drag-and-drop, turn-on-a-dime, smoke-the-tires kind of exploratory tool, particularly for leveraging big data. SAS Enterprise Miner is more of a process flow, drag-and-drop, batch-driven application. There are some overlaps in the algorithms. SAS Visual Statistics does generate SAS code so you can do integrated model comparisons in SAS Enterprise Miner. So they work in tandem."

> "Visualization is a great way to explore data and discover unknown facts, which is why it's a great fit for the discovery analytics typically done with big data. In addition, leading data visualization tools work directly with Hadoop data, so that large volumes of big data need not be processed and transferred to another platform."
>
> Philip Russom, TDWI
> *Integrating Hadoop Into Business Intelligence and Data Warehousing*

For Data Scientists Who Prefer a Programming Environment

At this point in the demo, data scientists might protest. A visual environment is quick and easy, but what if you like writing code? What if you need levels of control and flexibility that only custom programming can provide? For those types, there is SAS In-Memory Statistics for Hadoop: a single, interactive programming environment for analytical data preparation, variable transformation, exploratory analysis, statistical modeling and machine-learning techniques, integrated model comparison and scoring – all inside the Hadoop environment.

Interactive programming enables multiple users to quickly analyze data in Hadoop. "As with SAS Visual Statistics, the process is interactive and visual – dragging and dropping terms – but I'm writing code," said Thompson. Thompson demonstrated the product's versatility with a hypothetical business problem: what constitutes a lemon vehicle and how to avoid buying one at an online auto auction.

Thompson's demo database has more than 11 million observations and contains variables such as odometer reading, price, buyer number and whether or not the vehicle was an online purchase – joined with additional car information from a dimensional table. From this data, Thompson builds a supervised classification model. The process starts with exploratory analysis to better understand the data. Data is loaded into memory, explorations are interactive, and responses come back almost instantly. In the demo, the system runs distinct counts and computes centrality measures on 11 million observations in 3.26 seconds.

"As I'm analyzing the data, rather than writing to disk, temp tables are being created, and I can add new columns to these temp tables on the fly," said Thompson. Thompson creates new variables – vehicle age and average odometer reading, both computed from other variables – then targets the exploration to vehicles of a certain age and use pattern. Clicks to run it. In seconds, we see that average odometer reading is higher for "bad" older cars – no surprise there – but it's higher for "good" newer cars. This finding points the way to further investigation.

"It's very easy to work in this environment," said Thompson. "This is the way a lot of data scientists work, but it's very easy to work with the language and get back detailed information. It's very simple to look at, and you get results within seconds. There's no time to look at the log to see if it's running, because about as soon as I submit it, I already have output."

Thompson then gets additional summarizations, creates a few new attributes to add to the model, strips out other variables, joins tables and displays an analysis-ready table. All within minutes. He then runs a multipass algorithm – a logistic regression with backward elimination in this case – and gets results back in eight seconds. Then he computes assessment statistics such as lift, so we can see how well the model is performing for scoring rankings. Next he creates a random forest consisting of 20 decision trees. The algorithm randomly swaps in variables during the construction of the 20 trees while also using boot strap samples. The final model represents an averaging of the trees with out-of-bag samples used to see how well the random forest generalizes.

All in not much more time than it took to read this page.

> An in-memory infrastructure running on top of Hadoop eliminates costly data movement and persists data in-memory for the entire analytic session. This significantly reduces data latency and provides rapid analysis at lightning-fast speeds.

Better Answers in Seconds, Instead of Hours or Days

The demonstrations used data sources with millions of rows. What happens if you have billions of rows, too much data to fit in-memory? No problem, says Thompson. "First, in machine learning and statistics, the number of rows is not as meaningful as the number of columns; the 'width' of the data matters more. At a recent analytics conference, we ran live demos on stage with 70 million observations and got very much the same kind of scalability – almost instantaneous."

For truly heavy-duty processing, there are ways to boost response times, such as adding nodes to the computing cluster or doing some caching back to disk when needed.

> "Big data and analytics go together because analytic methods help user organizations get value from big data (which is otherwise a cost center) in the form of more numerous and accurate business insights."
>
> Philip Russom, TDWI

Closing Thoughts

"Getting relevant information from big data sources such as Hadoop requires a different approach," said Georgia Mariani, Principal Product Marketing Manager for Statistics at SAS. "If you're just looking at reports, doing some data discovery or turning out a couple of analytical models, that's really not going to cut it.

"Getting insights out of Hadoop in a timely manner requires in-memory analytics and an interactive, end-to-end process that addresses analytical data preparation, exploration, modeling and scoring."

SAS marries the power of world-class analytics with Hadoop's ability to perform distributed processing on low-cost commodity hardware. For data exploration and analysis, you have the choice of an intuitive graphical user interface with SAS Visual Statistics or an interactive programming approach with SAS In-Memory Statistics for Hadoop.

Whichever approach is used, SAS and Hadoop integration provides important benefits for extracting the most value from their big data assets:

- **Precision.** Apply the most proven and state-of-the-art analytical algorithms and machine-learning techniques to get the best business results.
- **Scalability.** As data and the number of users grow and problems get more complex, the SAS and Hadoop implementation can scale to match.
- **Speed.** The SAS and Hadoop approach is memory-efficient and data-efficient, so you can rapidly analyze very large and complex data in Hadoop.
- **Interactivity.** A multiuser, interactive analytics environment supports increased productivity.

Big data and analytics go hand in hand. Hadoop and SAS redefine the art of the possible, thanks to a naturally close relationship.

> "Advances such as in-memory analytics and in-database analytics have helped to make analytics computations faster. This has helped organizations more effectively analyze data in order to compete."
>
> Fern Halper, Research Director for Advanced Analytics, TDWI

Learn More

Download the TDWI Best Practices Report, *Integrating Hadoop Into Business Intelligence and Data Warehousing* by Philip Russom, 2Q2013: sas.com/integrate-hadoop

Download the TDWI Checklist Report, *Eight Considerations for Utilizing Big Data Analytics with Hadoop* by Fern Halper, March 2014: sas.com/consider-hadoop

Learn more about SAS and Hadoop: sas.com/hadoop

Learn more about SAS Visual Statistics: sas.com/vis-stat

Learn more about SAS In-Memory Statistics for Hadoop: sas.com/in-mem

Follow us on Twitter: @sasanalytics

Like us on Facebook: SAS Software

About the Presenters

Georgia Mariani
Principal Product Marketing Manager for Statistics, SAS

Over the course of her 16 years at SAS, Georgia Mariani has supported various areas within product marketing, including the education industry and analytical marketing strategy for the public sector business unit. She began her career at SAS as a data mining systems engineer. Mariani received her MS degree in mathematics with a concentration in statistics, and her BS degree in mathematics, both from the University of New Orleans. During her master's program, she was awarded a fellowship with NASA.

Wayne Thompson
Manager of Data Science Technologies, SAS

Over the course of his 20-year tenure at SAS, Wayne Thompson has been credited with bringing to market analytics technologies such as SAS Text Miner, SAS Credit Scoring for Enterprise Miner, SAS Model Manager, SAS Rapid Predictive Modeler, SAS Scoring Accelerator for Teradata and SAS Analytics Accelerator for Teradata. Thompson received his PhD and MS from the University of Tennessee. During his PhD program, he was also a visiting scientist at the Pasteur Institute in Lille, France.

To contact your local SAS office, please visit: sas.com/offices

THE POWER TO KNOW®