From Big Data to Meaningful Information

Insights from a webinar sponsored by KMWorld Magazine and SAS

Featuring:

Presenter:
David Pope, Principal Solutions Architect,
SAS® High-Performance Analytics

Contributor:
Fiona McNeill, Principal Product Marketing Manager,
SAS® Text Analytics
# Table of Contents

**Introduction** .................................................. 1  
The Rising Tide of Big Data .................................. 1  
**What Exactly Is Big Data?** ................................. 3  
Big Data Technologies ........................................... 3  
**What Should You Capture, and What Should You Keep?** ............. 4  
Smart Filters Identify What to Store ................................. 4  
Smart Filters Determine Where to Keep What You Capture .............. 5  
Capture and Correlate Data on the Fly .............................. 5  
**From Hindsight to Insight to Foresight** ......................... 5  
**New Thinking About Data and Model Management** ............... 7  
Evolve from Being Data-Focused to Analytics-Focused ..................... 7  
Consider That Data Preparation Is Different for Analytics ............. 7  
Than for Reporting ............................................. 7  
Manage Models as Critical Information Assets ....................... 7  
Use All the Data, if It Is Relevant ................................ 8  
**How to Get Started with Big Data Analytics** ................. 8  
Determine the Analytical Maturity of the Organization .................. 8  
Get Executive and Management Buy-In ............................ 9  
Consider an Analytics Center of Excellence ......................... 9  
**Closing Thoughts** ............................................ 10  
**About the Presenter** ......................................... 10  
**About KMWorld Magazine** ..................................... 10  
**For More Information** .......................................... 11
Introduction

A retail organization was seeking to increase ROI on marketing campaigns by 15 percent. The company was already using analytics to determine the right offers to make to the right prospects at the right time – but the modeling process took 11 hours to run. Using grid computing, the company reduced that to 10 seconds.

Catalina Marketing found that 10 percent of customers receiving coupons at the grocery checkout redeemed them on a future visit – not bad. But by moving analytics to where the data lived, the company could refresh models much faster. Scoring for millions of customers, which previously took 4.5 hours, is now done in 60 seconds. The result is more tailored coupons based on fresher information – and a 250 percent increase in coupon redemption.

A financial services firm required 167 hours – nearly a week – to assess its risk portfolio across the organization. By adapting infrastructure and processes to support high-performance analytics, the firm reduced that to 84 seconds. The result: better decisions that contributed tens of millions of dollars to the bottom line.

Problems that were difficult or impossible to solve before are now manageable. Organizations can analyze all their data – not just a subset of it – and are empowered to analyze it more extensively, iteratively and frequently. The end result is better business decisions in a fraction of the time.

That’s the promise of big data analytics – advanced analytics applied in a high-performance computing infrastructure to address business questions that are best answered with a vast amount of diverse data sources.

The Rising Tide of Big Data

Organizations are awash in data – gigabytes and terabytes and petabytes of it – churned out daily by operational/transactional systems, imported from purchased databases and propagated through analysis and reporting.

But that’s only the tip of the data iceberg.

By some estimates, this structured (numerical) data represents only about 10 percent of the information in an organization. As much as 90 percent of data is actually unstructured data – freeform text, images, audio and video. This unstructured data comes from websites, correspondence, contact center records, social media, blogs, claims, customer complaints and any number of other sources. It is contained in document repositories, emails, PowerPoint presentations, spreadsheets, PDFs, XML documents, SharePoint sites, website interactions, social media sites and texting channels such as SMS and IM. It is everywhere, and it is growing fast.
What would your organization do if you could harness the insights hidden within that vast sea of words and images? Imagine how much better the business decisions would be if they were based on four, five or even 10 times more data. Picture how you could improve knowledge sharing and decisions if the right unstructured data was easy to find, useable and intelligently embedded into analytical processes.

That was the topic of a webinar hosted by *KMWorld Magazine* and SAS. In the hour-long event, David Pope of SAS explained how organizations can exploit this tidal wave of unstructured data, and how big data technologies redefine what is possible with these huge volumes.

“That the sheer volume of data resources available to us causes a scarcity of human attention,” said Pope. Excessive data dilutes focus. “Once you have transformed all this wealth of data into some consumable form, such as a report, what happens to it? What should you pay attention to? What do you want your customers to pay attention to? The key to deriving insight and not just information from data, regardless of size, really comes down to analytics.”

For unstructured data (particularly unstructured text) this is where text analytics comes in. Text analytics identifies and extracts the relevant information and interprets, mines and structures it to reveal patterns, sentiments and relationships within and amongst documents.

- **Automated content categorization** makes information searches far faster and more effective than manual or retrospective tagging methods.
- **Ontology management** links text repositories together, enforcing data quality with consistent and systematically defined relationships.
- **Sentiment analysis** automatically locates and identifies sentiment expressed in online materials, such as social networking sites, comments and blogs on the Internet, as well as from internal electronic documents.
- **Text mining** provides powerful ways to explore unstructured data collections and discover previously unknown concepts and patterns.

These capabilities have been available for some time and are proving their value. In a 2012 AIM study, *Big Data – Extracting Value from Your Digital Landfills* (Doug Miles, aim.org/research), more than 30 percent of survey respondents said their organizations use analytics to derive insights from document repositories or enterprise content management (ECM) systems. Another 50 percent are planning to do so, or wish they could.

In many organizations the legacy data infrastructure is straining to keep pace even with the existing structured data, never mind the new pressures: escalating data volumes and demands on data, complexity of data usage, a growing user base and faster response time expectations. In other words, organizations are grappling with big data.

“What information consumes is rather obvious. It consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention … the only factor becoming scarce in a world of abundance is human attention.”

Herbert Simon
Economist

“We’ve had a lot of data about our customers and partners for many years. What’s different now is the growth in nontraditional data sources (particularly unstructured data), the degree to which business partners and industry consortiums are willing to share data with others in the industry, and the speed at which we are expected to access and process the data to get business-changing insights.”

A vice president of marketing analytics for a retail finance company
What Exactly Is Big Data?

Big data is defined less by volume – which is a constantly moving target – than by the ever-increasing variety, complexity, velocity and variability of the data. “When you’re talking about unstructured data, the concept of data variety can become more significant than volume,” said Pope. “Organizations must be able to fold unstructured data into quantitative analysis and decision making. Yet text, video and other unstructured media require different architecture and technologies for analysis.

“Legacy data infrastructures are really not designed to effectively handle big data, and that’s why new technologies are coming online to help deal with that. With big data technologies, information users can now examine and analyze more complex problems than ever before. The ability to quickly analyze big data can redefine so many important business functions, such as risk calculation, prize optimization, customer experience and social learning. It’s hard to imagine any forward-looking company that is not considering its big data strategy, regardless of actual data volume.”

Some organizations will have to rethink their data management strategies when they face hundreds of gigabytes of data for the first time; others might be OK until they reach tens or hundreds of terabytes. But whenever an organization reaches the critical mass defined as big data for them, change is inevitable.

Big Data Technologies

Accelerated processing with huge data sets is made possible by four primary technologies:

- **High-performance computing** makes it possible to analyze all available data, for cases where analyzing just a subset or samples would not yield as accurate a result. High-performance computing enables you do things you never thought about before because the data was just way too big.

- **In-database analytics**, an element of high-performance computing, moves relevant data management, analytics and reporting tasks to where the data resides. This approach improves speed, reduces data movement and promotes better data governance.

- **In-memory analytics** can solve complex problems and provide answers more rapidly than traditional disk-based processing because data can be quickly pulled into memory.

- **The Hadoop framework** stores and processes large volumes of data on grids of low-cost commodity hardware.

“The concept of high-performance analytics is about using these high-performance computing techniques specifically with analytics in mind,” said Pope. “It’s a bit of a nuance, but it refers to applying advanced analytics as a core piece of the infrastructure.”

Big data: The point when the volume, velocity, variability and variety of data exceeds an organization’s storage or compute capacity for accurate and timely decision making.

Quickly solve complex problems using big data and sophisticated analytics in a distributed, in-memory and parallel environment.
What Should You Capture, and What Should You Keep?

Technology enables you to capture every bit and byte, but should you? No. Not all of the data in the big data ocean will be relevant or useful. Organizations must have the means to separate the wheat from the chaff and focus on what counts, instead of boiling the proverbial ocean.

“Organizations shouldn’t try to analyze the world just to answer one question,” said Pope. “They need to first isolate the relevant data, then further refine the analysis, and be able to iterate large amounts of complex data. These requirements are not mere technical problems; they are central to creating useful knowledge that supports effective decisions.”

Smart Filters Identify What to Store

With smart content extraction, the organization captures and stores only what is suspected of being relevant for further processing, and filters out unnecessary documents during the initial retrieval. The goal is to reduce data noise and store only what is needed to answer business questions.

“Smart filters help identify the relevant data, so you don’t spend time searching large data stores simply because you don’t know what subsection of data could contain value,” said Pope. Smart filters can apply natural language processing (NLP) and advanced linguistic techniques to identify and extract only the text that is initially believed to be relevant to the business question at hand.

Pope provided an example of smart content extraction for a SAS customer that monitors scientific information sources across disciplines and media outlets to identify potential risks to food production, creating notifications and reports for advance notice to government and production agencies.

“This organization assesses more than 15 million unique texts looking for relationships between chemicals in the food production chain and possible side effects,” said Pope. “Historically, the organization was restricted to running this analysis once a month. Given that there’s a time value to safety-related information and reports, month-old data is not going to be as effective as more recent data, especially if there could be public health risks at stake.”

Now the organization can customize information retrieval calls on those millions of texts across the entire food chain, honing in on the most relevant information before download. As search functions crawl the Web, smart filters with embedded extraction rules filter out the irrelevant content. “This customer found out that only about 10 percent of the data they previously stored was what they were interested in,” said Pope. “By narrowing down the data store and analysis to that critical 10 percent, they can now report much more frequently and deliver better and more timely alerts of emerging contaminants or other safety risks, for government agencies to take action.”

“Say it takes five hours to run a marathon. You can train and train and train and make small, incremental performance improvements. Or you can get 26 people to each run one mile. The marathon is completed much faster and there is no single point of failure. That is essentially what grid computing does.”

David Pope
Principal Solutions Architect, SAS High-Performance Analytics
Smart Filters Determine Where to Keep What You Capture

In addition to identifying the most relevant nuggets of information from the available universe of information, smart filters can help determine where to store this data. Is it highly relevant? Then you’d want to have it readily accessible in an operational database type of storage. Or is it lower relevance? If so, it can be stored in lower-cost storage, such as a Hadoop cluster.

Now organizations have a way to analyze data up front, determine its relative importance, and use analytics to support automated processes that move data to the most appropriate storage location as it evolves from low to high relevance, or vice versa.

Capture and Correlate Data on the Fly

Often it’s not a matter of storing the data somewhere, but how to manage it in flight, for instance, when capturing website activity to optimize the online customer experience. “We may be capturing deep and broad information about a person or product from the Web or other sources – getting complete and accurate, detailed data on everything they view, everything they do and everything that happens, timed to the millisecond,” said Pope. “Once we bring in that data from online applications, we want to be able to tie it to other data sources. We might want to tie it to the customer relationship management system, or to an in-store promotion or contact center script. So the big data challenge is two-pronged: There’s a need for extremely high efficiency in processing data into insight, and speed in delivering that insight to the point of action.”

From Hindsight to Insight to Foresight

“Raw data has the potential to do a lot of things, ranging from static reporting about what happened in the past to predictive insight about what will happen in the future,” said Pope. “Business intelligence (BI) helps keep your business running this year; business analytics help you keep running your business three to five years from now.”

Most companies that think they have analytics actually just have operational reports that tell them about what has happened in the past. Such hindsight reports are important to an organization, because they describe the current pulse of the organization and inform decisions to react to it. For instance, you may need to know how many people have downloaded articles that mention your company, how customer sentiment about your brand has changed in social media, and which keywords drive the best prospects to your website.

“A proactive report, on the other hand, not only gives you that operational view of what happened in the past or present – such as how many website visitors downloaded which articles – but also gives you a prediction into the future – what visitors will most likely want to download next week. You gain foresight to help determine which content to generate, how to optimize the website design and so on.”

“The big data challenge is two-pronged: There’s a need for extremely high efficiency in processing data into insight, and for speed in delivering that insight to the point of action.”

David Pope
Principal Solutions Architect, SAS High-Performance Analytics
Is your organization using the data for hindsight as well as foresight? And is it using all the data it could to its best advantage? If we can assume that (A) more data can lead to more insight and hence is better than less data, and (B) analytics provides more forward-looking insight than point-in-time reporting, then the business value the organization gets from its data can be conceptualized in four quadrants. The sophistication of the data infrastructure is plotted on the x axis, and the sophistication of analytic techniques on the y axis in Figure 1 below:

- The lower left quadrant represents traditional business intelligence – hindsight reporting on current or past conditions with conventional data volumes to get answers in established time frames.
- In the upper left quadrant, you have traditional analytic processing technologies performing more complex assessments, such as predictive modeling or forecasting – yielding good answers but often taking a long time to do it. “Predictive analytics on even small data can take a lot of computational power,” said Pope.
- The lower right quadrant represents the use of big data technologies to expedite hindsight reporting (or enable more iterations) with much more data. Better answers, delivered faster than conventional BI.
- The upper right quadrant is the sweet spot – big data analytics – the combination of big data technologies with predictive and hybrid analytics.

“This is where you start really getting the value out of your data,” said Pope. “The organization at this last stage can quickly solve complex problems using big data and sophisticated analytics in an unfettered manner. Big data analytics enables you to iterate on new scenarios with complex analytical computations, instantly explore and visualize all of the data, and rapidly solve very specific business challenges.”

Figure 1. The business value of data is a factor of processing capacity and analytic sophistication.
New Thinking About Data and Model Management

In an on-the-fly, on-demand data world, organizations may find themselves having to rethink how they do data preparation and how they manage the analytical models that transform data into insight.

Evolve from Being Data-Focused to Analytics-Focused

“In the typical IT-focused organization, application design is driven by a data focus,” said Pope. “This is not a slight on the IT organization, just that applications are designed for a known outcome that you want to deliver to the organization over and over again. That approach is great for automating repetitive delivery of a fact or a standard report, but it isn’t adaptable for developing new insights. If the data sources change, you would have to change all the models and applications as well.

“In an analytic organization, on the other hand, application design is driven by an analytics focus. End users are looking to the IT infrastructure to deliver new insights, not the same thing over and over. These new discoveries may arise from any type of data (often combinations of data), as well as different technologies for exploring and modeling various scenarios and questions. So there must be recognized interlinks between data, analytics and insights – and applications must make these connections accessible to users. With an analytics approach, you can add new data sources on the back end without having to change the application.”

Consider That Data Preparation Is Different for Analytics Than for Reporting

Different analytic methods require different data preparation. For example, with online analytical processing (OLAP) reporting, you would put a lot of effort into careful data cleansing, transformation through extract-transform-load (ETL) processes, dimension definition and so on.

However, with query-based analytics, users often want to begin the analysis very quickly in response to a sudden change in the business environment. The urgency of the analysis doesn’t allow time for much (if any) data transformation, cleansing and modeling. Not that you’d want to, because too much upfront data preparation may remove the data nuggets that would fuel discovery. For example, if you’re trying to identify fraud, you wouldn’t want a data cleansing routine to fix aberrations in names and addresses, since those very inconsistencies help spot potential fraud. For many such cases, you want to preserve the rich details in the relevant data that could reveal facts, relationships, clusters and anomalies.

Manage Models as Critical Information Assets

The proliferation of models – and the complexity of the questions they answer – call for a far more systematic, streamlined and automated way of managing the organization’s essential analytic assets. A predictive analytics factory formalizes ongoing processes for the requisite data management and preparation, model building, model management and deployment.
A predictive analytics factory closes the analytical loop in two ways, by:

- Providing a mechanism to automatically feed model results into decision-making processes – putting the model-derived intelligence to practical use.
- Monitoring the results of that intelligence to make sure the models continue to add value. When model performance has degraded – for example, due to customer behavior changes or changes in the marketplace – the model should be modified or retired.

**Use All the Data, if It Is Relevant**

Depending on your business goal, data landscape and technical requirements, your organization may have very different ideas about working with big data. Two scenarios are common:

- **In a complete data scenario**, entire data sets can be properly managed and factored into analytical processing, complete with in-database or in-memory processing and grid technologies.
- **Targeted data scenarios** use analytics and data management tools to determine the right data to feed into analytic models, for situations where using the entire data set isn’t technically feasible or adds little value.

The point is, you have a choice. Different scenarios call for different options. “Some of your analytic talent has been working under self-imposed or system-imposed constraints,” said Pope. “If you need to create subsets using analytics on huge data volumes, that is still valuable – if you’re doing it in a smart, analytically sound way. But when you do predictive modeling on all your data, and you have the infrastructure environment to support it, you don’t have to do all that work to find that valuable subset.”

**How to Get Started with Big Data Analytics**

**Determine the Analytical Maturity of the Organization**

Pope outlined a four-stage hierarchy that describes an organization’s maturity level in its use of analytics for decision making:

- **The Stage 1 organization** is analytically naive. Senior management has limited interest in analytics. Good luck with that.
- **The Stage 2 organization** uses analytics in a localized way. Line of business managers drive momentum on their own analytics projects, but there’s no enterprisewide cohesion, infrastructure or support.
- **The Stage 3 organization** has analytical aspirations. Senior executives are committed to analytics, and enterprisewide analytics capability is under development as a corporate priority.
- **A Stage 4 organization** uses analytics as a competitive differentiator. This organization routinely reaps the benefits of enterprisewide analytics for business benefit and continuous improvement.
Get Executive and Management Buy-In

That’s easy; show them the money, but pick the right emissary to do it. “The analysts who will be using high-performance technologies and big data analytics are typically not the best ones to explain the business value to executives,” said Pope. “If you tell executives you need to be able to do regression analysis, natural language processing or in-database computing, you will get kicked out of the boardroom pretty fast. As soon as you say ‘advanced analytics’ to non-statisticians, they stop listening, unless you can tie it to a business initiative.”

“Show them how predictive analytics will deliver better results, and how data-driven decisions will improve the day-to-day work of front-line employees and advance the organization’s overall agenda. You have to sell it, and you have to sell it iteratively, always tying text analytics and advanced analytics to the bottom line.” Maybe leave copies of Tom Davenport’s Competing on Analytics on their desks.

“Think of the billions of dollars organizations have spent on infrastructure that stores their data,” said Pope. “Storing data does not help you run your business; deriving insight from that data does. Yes there’s that initial investment in order to get insight from analytics, but most smart executives understand the difference between cost and value. When you actually apply analytics to a use case, it will pay for itself many times over.”

Consider an Analytics Center of Excellence

A center of excellence is a cross-functional team with a permanent, formal organizational structure that:

- Collaborates with the business stakeholders to plan and prioritize information initiatives.
- Manages and supports those initiatives.
- Promotes broader use of information throughout the organization through best practices, user training and knowledge sharing.

Several different types may exist within a single organization, Pope explained. For example, a data management center of excellence focuses on issues pertaining to data integration, data quality, master data, enterprise data warehousing schema, etc. A traditional business intelligence (BI) center of excellence focuses on reporting, querying and other issues associated with distributing information to business users across the organization. In contrast, an analytics center of excellence focuses on the proper use and promotion of advanced analytics, including big data analytics, to produce ongoing value to decision makers at both an operational and strategic level.

Forming an analytics center of excellence will not solve all the problems and challenges that may exist in the information environment today, but it will lead the way toward alignment – shaping the analytic evolution from project to process, from unit-level to enterprise-level perspective.

“Think of the billions of dollars organizations have spent on infrastructure that stores their data. Storing data does not help you run your business; deriving insight from that data does. … When you actually apply analytics to a use case, it will pay for itself many times over.”

David Pope
Principal Solutions Architect, SAS High-Performance Analytics Practice
Closing Thoughts

Big data technologies – such as grid computing, in-database analytics and in-memory analytics – can deliver answers to complex questions with very large data sets in minutes and hours, compared to days or weeks. You can also analyze all available data (not just a subset of it) to get more accurate answers for hard-to-solve problems, uncover new growth opportunities and manage unknown risks – all while using IT resources very effectively.

“Using a combination of advanced statistical modeling, machine learning and advanced linguistic analysis, you can quickly and automatically decipher large volumes of structured and unstructured data to discover hidden trends and patterns,” said Pope. “Whether you need to analyze millions of social media posts to determine sentiment trends, enrich your customer segmentation with information from unstructured sources, or distill meaningful insights from millions of documents and diverse content sources, big data technologies redefine the possibilities.”

About the Presenter

David Pope, Principal Solutions Architect, SAS® High-Performance Analytics

David Pope has more than 21 years of experience working with and at SAS, ranging from research and development to management information systems to working with sales and marketing for SAS High-Performance Analytics solutions. He has experience in multiple industries, including communications, media, finance, health care, government, retail and education. His background in data integration and business intelligence – combining expertise in statistics, modeling and forecasting – enables him to describe new or innovative ways to solve business issues.

About KMWorld Magazine

KMWorld is the leading publisher, conference organizer and information provider serving the knowledge management, content management and document management markets. KMWorld informs more than 50,000 subscribers about the components and processes — and related success stories — that together offer solutions for improving business performance.
For More Information

View the on-demand recording of the webinar, From Big Data to Meaningful Information: kmworld.com/Webinars/487-From-Big-Data-To-Meaningful-Information.htm

Read more about SAS High-Performance Analytics: sas.com/software/high-performance-analytics

Download the SAS white paper, Big Data Meets Big Data Analytics: sas.com/reg/wp/corp/46345

Learn more about text analytics capabilities from SAS: sas.com/text-analytics
About SAS

SAS has reinvented its architecture and software to satisfy the demands of big data, larger problems and more complex scenarios, and to take advantage of new technology advancements. SAS High-Performance Analytics is specifically designed for big data initiatives, with support for in-memory, in-database and grid computing.

SAS OnDemand delivers any SAS solution on a SAS-hosted infrastructure or private cloud. The SAS High-Performance Analytics solution on dedicated high-performance appliances provides yet another option for applying advanced analytics to big data.

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 60,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW® For more information on SAS® Business Analytics software and services, visit sas.com.