THE POWER TO KNOW®

Big Data Meets Big Data Analytics

Three Key Technologies for Extracting Real-Time Business Value from the Big Data That Threatens to Overwhelm Traditional Computing Architectures

WHITE PAPER

Table of Contents

Content for this paper, *Big Data Meets Big Data Analytics*, was provided by Mark Troester, IT/CIO Thought Leader and Strategist at SAS. Troester oversees the company’s marketing efforts for information management and for the overall CIO and IT vision. He began his career in IT and has worked in product management and product marketing for a number of startups and established software companies.

Introduction

> ***Wal-Mart handles more than a million customer transactions each hour and imports those into databases estimated to contain more than 2.5 petabytes of data.***
>
> ***Radio frequency identification (RFID) systems used by retailers and others can generate 100 to 1,000 times the data of conventional bar code systems.***
>
> ***Facebook handles more than 250 million photo uploads and the interactions of 800 million active users with more than 900 million objects (pages, groups, etc.) – each day.***
>
> ***More than 5 billion people are calling, texting, tweeting and browsing on mobile phones worldwide.***

Organizations are inundated with data – terabytes and petabytes of it. To put it in context, 1 terabyte contains 2,000 hours of CD-quality music and 10 terabytes could store the entire US Library of Congress print collection. Exabytes, zettabytes and yottabytes definitely are on the horizon.

Data is pouring in from every conceivable direction: from operational and transactional systems, from scanning and facilities management systems, from inbound and outbound customer contact points, from mobile media and the Web.

According to IDC, "In 2011, the amount of information created and replicated will surpass 1.8 zettabytes (1.8 trillion gigabytes), growing by a factor of nine in just five years. That's nearly as many bits of information in the digital universe as stars in the physical universe." (Source: IDC Digital Universe Study, sponsored by EMC, June 2011.)

The explosion of data isn't new. It continues a trend that started in the 1970s. What has changed is the velocity of growth, the diversity of the data and the imperative to make better use of information to transform the business.

The hopeful vision of big data is that organizations will be able to harvest and harness every byte of relevant data and use it to make the best decisions. Big data technologies not only support the ability to collect large amounts, but more importantly, the ability to understand and take advantage of its full value.

What Is Big Data?

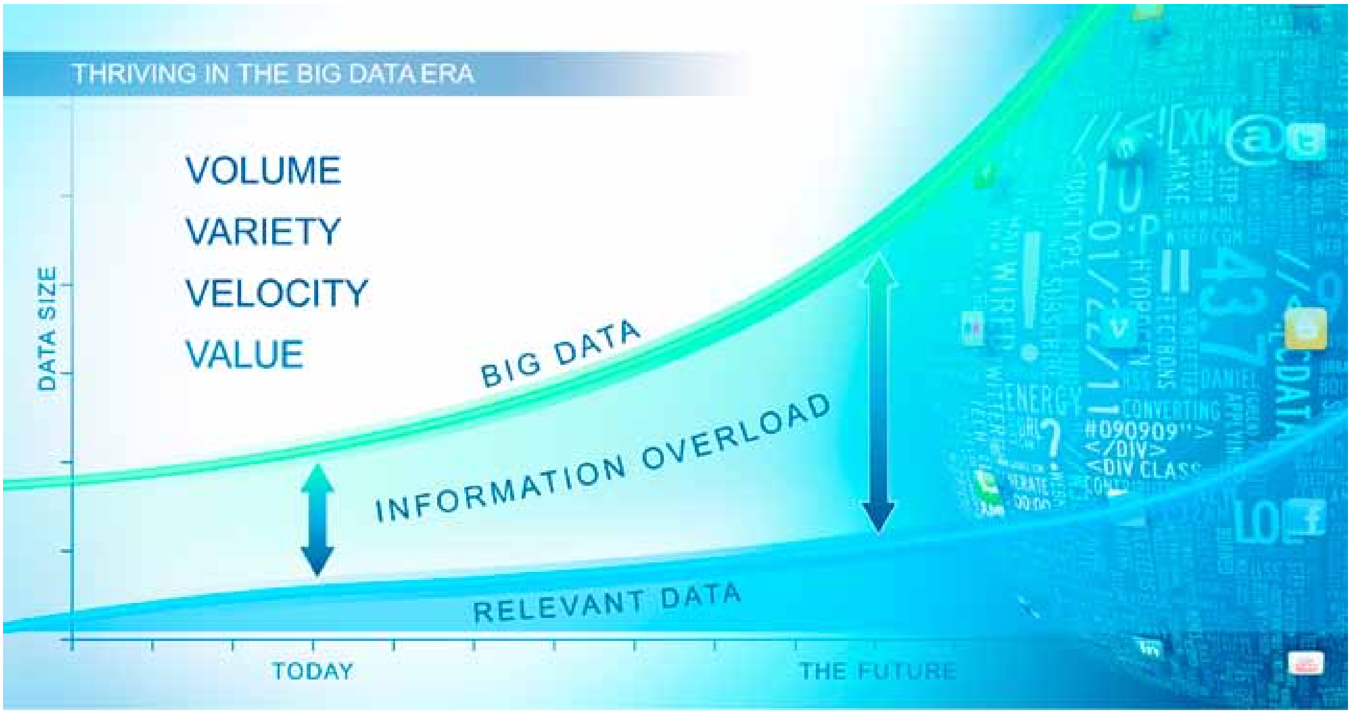
Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making.

Some of this data is held in transactional data stores – the byproduct of fast-growing online activity. Machine-to-machine interactions, such as metering, call detail records, environmental sensing and RFID systems, generate their own tidal waves of data. All these forms of data are expanding, and that is coupled with fast-growing streams of unstructured and semistructured data from social media.

That's a lot of data, but it is the reality for many organizations. By some estimates, organizations in all sectors have at least 100 terabytes of data, many with more than a petabyte. "Even scarier, many predict this number to double every six months going forward," said futurist Thornton May, speaking at a SAS webinar in 2011.

Big Data

When the volume, velocity, variability and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making.



Determining relevant data is key to delivering value from massive amounts of data.

However, big data is defined less by volume – which is a constantly moving target – than by its ever-increasing variety, velocity, variability and complexity.

- **Variety**. Up to 85 percent of an organization's data is unstructured – not numeric – but it still must be folded into quantitative analysis and decision making. Text, video, audio and other unstructured data require different architecture and technologies for analysis.

- **Velocity**. Thornton May says, "Initiatives such as the use of RFID tags and smart metering are driving an ever greater need to deal with the torrent of data in near-real time. This, coupled with the need and drive to be more agile and deliver insight quicker, is putting tremendous pressure on organizations to build the necessary infrastructure and skill base to react quickly enough."
- **Variability**. In addition to the speed at which data comes your way, the data flows can be highly variable – with daily, seasonal and event-triggered peak loads that can be challenging to manage.
- **Complexity**. Difficulties dealing with data increase with the expanding universe of data sources and are compounded by the need to link, match and transform data across business entities and systems. Organizations need to understand relationships, such as complex hierarchies and data linkages, among all data.

A data environment can become extreme along any of the above dimensions or with a combination of two or all of them at once. However, it is important to understand that not all of your data will be relevant or useful. Organizations must be able to separate the wheat from the chaff and focus on the information that counts – not on the information overload.

Big data refers to enormity in five dimensions:

- Volume – from terabytes to petabytes and up.
- Variety – an expanding universe of data types and sources.
- Velocity – accelerated data flow in all directions.
- Variability – inconsistent data flows with periodic peaks.
- Complexity – the need to correlate and share data across entities.

Rethinking Data Management

The necessary infrastructure that May refers to will be much more than tweaks, upgrades and expansions to legacy systems and methods.

"Because the shifts in both the amount and potential of today's data are so epic, businesses require more than simple, incremental advances in the way they manage information," wrote Dan Briody in *Big Data: Harnessing a Game-Changing Asset* (Economist Intelligence Unit, 2011). "Strategically, operationally and culturally, companies need to reconsider their entire approach to data management, and make important decisions about which data they choose to use, and how they choose to use them. … Most businesses have made slow progress in extracting value from big data. And some companies attempt to use traditional data management practices on big data, only to learn that the old rules no longer apply."

> **"Most businesses have made slow progress in extracting value from big data. And some companies attempt to use traditional data management practices on big data, only to learn that the old rules no longer apply."**
>
> **Dan Briody**
> "Big Data: Harnessing a Game-Changing Asset," Economist Intelligence Unit, 2011

Some organizations will need to rethink their data management strategies when they face hundreds of gigabytes of data for the first time. Others may be fine until they reach tens or hundreds of terabytes. But whenever an organization reaches the critical mass defined as big data for itself, change is inevitable.

From Standalone Disciplines to Integrated Processes

Organizations are moving away from viewing data integration as a standalone discipline to a mindset where data integration, data quality, metadata management and data governance are designed and used together. The traditional extract-transform-load (ETL) data approach has been augmented with one that minimizes data movement and improves processing power.

Organizations are also embracing a holistic, enterprise view that treats data as a core enterprise asset. Finally, many organizations are retreating from reactive data management in favor of a managed and ultimately more proactive and predictive approach to managing information.

Cheap storage has driven a propensity to hoard data, but this habit is unsustainable. What organizations need is a better information engineering pipeline and a better governance process.

From Sample Subsets to Full Relevance

The true value of big data lies not just in having it, but in harvesting it for fast, fact-based decisions that lead to real business value. For example, disasters such as the recent financial meltdown and mortgage crisis might have been prevented with risk computation on historical data at a massive scale. Financial institutions were essentially taking bundles of thousands of loans and looking at them as one. We now have the computing power to assess the probability of risk at the individual level. Every sector can benefit from this type of analysis.

"Big data provides gigantic statistical samples, which enhance analytic tool results," wrote Philip Russom, Director of Data Management Research for TDWI in the fourth quarter 2011 TDWI Best Practices Report, *Big Data Analytics*. "The general rule is that the larger the data sample, the more accurate are the statistics and other products of the analysis."

However, organizations have been limited to using subsets of their data, or they were constrained to simplistic analysis because the sheer volume of data overwhelmed their IT platforms. What good is it to collect and store terabytes of data if you can't analyze it in full context, or if you have to wait hours or days to get results to urgent questions? On the other hand, not all business questions are better served by bigger data. Now, you have choices to suit both scenarios:

Organizations do not have to grapple with overwhelming data volumes if that won't better serve the purpose. Nor do they have to rely solely on analysis based on subsets of available data.

- **Incorporate massive data volumes in analysis**. If the business question is one that will get better answers by analyzing all the data, go for it. The game-changing technologies that extract real business value from big data – all of it – are here today. One approach is to apply high-performance analytics to analyze massive amounts of data using technologies such as grid computing, in-database processing and in-memory analytics. SAS has introduced the concept of an analytical data warehouse that surfaces for analysis only the relevant data from the enterprise data warehouse, for simpler and faster processing.
- **Determine upfront which data is relevant**. The traditional modus operandi has been to store everything; only when you query it do you discover what is relevant. SAS provides the ability to apply analytics on the front end to determine data relevance based on enterprise context. This analysis can be used to determine which data should be included in analytical processes and which can be placed in low-cost storage for later availability if needed.

Three Key Technologies for Extracting Business Value from Big Data

According to Philip Carter, Associate Vice President of IDC Asia Pacific, "Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery and/or analysis." (Source: IDC. *Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO*, September 2011.) Furthermore, this analysis is needed in real time or near-real time, and it must be affordable, secure and achievable.

A "stream it, store it, score it" approach determines the 1 percent that is truly important in all the data an organization has. The idea is to use analytics to determine relevance instead of always putting all data in storage before analyzing it.

Fortunately, a number of technology advancements have occurred or are under way that make it possible to benefit from big data and big data analytics. For starters, storage, server processing and memory capacity have become abundant and cheap. The cost of a gigabyte of storage has dropped from approximately $16 in February 2000 to less than $0.07 today. Storage and processing technologies have been designed specifically for large data volumes. Computing models such as parallel processing, clustering, virtualization, grid environments and cloud computing, coupled with high-speed connectivity, have redefined what is possible.

Here are three key technologies that can help you get a handle on big data – and even more importantly, extract meaningful business value from it.

- **Information management for big data**. Manage data as a strategic, core asset, with ongoing process control for big data analytics.
- **High-performance analytics for big data**. Gain rapid insights from big data and the ability to solve increasingly complex problems using more data.
- **Flexible deployment options for big data**. Choose between options for on-premises or hosted, software-as-a-service (SaaS) approaches for big data and big data analytics.

Information Management for Big Data

Many organizations already struggle to manage their existing data. Big data will only add complexity to the issue. What data should be stored, and how long should we keep it? What data should be included in analytical processing, and how do we properly prepare it for analysis? What is the proper mix of traditional and emerging technologies?

Big data will also intensify the need for data quality and governance, for embedding analytics into operational systems, and for issues of security, privacy and regulatory compliance. Everything that was problematic before will just grow larger.

SAS provides the management and governance capabilities that enable organizations to effectively manage the entire life cycle of big data analytics, from data to decision. SAS provides a variety of these solutions, including data governance, metadata management, analytical model management, run-time management and deployment management.

With SAS, this governance is an ongoing process, not just a one-time project. Proven methodology-driven approaches help organizations build processes based on their specific data maturity model.

SAS® Information Management technology and implementation services enable organizations to fully exploit and govern their information assets to achieve competitive differentiation and sustained business success. Three key components work together in this realm:

- Unified **data management** capabilities, including data governance, data integration, data quality and metadata management.
- Complete **analytics management**, including model management, model deployment, monitoring and governance of the analytics information asset.
- Effective **decision management** capabilities to easily embed information and analytical results directly into business processes while managing the necessary business rules, workflow and event logic.

Quickly solve complex problems using big data and sophisticated analytics in a distributed, in-memory and parallel environment.

High-performance, scalable solutions slash the time and effort required to filter, aggregate and structure big data. By combining data integration, data quality and master data management in a unified development and delivery environment, organizations can maximize each stage of the data management process.

Stream it, score it, store it. SAS is unique for incorporating high-performance analytics and analytical intelligence into the data management process for highly efficient modeling and faster results.

For instance, you can analyze all the information within an organization – such as email, product catalogs, wiki articles and blogs – extract important concepts from that information, and look at the links among them to identify and assign weights to millions of terms and concepts. This organizational context is then used to assess data as it streams into the organization, churns out of internal systems, or sits in offline data stores. This up-front analysis identifies the relevant data that should be pushed to the enterprise data warehouse or to high-performance analytics.

High-Performance Analytics for Big Data

High-performance analytics from SAS enables you to tackle complex problems using big data and provides the timely insights needed to make decisions in an ever-shrinking processing window. Successful organizations can't wait days or weeks to look at what's next. Decisions need to be made in minutes or hours, not days or weeks.

High-performance analytics also makes it possible to analyze all available data (not just a subset of it) to get precise answers for hard-to-solve problems and uncover new growth opportunities and manage unknown risks – all while using IT resources more effectively.

Whether you need to analyze millions of SKUs to determine optimal price points, recalculate entire risk portfolios in minutes, identify well-defined segments to pursue customers that matter most or make targeted offers to customers in near-real time, high-performance analytics from SAS forms the backbone of your analytic endeavors.

To ensure that you have the right combination of high-performance technologies to meet the demands of your business, we offer several processing options. These options enable you to make the best use of your IT resources while achieving performance gains you never would have thought possible.

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

- **Grid computing**. A centrally managed grid infrastructure provides dynamic workload balancing, high availability and parallel processing for data management, analytics and reporting. Multiple applications and users can share a grid environment for efficient use of hardware capacity and faster performance, while IT can incrementally add resources as needed.
- **In-database processing**. Moving relevant data management, analytics and reporting tasks to where the data resides improves speed to insight, reduces data movement and promotes better data governance. Using the scalable architecture offered by third-party databases, in-database processing reduces the time needed to prepare data and build, deploy and update analytical models.
- **In-memory analytics**. Quickly solve complex problems using big data and sophisticated analytics in an unfettered manner. Use concurrent, in-memory, multiuse access to data and rapidly run new scenarios or complex analytical computations. Instantly explore and visualize data. Quickly create and deploy analytical models. Solve dedicated, industry-specific business challenges by processing detailed data in-memory within a distributed environment, rather than on a disk.
- **Support for Hadoop**. You can bring the power of SAS Analytics to the Hadoop framework (which stores and processes large volumes of data on commodity hardware). SAS provides seamless and transparent data access to Hadoop as just another data source, where Hive-based tables appear native to SAS. You can develop data management processes or analytics using SAS tools – while optimizing run-time execution using Hadoop Distributed Process Capability or SAS environments. With SAS Information Management, you can effectively manage data and processing in the Hadoop environment.

In addition, a new product from SAS provides a Web-based solution that leverages SAS high-performance analytics technologies to explore huge volumes of data in mere seconds. Using **SAS Visual Analytics**, you can very quickly see correlations and patterns in big data, identify opportunities for further analysis and easily publish reports and information to an iPad®. Because it's not just the fact that you have big data, it's what you can do with the data to improve decision making that will result in organizational gains. SAS can cut through the complexities of big data and identify the most valuable insights so decision makers can solve complex problems faster than ever before.

High-performance analytics from SAS is optimized to address new business requirements and overcome technical constraints. In addition, SAS is leading the way in empowering organizations to transform their structured and unstructured data assets into business value using multiple deployment options.

> **"Today's rapid pace of business requires operational analytics that deliver answers before a question becomes obsolete; the sooner you act on a decision, the greater its potential value. SAS High-Performance Analytics can turn any data, including big data assets, into quicker, better business decisions and ultimately competitive advantage."**
>
> **Dan Vesset,**
> Program Vice President,
> Business Analytics, IDC

Flexible Deployment Options for Big Data

Flexible deployment models bring choice. High-performance analytics from SAS can be deployed in the cloud (with SAS or another provider), on a dedicated high-performance analytics appliance or in the existing on-premises IT infrastructure – whichever best serves your organization's big data requirements.

Whatever the deployment environment – from a desktop symmetric multiprocessing (SMP) to massively parallel processing (MPP) running on tens, hundreds or even thousands of servers – high-performance analytics from SAS scales for the best performance. A flexible architecture enables organizations to take advantage of hardware advances and different processing options, while extending the value of original investments.

For some organizations, it won't make sense to build the IT infrastructure to support big data, especially if data demands are highly variable or unpredictable. Those organizations can benefit from cloud computing, where big data analytics is delivered as a service and IT resources can be quickly adjusted to meet changing business demands.

SAS Solutions OnDemand provides customers with the option to push big data analytics to the SAS infrastructure, greatly eliminating the time, capital expense and maintenance associated with on-premises deployments.

Flexible deployment models bring choice. High-performance analytics from SAS can be deployed in the cloud (with SAS or another provider), on a dedicated high-performance analytics appliance or in the existing on-premises IT infrastructure – whatever best serves your organization's big data requirements.

SAS Differentiators at a Glance

- **Flexible architecture approach**. SAS provides flexible architecture approaches that are optimized based on business requirements and technical constraints.
- **Ability to manage and leverage many models**. Multiple deployment models include on-premises, cloud-hosted or hybrid options that provide the flexible capabilities required in many big data scenarios.
- **Solutions that are enabled for big data**. SAS provides comprehensive big data analytics capabilities, from robust information management support (data, analytics and decision management) to high-performance analytics infrastructure support, big data visualization and exploration capabilities, solutions that integrate structured and unstructured data, and prepackaged business solutions.
- **Proven, trusted adviser status**. SAS is uniquely positioned to help organizations turn big data and big data analytics into business value and differentiation based on our unparalleled leadership, product and solution offerings, and domain expertise.
- **Comprehensive information management approach supports the entire analytics life cycle**. Our graduated big data analytics maturity curve approach allows organizations to address their current and future needs in an optimal fashion.

High-performance analytics lets you do things you never thought about before because the data volumes were just way too big. For instance, you can get timely insights to make decisions about fleeting opportunities, get precise answers for hard-to-solve problems and uncover new growth opportunities – all while using IT resources more effectively.

Conclusion

“One-third of organizations (34 percent) do big data analytics today, although it’s new,” wrote Russom of TDWI. “In other words, they practice some form of advanced analytics, and they apply it to big data. This is a respectable presence for big data analytics, given the newness of the combination of advanced analytics and big data.”

Given that more than one-third of organizations in Russom’s research reported having already broken the 10-terabyte barrier, big data analytics will see more widespread adoption. Organizations that succeed with big data analytics will be those that understand the possibilities, see through the vendor hype and choose the right deployment model.

Big Data and Big Data Analytics – Not Just for Large Organizations

If we define big data as the data volume, variety and velocity that exceed an organization’s ability to manage and analyze it in a timely fashion, then there are candidates in any industry. It doesn’t matter if the breaking point is reached at hundreds of gigabytes or tens or hundreds of terabytes. The principles that apply to big data and big data analytics are similar and can help the smaller organization extract more value from its data assets and IT resources.

It Is Not Just About Building Bigger Databases

Big data is not about the technologies to store massive amounts of data. It is about creating a flexible infrastructure with high-performance computing, high-performance analytics and governance – in a deployment model that makes sense for the organization.

SAS can run in a symmetric multiprocessing (SMP) or grid environment – on-premises, in a cloud environment or on an appliance. Organizations can choose the approach that meets their needs today and scales for the future.

Choose the Most Appropriate Big Data Scenario

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:

- **A complete data scenario** whereby 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** that 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.

SAS can help assess, provide guidance and deliver solutions that support the best approach for any organization.

> **“Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery and/or analysis.”**
>
> **Philip Carter,**
> Associate Vice President of IDC Asia Pacific
> “Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO,”
> September 2011

> **“The new technologies and new best practices are fascinating, even mesmerizing, and there’s a certain macho coolness to working with dozens of terabytes. But don’t do it for the technology. Put big data and discovery analytics together for the new insights they give the business.”**
>
> **Philip Russom,**
> Director of Data Management Research, TDWI
> “Big Data Analytics, TDWI Best Practices Report,” Fourth Quarter 2011

Moving Processing to the Data Source Yields Big Dividends

SAS was one of the first vendors to move data preparation and analytical processing to the actual data source, taking advantage of the massive parallel processing (MPP) capabilities in some databases. This approach eliminates the need to move the data, which in turn reduces demand on processing and network resources and accelerates performance. In-database processing will pay additional dividends as data volumes continue to grow.

Big Data and Big Data Analytics Don't Have to Be Difficult

Big data technologies don't have to be complex and require specialized skills. SAS provides an extensive array of preconfigured business solutions and business analytics solutions that greatly simplify the most complex analytical problems, including those based on big data. With cloud computing, big data analytics becomes an on-demand service. And of course, SAS offers technical support, professional services, training and partnerships to ease the way into big data analytics.

Closing Thoughts

Big data is not just about helping an organization be more successful – to market more effectively or improve business operations. It reaches to far more socially significant issues as well. Could we have foreseen the mortgage meltdown, the financial institution crisis and the recession, if only we had gotten our arms around more data and done more to correlate it? Could we trim millions of dollars in fraud from government programs and financial markets? Could we improve the quality and cost of health care and save lives?

The possibilities are wide open. At SAS, we are optimistic about the potential for deriving new levels of value from big data with big data analytics. That's why we reinvented our architecture and software to satisfy the demands of big data, larger problems and more complex scenarios, and to take advantage of new technology advancements.

High-performance analytics from SAS is specifically designed to support big data initiatives, with in-memory, in-database and grid computing options. SAS Solutions OnDemand delivers SAS solutions on an infrastructure hosted by SAS or on a private cloud. The SAS High-Performance Analytics solution for Teradata and EMC Greenplum appliances provides yet another option for applying high-end analytics to big data.

So, bring on the petabytes. Big data analytics has arrived.

Learn more

Explore SAS high-performance solutions to learn how to turn your big data into bigger opportunities.
sas.com/hpa

White paper:
SAS® High-Performance Analytics: What Could You Do with Faster, Better Answers? Transform Your Organization and Gain Competitive Advantage
sas.com/reg/wp/corp/41948

White paper:
In-Memory Analytics for Big Data: Game-Changing Technology for Faster, Better Insights
sas.com/reg/wp/corp/42876

About SAS

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 55,000 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**.

SAS Institute Inc. World Headquarters +1 919 677 8000
To contact your local SAS office, please visit: **sas.com/offices**