





The future of Big Data is Data Management

The future of Big Data is Data Management

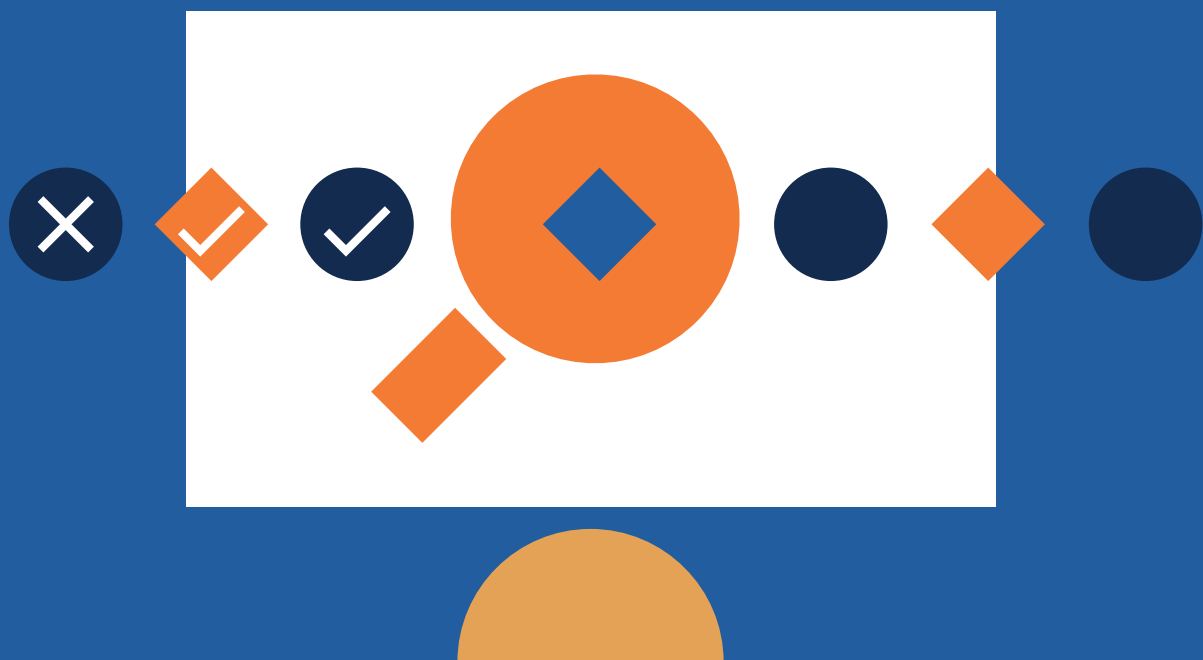
You are already aware of the importance of data, but do you truly use it effectively?

index

	Introduction Data Management: finding the gold in your data	<u>04</u>
	Section 1 Overview of Data Management and its objectives	<u>07</u>
	Section 2 What is Data Quality?	<u>10</u>
	Section 3 Data Integration	<u>15</u>

Introduction

Data Management: finding the gold in your data



The gold rush started in the United States in the middle of the nineteenth century, when a worker was lucky enough to find tiny specks of gold while digging a ditch. The discovery brought many changes, first in California and then in many other US states; the peak of the Gold Rush was when large gold fields were found in the Klondike in Canada at the end of the century. It is safe to say that, more than 250 years after James Marshall's discovery of gold in California, our outlook on the world in terms of requirements, technologies and behaviors have changed. **Meaningful data is the gold of our era** and understanding how to exploit it is the real socio-economic revolution of our century. Gold makes the owner a lucky person; the same principle does not apply to all data: companies can only maximize the value of their data if the data is collected, cleaned, interpreted and implemented correctly in the business.

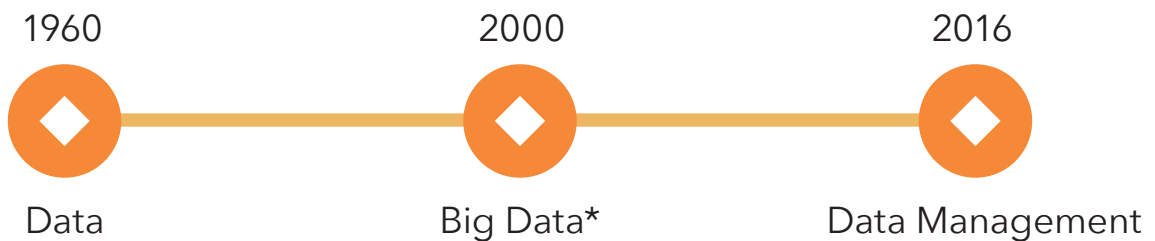
We've known about Big Data for many years now; they are nothing new. However, the way data are produced and processed has changed. **The growth of the Big Data is exponential** and has been going on since 2001; many of the connected devices that we use to work, play and exercise everyday continuously produce different types of data.

The situation has changed rapidly, though, and what at the beginning of the millennium looked like the next big thing, i.e. the possibility to collect and process huge amounts of data, has now evolved into the foundation of a more structured approach to data analysis. Data Management is a must-have for present and future trends and allows business processes and professionals

The future of Big Data is Data Management

to be supplied at all times with correct and functional data that are necessary to reach their objectives and **support strategic decisions.**

The future of Big Data is in the hands of Data Management. It must be applied not only to the existing corporate infrastructures and processes, but also to those that are implemented in the future.






* Academic origin of the term: <http://www.ssc.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF>

Section 1

Overview of Data Management and its objectives





Data Management is a process in which the data made available to the company processes and roles are always correct and useful for reaching business goals.

Data Management simplifies the work of data professionals (e.g. Data Analysts or Data Scientists) and supports the processes of producing **reliable data in the right place and at the right time.**

In the right place, because data is interconnected with multiple systems and departments within an organization: thanks to Data Management solutions, IT and Business areas no longer work as separate entities, but instead work together with a strategic approach. This is possible through the technique of combining tools, technologies and knowledge, which allows them to work to common purposes. At the right time because, thanks to this process which can be easily implemented in any data architecture, information can be used in *near real time*.

Similarly to gold, which must be accurately sifted, Data Management provides data professionals with data that have already been cleansed of “noxious” elements which might lead to unreliable results during the analytical process.

“Refined” data, ready for Analytics, are crucial for the survival of organizations and, at the same time, they must be easy and fast to access and use for the users. For the information to be effective, it must be contextualized; the data must be given to the user

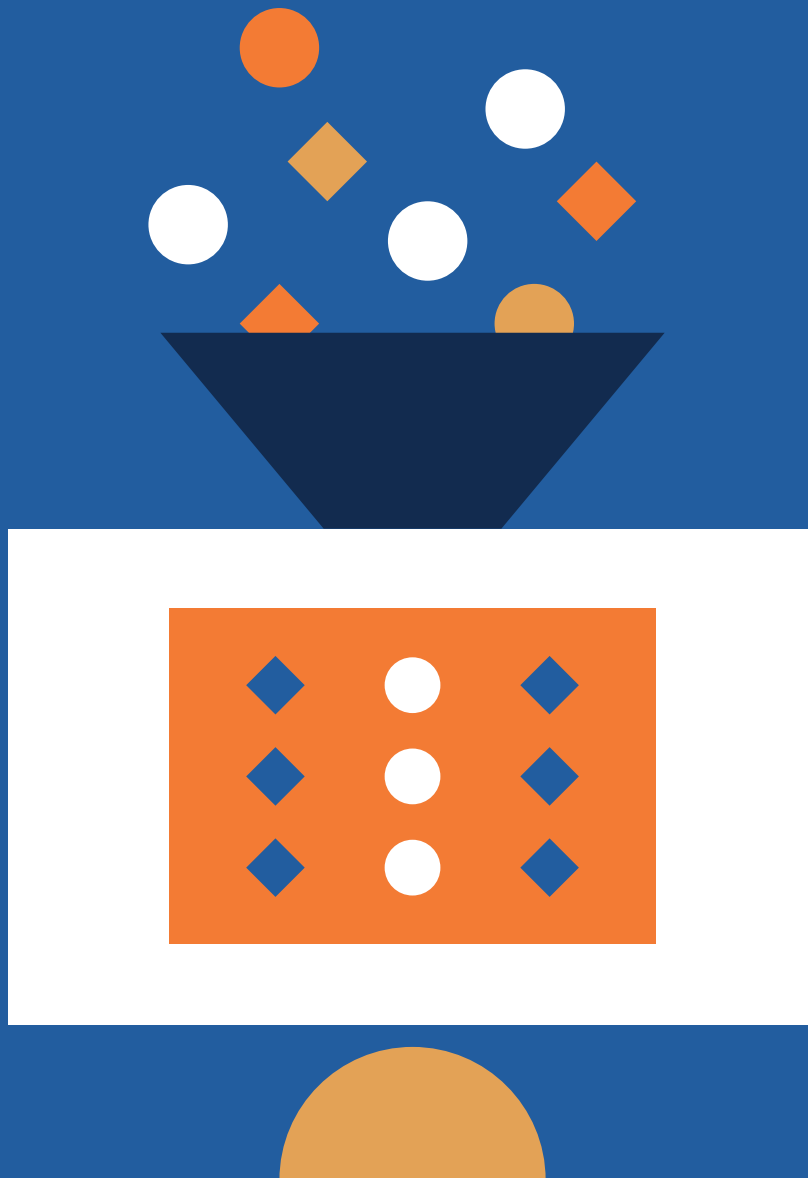
in a **defined format and in a consistent way**, after integrating and combining multiple sources. Here is a practical example to show the role that Data Management can play in daily business decisions: John is the manager of a large resort in Las Vegas, with approximately one hundred rooms, a wellness area with a spa and a pool, an exclusive restaurant and a Casino with Slot Machines and gambling tables. Thanks to all these services, John can count on a broad customer base, for example families visiting the city or professional poker players who need a room for the weekend. The hotel uses different data management systems for different services: if a traditional analytical approach is followed, data assignment problems might occur, due to the dissimilar nature of the sources being used. For example, John might lose a lot of money if, by mistake, the spa customers' data were associated with the people playing at the gambling tables. Data Management can be used to prevent John from running this risk. He will be get the accurate data that is necessary to make the appropriate business decisions. Moreover, if this is applied continuously, John will be able to tailor his customers' experience and better meet their needs.

Good Data Management is based on two key tools:
Data Quality and Data Integration.



Section 2

What is Data Quality?



Data Quality is a combination of processes that allow the value of data to be aligned with the business needs..

When choosing any product or service, a buyer often unknowingly thinks about how much he or she is willing to give up (in both a financial and non-financial sense) to get that commodity. Once the deal has been made, the buyer expects the commodity to be consistent with the price he/she paid for it and meet his/her usage expectations.

The same standard is also applied to the data: why commit large amounts of money to obtain huge quantities of data that do not generate an equivalent amount of business value?

Too much unreliable or “dirty” data is a huge waste of time for the company. In the business world, wasted time goes hand in hand with financial losses and a waste of resources.

Good Data Quality reduces the time needed for the manual checks and reconciliation of the data by 90%, so that analytical objectives can be accurately and rapidly reached.

Multichannel approaches, web marketing and apps continuously generate a huge amount of non-integrated data which may hide the nuggets of data gold that the organizations need for their strategic business decisions.

A Data Quality tool is essential to “clean” this data and

let “pure” information emerge, while getting rid of any dangerous elements which may jeopardize correct analysis.

Companies always want **to increase sales** and optimize the customer experience. Thanks to the introduction of Big Data, these objectives have become more achievable, but also more challenging.

Think about how people go about buying things today: millions of users search online for the product they are interested in and go to the either physical or virtual checkouts to finalize the purchase. They interact with companies through different channels and in different times. By **analyzing the company-customer touchpoints, the consumers’ habits and behaviors** can be identified, with the aim of developing customized and targeted marketing campaigns, thus improving the outcome and preventing wasted resources and useless costs.

Furthermore, customers talk to one another, exchange experiences and opinions on online goods and services (eg: via social media, communities and blogs). If the consumer is the center of a company’s business strategy, listening to the users is essential in order to obtain precious strategic insights and **develop the best offer and experience** for each customer.

Why is Data Quality important?

Data Quality is similar to the foundations of a building: its processes and instruments certify and control every stage of the data production and distribution cycle in order to guarantee the best possible quality for the data used by the business.

It is often seen, however, as a “tool of last resort” - used to only to solve problems such as managing “dirty” or incorrect data or data that needs to be fixed. This is not ideal approach: quality must be guaranteed all the time, in all company process and for all the users. It must become a habit, a daily “exercise”. This is the only way to increase the usability of data within company processes.

Upstream from Data Quality, the level of “appropriateness” of the data must be defined, according to the expected impact on the business decisions.

Different Data Quality measures are used to evaluate different types of Data Quality problems such as:



accuracy: is the data value correct?



completeness: to what extent are the data represented in my dataset?



consistency: can different data conflict with each other?



timeliness: are those data up-to-date?



uniqueness: are there any repetitions or duplicates of the same data?




To explain how the Data Quality works to provide accurate and reliable data for subsequent analyses, let's consider John's Resort Hotel once again. Upon arriving at the hotel, Michael T. Smith, a customer, gives the receptionist his ID, to be used for registration. Mr. Smith reserves a table for two at the hotel restaurant under the name Michael Smith and the following day he decides to register at the SPA using his nickname, Mike Smith. An efficient Data Quality approach will reduce ambiguities and Michael Smith will be recognized as one customer in the different data management systems, thus avoiding any errors in the downstream analytical phase.



Section 3

Data Integration





Data Integration is the combination of processes and technologies for the management of business data coming from different sources both within and from outside the company; the data is transformed and combined in a single format to harmonize them and maximize their value.

Data analysis has given companies a competitive edge on the market for many years. ETL (Extract, Transform and Load) has been a fundamental tool in this field for a long time. However, organizations have often had a hard time succeeding in implementing Data Integration projects. Access to data and processing it when it comes from different sources have often hindered implementations due to the selection of the incorrect tool for the purpose and the ease of using that specific tool with the different data sources.

Business and IT have traditionally worked in isolation and have independently chosen the tools that they considered the most appropriate ones for their goals. This has led to data and information that were difficult to integrate and also to very high ownership, maintenance and training costs.




Data Integration has recently been playing a more central role within the organizations, because its shared and optimized implementation affects all company processes, and promotes sharing of information and re-use of key components.

Nowadays, **standardization** is a key concept when

choosing a tool, starting with the same shared rules within an organization and going all the way through into production. Effective Data Integration, combined with appropriate Data Quality, unifies any type of data into one complete system, leading to lower costs and more benefits, from the operational and business decision-making point of view.

A large **variety** of Data Integration programs and/or business initiatives can be executed, keeping the data quality always at the center, at an integration strategy level. Any new data entering the company are not only added to the existing information, but also **increase the value** of that information, thanks to being combined with the current data..

The organizations that perform best are the ones that have already implemented a Data Integration-based strategy as a foundation for their competitive edge. This allows:

-  simplicity
-  flexibility
-  decreased reaction time

For an overview on Data Integration, let's take John's Las Vegas hotel into consideration again to understand the experience of Michael Smith, our reference customer, during his stay. Since quality data has already been guaranteed by using data quality, the restaurant, spa and casino data can easily be combined. This will provide an overview of his stay and, once the Data Management process has been completed, it will also allow us to understand whether Michael should be included as the target for a marketing campaign

offering a 100 USD casino bonus coupon for those guests who book a room for a minimum of two nights. Rather than gambling, Michael prefers to relax at the swimming pool after work. He might therefore be included in the target group that is interested in a discount on the resort spa and gym.

The democratization of Analytics and the creation of the Citizen Data Scientist

It is clear that a correct Data Management Strategy for the business requires the combination of different techniques and tools, such as Data Quality and Data Integration.

With an ever increasing amount of data and the improved opportunities offered by the Analytics to the business from the point of view of the strategic vision and decision-making process, the Data Management is a fundamental resource for the management of both **SME's** and large companies facing complex data sources and a large amount of data.

An effective and shared Data Management platform can act as a data filter and provides **guaranteed access and quality** as part of the same platform, so that the information to be passed on to downstream analytics can be processed and organized.

High-quality, integrated and properly managed data improve performance and decision-making.

Thanks to the new technologies and Open Source platforms like Hadoop, the integration of Data Management in the business is now simple and no reorganization of the IT department is needed. Self-service Data Integration and Data Quality solutions allow even non-experts to access and use Hadoop. Data is accessible to anyone and from any source; Data Scientists (versatile, creative professionals who are also data experts) will no longer have to go through the data preparation routine, which typically took 50% of their time. This time can now be used to generate *insights*.

The widespread presence of Data Management software, the increased amount of available data and the reduced storage costs are changing the way data are analyzed and approached. The viewpoint has changed: the need for data analysis has increased and business demand is targeted, at bringing more value into the company.

This trend has been defined as the “democratization of analytics” in a document from the International Institute for Analytics and will favor the creation of a new class of analysts: the **Citizen Data Scientist**.

The Citizen Data Scientist has a cross-section of knowledge about the business and a solid base of mathematics and statistics. The Citizen Data Scientist is a curious person who will act as a mediator between two main “sides” in the Big Data world: IT and Business. These two functions do not communicate easily because they speak two different (professional) languages, but they have always had to meet each other halfway in order to allow the company to grow.

The Citizen Data Scientist is crucial for small and medium enterprises that lack the necessary resources to support the cost of a Data Scientist, who is more difficult to recruit, due to their specific mathematical, statistical

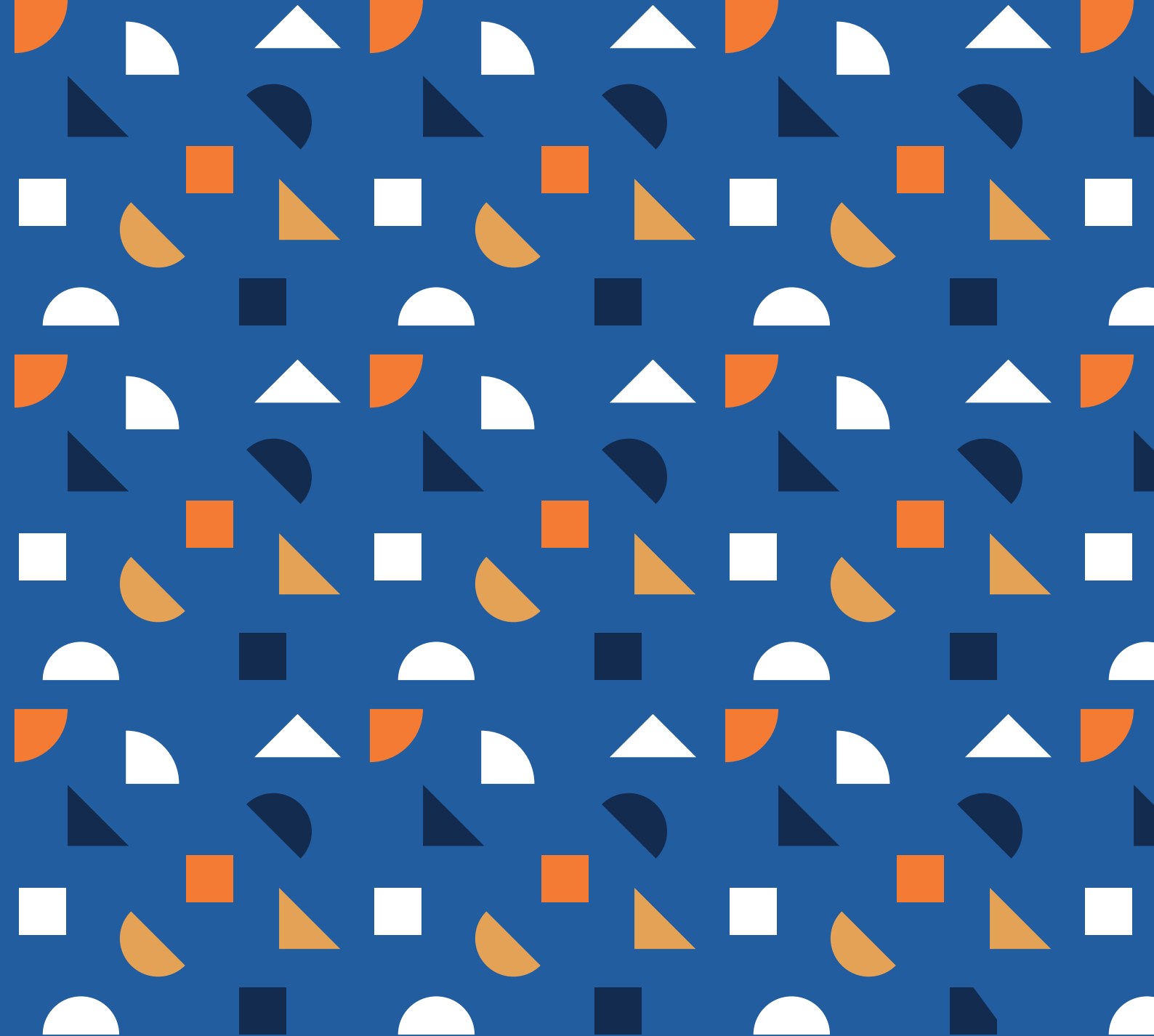
marketing and communication knowledge. However the Citizen Data Scientist alone cannot accomplish much and his/her work must be supported by high performance analytical tools.



This is the need that the Data Management can help customers meet: **a unique work environment, integrated with the present and future business** analytical data infrastructures, which communicates with other company software, such as the CRM, for advanced data analysis or visualization.

Data Management renders analyses reliable thanks to the integration of a higher number of data sources and the guarantee of their quality.





THE
POWER
TO KNOW.

SAS Analytical Solutions SRL - 29 Sfanta Vineri Street, Bectro Center Building,
4th Floor, District 3, Bucharest, 030203 Romania www.sas.com/romania

office_romania@sas.com

WORLD HEADQUARTERS CARY, NC 27153 USA +1 919 677 8000