From Big Data to Meaningful Information with SAS® High-Performance Analytics

Silvia BOLOHAN, Sebastian CIOBANU
SAS Analytical Solutions Romania
Silvia.Bolohan@eur.sas.com; Sebastian.Ciobanu@sas.com

This paper is about the importance of Big Data and What You Can Accomplish with the data that counts. Until recently, organizations have been limited to using subsets of their data, or they were constrained to simplistic analyses because the sheer volumes of data overwhelmed their processing platforms. But, what is the point of collecting and storing terabytes of data if you can't analyze it in full context, or if you have to wait hours or days to get results? On the other hand, not all business questions are better answered by bigger data.
How can you make the most of all that data, now and in the future? It is a twofold proposition. You can only optimize your success if you weave analytics into your solution. But you also need analytics to help you manage the data itself. There are several key technologies that can help you get a handle on your big data, and more importantly, extract meaningful value from it.

Keywords: *big data, data visualization, high-performance analytics, grid computing, in-memory analytics, in database computing, Hadoop*

1 Introduction - Big Data

At first glance, the term seems rather vague, referring to something that is large and full of information. That description does indeed fit the bill, yet it provides no information on what Big Data really is. Big Data is often described as extremely large data sets that have grown beyond the ability to manage and analyze them with traditional data processing tools. Searching the Web for clues reveals an almost universal definition, shared by the majority of those promoting the ideology of Big Data, that can be condensed into something like this: Big Data defines a situation in which data sets have grown to such enormous sizes that conventional information technologies can no longer effectively handle either the size of the data set or the scale and growth of the data set. In other words, the data set has grown so large that it is difficult to manage and even harder to garner value out of it. The primary difficulties are the acquisition, storage, searching, sharing, analytics, and visualization of data.

There is much more to be said about what Big Data actually is. The concept has evolved to include not only the size of the data set but also the processes involved in leveraging the data. Big Data has even become synonymous with other business concepts, such as business intelligence, analytics, and data mining.

Paradoxically, Big Data is not that new. Although massive data sets have been created in just the last two years, Big Data has its roots in the scientific and medical communities, where the complex analysis of massive amounts of data has been done for drug development, physics modeling, and other forms of research, all of which involve large data sets. Yet it is these very roots of the concept that have changed what Big Data has come to be.

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.

2. 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.

A number of recent technology advancements enable organizations to make the most of big data and big data analytics:

- Cheap, abundant storage.
- Faster processors.
- Affordable open source, distributed big data platforms, such as Hadoop.
- Parallel processing, clustering, MPP, virtualization, large grid environments, high connectivity and high throughputs.
- Cloud computing and other flexible resource allocation arrangements.

The goal of all organizations with access to large data collections should be to harness the most relevant data and use it for better decision making.

“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.”[1]

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." [1]

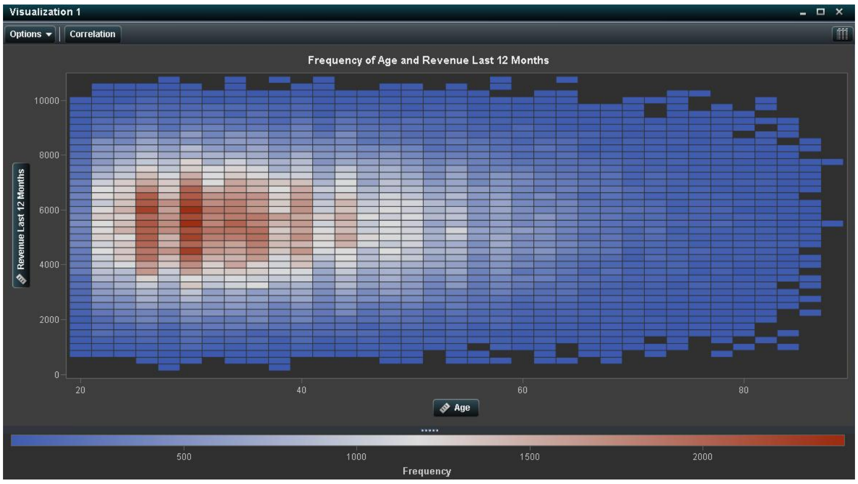

Fig. 1. Data Visualization: Correlation between Age & Revenue (in last 12 months)
Source: www.sas.com

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." [1]

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."

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."[1]

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.

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.[3]

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."[1]

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 enterprise-wide cohesion, infrastructure or support.
- The Stage 3 organization has analytical aspirations. Senior executives are committed to analytics, and enterprise-wide 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 enterprise-wide analytics for business benefit and continuous improvement.

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. 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. [4]

3. SAS® High-Performance Analytics

SAS High-Performance Analytics enables organizations to quickly and confidently seize new opportunities, make better choices ahead of the competition and create new value from big data. It will handle your most difficult challenges and quickly generate high-impact insights.

With SAS High-Performance Analytics you can:

- Get the timely insights needed to make decisions in an ever-shrinking window of opportunity. Processing that once took days or weeks now takes just minutes or seconds, enabling faster innovation.
- Discover precise answers for complex problems. Seize opportunities for growth that otherwise would remain unrecognized, and achieve better organizational performance.
- Optimally use and manage IT infrastructure resources to leverage big data and accommodate future growth while providing superior scalability and reliability.

3.1 Challenges

- Increasing volumes and varieties of data. Exploding data volumes hinder the completion of key analytic processes in a timely manner.
- Excessive data movement, and unnecessary data proliferation. Organizations struggle to determine what data should be stored where and for how long, what data should be used in analytical processing and how it should be prepared for analysis.
- Overwhelmed and poorly deployed IT resources. More requests for analytical processing mean longer waits for answers and unpredictable response times.
- Analytical processing complexities. The growing number of analytical models and data refreshes that are needed require an on-demand pool of distributed and parallel processing resources. Otherwise, it simply takes too long to get results.

3.2 How SAS® Can Help

Organizations are constantly seeking more effective ways to make decisions, relying increasingly on facts derived from a variety of data assets. But difficulties arise when data volumes grow ever-larger and there are hundreds or thousands of decisions to make each day.

Whether you need to analyze millions of price points, recalculate entire risk portfolios in minutes, identify well-defined customer segments or make attractive and targeted offers to customers in near-real time, SAS can help.

The scalability of SAS to handle huge volumes of data is unsurpassed. And SAS Analytics is considered best-in-class by both our customers and industry analysts. These advantages, combined with high-performance analytics, enable you to quickly exploit high-value opportunities from big data, while making the most of your existing investments or the latest advances in analytics infrastructure.

3.3 Benefits

- Immediately capture value and gain competitive advantage by exploiting big data, including existing information and new data collected from other sources, such as mobile devices and social media.
- Achieve incredibly fast response times and gain rapid insights to identify optimal actions and make the best decisions.
- Use more granular data and more complex analytical algorithms to produce new insights quickly, solve your most difficult problems, act confidently to seize new opportunities and better manage risks.

- Improve collaboration and productivity among your analytic and IT groups.
- Ensure data quality, improve data governance and enhance resource use by reducing data movement and redundancy.
- Quickly meet ever-changing business demands with flexible and dynamic workload balancing and high availability.
- Incrementally grow and optimize IT infrastructures to support faster time to value in a cost-effective manner.

Business Value

- Highly accurate and precise insights that lead to superior decisions.
- Near-real-time insights at the point of decision or embedded in business processes.
- The ability to act quickly and confidently to seize new opportunities and effectively manage risks.

IT Value

- Superior performance, scalability and reliability.
- Optimal resource usage.
- Better data governance.

3.4 Components

SAS Grid Computing enables you to automatically leverage a centrally managed grid infrastructure to achieve workload balancing, high availability of computing resources and parallel processing. Multiple applications and users can share a managed grid environment for better use of hardware capacity, while making incremental IT resource growth a possibility.

SAS Grid Manager allows individual SAS jobs to be split up, with each piece running in parallel across multiple SMP machines in the grid environment using shared physical storage. It enables organizations to create a managed, shared environment for processing large volumes of data and analytic programs. This makes it a perfect solution for managing multiple SAS users and jobs while enabling efficient use of IT resources and lower-cost commodity hardware.

SAS In-Database processing is a flexible, efficient way to get more value from increasing amounts of data by integrating select SAS technologies into your databases or data warehouses. It uses the massively parallel processing (MPP) architecture of the database or data warehouse for scalability and better performance.

Using SAS In-Database technologies, you can run scoring models, some SAS procedures and SQL queries inside a database. Moving relevant data integration, analytics and reporting tasks to where the data resides reduces unnecessary data movement, promotes better data governance and provides faster results.

SAS Scoring Accelerator takes SAS® Enterprise Miner™ models and publishes them as scoring functions inside a database. This exploits the parallel processing architecture offered by the database to achieve faster results. SAS Scoring Accelerator interfaces are currently available for Aster Data, EMC Greenplum, IBM DB2, IBM Netezza and Teradata. [2]

SAS Analytics Accelerator for Teradata is designed for users of SAS Enterprise Miner, SAS/STAT® and SAS/ETS® who want to build predictive and descriptive models for executing directly within the database environment. In-database analytics shortens the time needed to build, execute and deploy models, improving productivity for both analytic professionals and IT staff. They also help tighten data governance processes by giving analytic professionals access to consistent, fresh data.

SAS® In-Memory Analytics enables you to tackle previously unsolvable problems using big data and sophisticated analytics. It allows complex data exploration, model development and model deployment steps to be processed in-memory and distributed in parallel across a dedicated set of nodes. Because data can be quickly pulled into the memory, requests to run new scenarios or new analytical computations can be handled much faster and with better response times.

SAS High-Performance Analytics (available for EMC Greenplum and Teradata) is the only in-memory offering on the market that processes sophisticated analytics and big data to produce time-sensitive insights very quickly. SAS High-Performance Analytics is truly about applying high-end analytical techniques to solve complex business problems – not just about using query, reporting and descriptive statistics within an in-memory environment. For optimal performance, data is pulled and placed within the memory of a dedicated database appliance for analytic processing. Because the data is stored locally in the database appliance, it can be pulled into memory again for future analyses in a rapid manner.

SAS High-Performance Analytics addresses the entire model development and deployment life cycle. Unlike other offerings, SAS High-Performance Analytics can perform analyses that range from descriptive statistics and data summarizations to model building and scoring new data at breakthrough speeds.

SAS Visual Analytics is a high-performance, in-memory solution that empowers all types of users to visually explore big data, execute analytic correlations on billions of rows of data in minutes or seconds, gain insights into what the data means and deliver results quickly wherever needed.

Giving multiple users the ability to dynamically examine huge volumes of data simultaneously, combined with SAS software's powerful high-performance analytics, provides organizations with an unprecedented way to tap into big data and identify new and better courses of action more quickly. Users can easily spot opportunities for further investigation and analysis, and then convey visual results to decision makers via Web reports or the iPad.

SAS® In-Memory Industry Solutions:

SAS High-Performance Markdown Optimization is part of the SAS Revenue Optimization Suite. It analyzes massive amounts of data in parallel and enables retailers to identify and implement optimal pricing strategies. Retailers can quickly determine which products to mark down, how much to mark them down, and when and where to adjust pricing to maximize revenues.

SAS High-Performance Risk delivers faster risk calculations. Global market volatility and economic uncertainty require financial services firms to be quick and agile. SAS High-Performance Risk helps rapidly answer complex questions in areas such as market risk, counterparty exposure, liquidity risk management, credit risk, stress testing and scenario analysis.

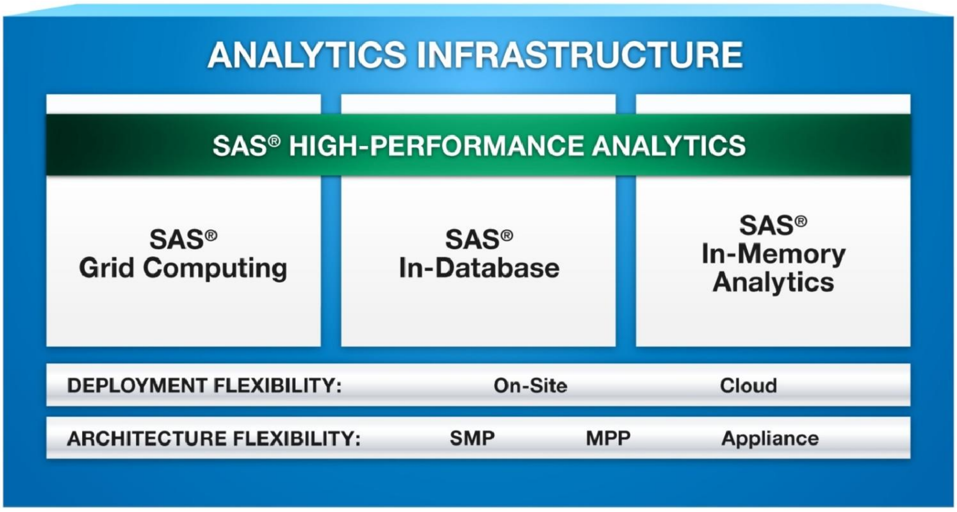

Fig. 2. SAS® High-Performance Analytics – Key Components
Source: www.sas.com

4. Conclusions

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. 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.

References

[1] *From Big Data to Meaningful Information* - Webinar: kmworld.com/Webinars/487-From-Big-Data-To-Meaningful-Information.htm, David Pope, Principal Solutions Architect, SAS® High-Performance Analytics

[2] *SAS High-Performance Analytics*: www.sas.com/hpa

[3] SAS white paper, *Big Data Meets Big Data Analytics: www.*sas.com/reg/wp/corp/46345

[4] *Thomas H. Davenport and Jill Dyche, "Big Data in Big Companies," May 2013.*

Silvia Bolohan is Marketing Manager at SAS Romania for 8 years. She leads and functions in the creation or production of marketing content for internal and external use in area of assignment. Silvia is responsible for developing and executing marketing strategy and/or programs for SAS products and services. Silvia has given support for data analysis based projects such as customer segmentation, attrition modeling, customer lifetime value, etc. She contributes to the efforts of building and maintaining a comprehensive reporting and tracking strategy for campaign response.

Sebastian CIOBANU is SAS consultant for over 4 years in the Business Intelligence and Data Integration domain for the Banking sector. Projects he has worked for include Analytical CRM solutions, Data Mining and Sales Data marts. He has a BA in Economic Informatics and MsC on Databases from the Academy of Economic Studies of Bucharest. His areas of interest are: Databases, Data Modeling, Business Intelligence solutions and the Banking area.