## Database Systems Journal BOARD

**Contact**
Calea Dorobanţilor, no. 15-17, room 2017, Bucharest, Romania
Web: http://dbjournal.ro
E-mail: editor@dbjournal.ro

# CONTENTS

# Enhancing the Ranking of a Web Page in the Ocean of Data

Hitesh KUMAR SHARMA
University of Petroleum and Energy Studies, India
hkshitesh@gmail.com

*In today's world, web is considered as ocean of data and information (like text, videos, multimedia etc.) consisting of millions and millions of web pages in which web pages are linked with each other like a tree. It is often argued that, especially considering the dynamic of the internet, too much time has passed since the scientific work on PageRank, as that it still could be the basis for the ranking methods of the Google search engine. There is no doubt that within the past years most likely many changes, adjustments and modifications regarding the ranking methods of Google have taken place, but PageRank was absolutely crucial for Google's success, so that at least the fundamental concept behind PageRank should still be constitutive. This paper describes the components which affects the ranking of the web pages and helps in increasing the popularity of web site. By adapting these factors website developers can increase their site's page rank and within the PageRank concept, considering the rank of a document is given by the rank of those documents which link to it. Their rank again is given by the rank of documents which link to them. The PageRank of a document is always determined recursively by the PageRank of other documents.*

***Keywords:*** *SEO, Search Engine, Page Rank, Inbound, Outbound, DBMS*

## 1 Introduction

PageRank was developed by Google founders Larry Page and Sergey Brin at Stanford. At the time that Page and Brin met, search engines typically linked to pages that had the highest keyword density, which meant people could game the system by repeating the same phrase over and over to attract higher search page results. The rapidly growing web graph contains several billion nodes, making graph-based computations very expensive. One of the best known web-graph computations is Page-Rank, an algorithm for determining the "importance" of Web pages. [7]. Page and Brin's theory is that the most important pages on the Internet are the pages with the most links leading to them. [1] PageRank thinks of links as votes, where a page linking to another page is casting a vote.

## 2. Page Rank

PageRank is the algorithm used by the Google search engine, originally formulated by Sergey Brin and Larry Page in their paper "The Anatomy of a Large-Scale Hypertextual Web Search Engine".It is based on the premise, prevalent in the world of academia, that the importance of a research paper can be judged by the number of citations the paper has from other research papers. Brin and Page have simply transferred this premise to its web equivalent: the importance of a web page can be judged by the number of hyperlinks pointing to it from other web pages. [2] Now web graph has huge dimensions and is subject to dramatic updates in terms of nodes and links, therefore the PageRank assignment tends to became obsolete very soon [4]. It may look daunting to non-mathematicians, but the PageRank algorithm is in fact elegantly simple and is calculated as follows:

$PR(A)=(1-d)+d(PR(T_1)/C(T_1)+...+PR(T_n)/C(T_n))$

where:
PR(A) is the PageRank of a page A
$PR(T_1)$ is the PageRank of a page $T_1$
$C(T_1)$ is the number of outgoing links from the page $T_1$, d is a damping factor in the range $0 < d < 1$, usually set to 0.85.The PageRank of a web page is therefore calculated as a sum of the PageRanks of all pages linking to it (its *incoming links*), divided by the number of links on each of those pages (its *outgoing links*). From a search engine marketer's point of view, this means there are two ways in which PageRank can affect the position of your page on Google:

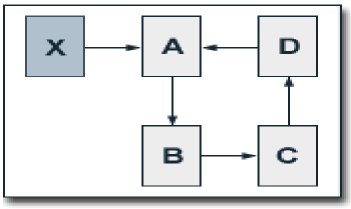
- *The number of incoming links.* Obviously the more of these the better. But there is another thing the algorithm tells us: no incoming link can have a negative effect on the PageRank of the page it points at. At worst it can simply have no effect at all.
- *The number of outgoing links on the page which points at your page.* The fewer of these the better. This is interesting: it means given two pages of equal PageRank linking to you, one with 5 outgoing links and the other with 10, you will get twice the increase in PageRank from the page with only 5 outgoing links. At this point we take a step back and ask ourselves just how important PageRank is to the position of your page in the Google search results. The next thing we can observe about the PageRank algorithm is that it has nothing whatsoever to do with relevance to the search terms queried. It is simply one single (admittedly important) part of the entire Google relevance ranking algorithm. Perhaps a good way to look at PageRank is as a multiplying factor, applied to the Google search results after all its other computations have been completed. The Google algorithm first calculates the relevance of pages in its index to the search terms, and then multiplies this relevance by the PageRank to produce a final list. The higher your PageRank therefore the higher up the results you will be, but there are still many other factors related to the positioning of words on the page which must be considered first.[2]

## 3. The Effect of Inbound Links

It has already been shown that each additional inbound link for a web page always increases that page's PageRank. Taking a look at the PageRank algorithm, which is given by:

$PR(A)=(1-d)+d(PR(T1)/C(T1)+...+PR(Tn)/C(Tn))$

one may assume that an additional inbound link from page X increases the PageRank of page A by d × PR(X) / C(X) where PR(X) is the PageRank of page X and C(X) is the total number of its outbound links. But page A usually links to other pages itself. Thus, these pages get a PageRank benefit also. If these pages link back to page A, page A will have an even higher PageRank benefit from its additional inbound link. The single effects of additional inbound links shall be illustrated by an example.



We regard a website consisting of four pages A, B, C and D which are linked to each other in circle. Without external inbound links to one of these pages, each of them obviously has a PageRank of 1. We now add a page X to our example, for which we presume a constant Page Rank PR(X) of 10. Further, page X links to page A by its only outbound link. Setting the damping factor d to 0.5, we get the following equations for the PageRank values of the single pages of our site:

$PR(A)=0.5+0.5(PR(X)+PR(D))=5.5+0.5\ PR(D)$

$PR(B)=0.5+0.5\ PR(A)$

$PR\copyright=0.5+0.5\ PR(B)$

$PR(D)=0.5+0.5\ PR\copyright$

Since the total number of outbound links for each page is one, the outbound links do not need to be considered in the equations. Solving them gives us the following PageRank values:

$PR(A)=19/3=6.33$

$PR(B)=11/3=3.67$

$PR\copyright=7/3=2.33$

$PR(D)=5/3=1.67$

We see that the initial effect of the additional inbound link of page A, which was given by

$d\times PR(X)/C(X)=0,5\times 10/1=5$

is passed on by the links on our site.

### *3.1 The Influence of the Damping Factor*

The degree of PageRank propagation from one page to another by a link is primarily determined by the damping factor d. If we set d to 0.75 we get the following equations for our above example:

$PR(A)=0.25+0.75(PR(X)+PR(D))=7.75+0.75PR(D)$

$PR(B)=0.25+0.75PR(A)$

$PR\copyright=0.25+0.75PR(B)$

$PR(D)=0.25+0.75PR\copyright$

Solving these equations gives us the following PageRank values:

$PR(A)=419/35=11.97$

$PR(B)=323/35=9.23$

$PR\copyright=251/35=7.17$

$PR(D)=197/35=5.63$

First of all, we see that there is a significantly higher initial effect of additional inbound link for page A which is given by

$d \times PR(X) / C(X) = 0.75 \times 10 / 1 = 7.5$

We remark that the way one handles the dangling node is crucial, since there can be a huge number of them. According to Kamvar et al. [Kamvar et al. 03b], a 2001 sample of the web containing 290 million pages had only 70 million nondangling nodes. This large amount of nodes without out-links includes both pages that do not point to any other page and also pages whose existence is inferred by hyperlinks but not yet reached by the crawler. Besides, a dangling node can represent a pdf, ps, txt, or any other file format gathered by a crawler but with no hyperlinks pointing outside [4]. This initial effect is then propagated even stronger by the links on our site. In this way, the PageRank of page A is almost twice as high at a damping factor of 0.75 than it is at a damping factor of 0.5. At a damping factor of 0.5 the PageRank of page A is almost four times superior to the PageRank of page D, while at a damping factor of 0.75 it is only a little more than twice as high. So, the higher the damping factor, the larger is the effect of an additional inbound link for the PageRank of the page that receives the link and the more evenly distributes PageRank over the other pages of a site.

### *3.2 The Actual Effect of Additional Inbound Links*

At a damping factor of 0.5, the accumulated PageRank of all pages of our site is given by

$PR(A)+PR(B)+PR(C)+PR(D)=14$

Hence, by a page with a PageRank of 10 linking to one page of our example site by its only outbound link, the accumulated PageRank of all pages of the site is increased by 10. (Before adding the link, each page has had a PageRank of 1.) At a damping factor of 0.75 the accumulated PageRank of all pages of the site is given by

$PR(A)+PR(B)+PR(C)+PR(D)=34$

This time the accumulated PageRank increases by 30. The accumulated PageRank of all pages of a site always increases by

(d/(1-d))×(PR(X)/C(X))

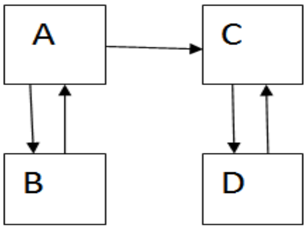
where X is a page additionally linking to one page of the site, PR(X) is its PageRank and C(X) its number of outbound links. The formula presented above is only valid, if the additional link points to a page within a closed system of pages, as, for instance, a website without outbound links to other sites. As far as the website has links pointing to external pages, the surplus for the site itself diminishes accordingly, because a part of the additional PageRank is propagated to external pages. The justification of the above formula is given by Raph Levien and it is based on the Random Surfer Model. The walk length of the random surfer is an exponential distribution with a mean of (d/(1-d)). When the random surfer follows a link to a closed system of web pages, he visits on average (d/(1-d)) pages within that closed system. So, this much more PageRank of the linking page – weighted by the number of its outbound links – is distributed to the closed system. For the actual PageRank calculations at Google, Lawrence Page und Sergey Brin claim to usually set the damping factor d to 0.85. Thereby, the boost for a closed system of web pages by an additional link from page X is given by

(0.85/0.15)×(PR(X)/C(X))=5.67×(PR(X)/C(X))

So, inbound links have a far larger effect than one may assume. [2]

## 4. The Effect Of Outbound Links

Since PageRank is based on the linking structure of the whole web, it is inescapable that if the inbound links of a page influence its PageRank, its outbound links do also have some impact. To illustrate the effects of outbound links, we take a look at a simple example.



We regard a web consisting of two websites, each having two web pages. One site consists of pages A and B, the other consists of pages C and D. Initially, both pages of each site solely link to each other. It is obvious that each page then has a PageRank of one.[6] Now we add a link which points from page A to page C. At a damping factor of 0.75, we therefore get the following equations for the single pages' PageRank values:

PR(A)=0.25+0.75PR(B)

PR(B)=0.25+0.375PR(A)

PR(C)=0.25+0.75PR(D)+0.375PR(A)

PR(D)=0.25+0.75PR(C)

Solving the equations gives us the following PageRank values for the first site:

PR(A)=14/23

PR(B)=11/23

We therefore get an accumulated PageRank of 25/23 for the first site. The PageRank values of the second site are given by

PR(C)=35/23

PR(D)=32/23

So, the accumulated PageRank of the second site is 67/23. The total PageRank for both sites is 92/23 = 4. Hence, adding a link has no effect on the total PageRank of the web. Additionally, the PageRank benefit for one site equals the PageRank loss of the other.

### *4.1 The Actual Effect of Outbound Links*

As it has already been shown, the PageRank benefit for a closed system of web pages by an additional inbound link is given by

```
(d/(1-d))×(PR(X)/C(X))
```

where X is the linking page, PR(X) is its PageRank and C(X) is the number of its outbound links. Hence, this value also represents the PageRank loss of a formerly closed system of web pages, when a page X within this system of pages now points by a link to an external page.

The validity of the above formula requires that the page which receives the link from the formerly closed system of pages does not link back to that system, since it otherwise gains back some of the lost PageRank. Of course, this effect may also occur when not the page that receives the link from the formerly closed system of pages links back directly, but another page which has an inbound link from that page. Indeed, this effect may be disregarded because of the damping factor, if there are enough other web pages in-between the link-recursion. The validity of the formula also requires that the linking site has no other external outbound links. If it has other external outbound links, the loss of PageRank of the regarded site diminishes and the pages already receiving a link from that page lose PageRank accordingly.[6]

Even if the actual PageRank values for the pages of an existing web site were known, it would not be possible to calculate to which extend an added outbound link diminishes the PageRank loss of the site, since the above presented formula regards the status after adding the link.

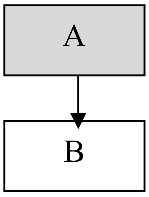
### *4.2 Intuitive Justification of the Effect of Outbound Links*

The intuitive justification for the loss of PageRank by an additional external outbound link according to the Random Surfer Modell is that by adding an external outbound link to one page the surfer will less likely follow an internal link on that page. So, the probability for the surfer reaching other pages within a site diminishes. If those other pages of the site have links back to the page to which the external outbound link has been added, also this page's PageRank will deplete.

We can conclude that external outbound links diminish the totalized PageRank of a site and probably also the PageRank of each single page of a site. But, since links between web sites are the fundament of PageRank and indispensable for its functioning, there is the possibility that outbound links have positive effects within other parts of Google's ranking criteria. Lastly, relevant outbound links do constitute the quality of a web page and a webmaster who points to other pages integrates their content in some way into his own site.

### *4.3 Dangling Links*

An important aspect of outbound links is the lack of them on web pages. When a web page has no outbound links, its PageRank cannot be distributed to other pages. Lawrence Page and Sergey Brin characterize links to those pages as dangling links.



The effect of dangling links shall be illustrated by a small example website. We take a look at a site consisting of three pages A, B and C. In our example, the pages A and B link to each other. Additionally, page A links to page C. Page C itself has no outbound links to other pages. At a damping factor of 0.75, we get the following equations for the single pages' PageRank values:

```
PR(A)=0.25+0.75PR(B)
PR(B)=0.25+0.375PR(A)
PR(C)=0.25+0.375PR(A)
```

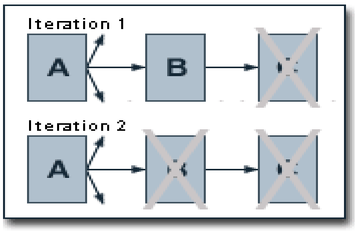
Solving the equations gives us the following PageRank values:

PR(A)=14/23

PR(B)=11/23

PR(C)=11/23

So, the accumulated PageRank of all three pages is 36/23 which is just over half the value that we could have expected if page A had links to one of the other pages. According to Page and Brin, the number of dangling links in Google's index is fairly high. A reason therefore is that many linked pages are not indexed by Google, for example because indexing is disallowed by a robots.txt file. Additionally, Google meanwhile indexes several file types and not HTML only. PDF or Word files do not really have outbound links and, hence, dangling links could have major impacts on PageRank.



In order to prevent PageRank from the negative effects of dangling links, pages without outbound links have to be removed from the database until the PageRank values are computed. According to Page and Brin, the number of outbound links on pages with dangling links is thereby normalized. As shown in our illustration, removing one page can cause new dangling links and, hence, removing pages has to be an iterative process. After the PageRank calculation is finished, PageRank can be assigned to the formerly removed pages based on the PageRank algorithm. Therefore, as many iterations are needed as for removing the pages. Regarding our illustration, page C could be processed before page B. At that point, page B has no PageRank yet and, so, page C will not receive any either. Then, page B receives PageRank from page A -and during the second iteration, also page C gets its PageRank. Regarding our example website for dangling links, removing page C from the database results in page A and B each having a PageRank of 1. After the calculations, page C is assigned a PageRank of 0.25+0.375PR(A)=0.625. So, the accumulated PageRank does not equal the number of pages, but at least all pages which have outbound links are not harmed from the dangling links problem. The definition of PageRank above has another intuitive basis in random walks on graphs. The simplified version corresponds to the standing probability distribution of a random walk on the graph of the Web. Intuitively, this can be thought of as modeling the behavior of a “random surfer”. The “random surfer” simply keeps clicking on successive links at random. However, if a real Web surfer ever gets into a small loop of web pages, it is unlikely that the surfer will continue in the loop forever. Instead, the surfer will jump to some other page [5]. By removing dangling links from the database, they do not have any negative effects on the PageRank of the rest of the web. Since PDF files are dangling links, links to PDF files do not diminish the PageRank of the linking page or site. So, PDF files can be a good means of search engine optimization for Google.

## 5. Conclusion

So what we conclude from here is the main factors influencing the page rank is the inbound links and the outbound links including the dangling links. Future work that can be done is the total no of pages affecting the page rank of a web site.

## References

[1] The PageRank Citation Ranking: Bringing Order to the Web (PDF, 1999) by Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd.

[2] Sergey Brin, Larry Page (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine". *Proceedings of the 7th international conference on World Wide Web (WWW)*.

[3] Taher Haveliwala and Sepandar Kamvar. (March 2003). "The Second Eigenvalue of the Google Matrix" *Stanford University Technical Report*: 7056.

[4] Gianna M. Del Corso, Antonio Gullí, Francesco Romani (2005). "Fast PageRank Computation via a Sparse Linear System".

[5] What can you do with a Web in your Pocket (PS, 1998) by Sergey Brin, Rajeev Motwani, Larry Page and Terry Winograd.

[6] Efficient Crawling Through URL rdering (PDF, 1998) by Junghoo Cho, Hector Garcia-Molina and Lawrence Page.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd. The Page- Rank citation ranking: Bringing order to the web. *StanfordDigital Libraries Working Paper*, 1998.

**Hitesh KUMAR SHARMA** is an Assistant Professor in University of Petroleum & Energy Studies, Dehradun. He has published 8 research papers in National Journals and 5 research papers in International Journal. Currently He is pursuing his Ph.D. in the area of database tuning**.**

# About Big Data and its Challenges and Benefits in Manufacturing

Bogdan NEDELCU
University of Economic Studies, Bucharest, Romania
bogdannedelcu@hotmail.com

*The aim of this article is to show the importance of Big Data and its growing influence on companies. It also shows what kind of big data is currently generated and how much big data is estimated to be generated. We can also see how much are the companies willing to invest in big data and how much are they currently gaining from their big data. There are also shown some major influences that big data has over one major segment in the industry (manufacturing) and the challenges that appear.*

***Keywords:*** *Big Data, manufacturing, challenges, benefits*

## 1 About Big Data

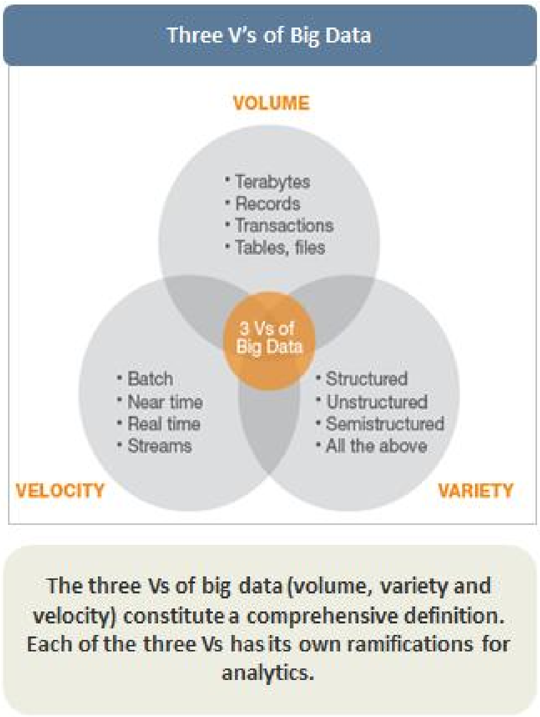
Big data is a new power that changes everything it interacts with and it is considered by some to be the electricity of the 21st century.

It was in the early 21st century when we first heard about the concept of big data. It was the first time when attributes like too large, too unstructured and too fast-moving where used for describing the nature of the data.

Big data's first main attribute is the volume. Some quantified data by counting there records, transactions, tables or file, but some found it more useful to quantify big data in terms of time. For example, in the U.S. some prefer to keep data available for legal analysis for seven years which is statute of limitations.

The second attribute is the variety of data. This happens because data come from a variety of sources like logs, streams, social media, text data, semi-structured data from B2B processes.

The last attribute of big data is the velocity which refers to the low-latency, real-time speed at which analytics need to be applied.



The three Vs of big data (volume, variety and velocity) constitute a comprehensive definition. Each of the three Vs has its own ramifications for analytics.

**Fig.1.** The three V's of Big Data [1]

Big data involves more than simply the ability to handle large volumes of data.

Firms like Google, eBay, LinkedIn and Facebook were the first organizations to embrace it, and were built from the beginning around big data. These firms had huge amounts of data in a new and less structured format (click streams, web server logs, social network relationships) and had no choice but to implement new technologies and management approaches.

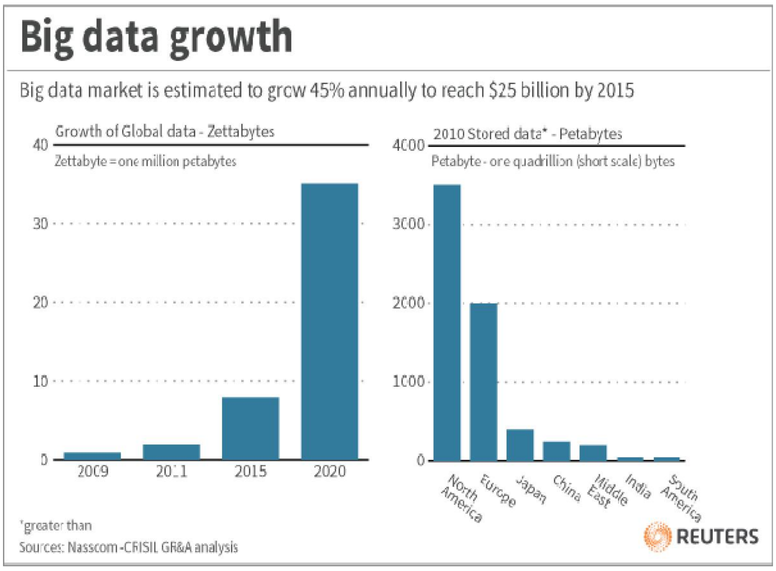
These companies were not the only ones facing such problems, in the travel industry everybody had the same issues. Every airline reservation, hotel stay, rental car or train reservation leaved a data trail, and that data over the years adds up to hundreds of terabytes or petabytes of structured transaction data.

Many research teams were assembled to gather information and study the total amount of data generated, stored, and consumed in the world. Though they had different estimates purposes and therefore their result vary, all point to an exponential growth in the future years. [2] MGI estimates that enterprises globally stored more than 7 exabytes of new data on disk drives in 2010, while consumers stored more than 6 exabytes of new data on devices such as PCs and notebooks. One xabyte of data is the equivalent of more than 4,000 times the information stored in the US Library of Congress [3]. Indeed, we are generating so much data today that it is physically impossible to store it all. [4]

Nowadays every sector in the global economy faces the big data problem. By 2009, nearly all sectors in the US economy had at least an average of 200 terabytes of stored data per company with more than 1,000 employees.

In total, European organizations have about 70% of the storage capacity of the entire United States at almost 11 exabytes.

According to Reuters [5], Big Data will grow from $3.2 billion in 2010 to reach a $25 billion industry by 2015.



Reuters graphic/Catherine Trevethan 05/10/12

**Fig.2.** Big Data Growth [5]

## 2. The Structure Of Big Data

Companies are using different types of data, therefore this data has been categorized based on the dimension of the data structure and on the dimension of the data source.

Based on the dimension of data structure, we distinguish structured (the data from fixed fields – spreadsheets or relational databases), unstructured (the data that does not reside in fixed fields – text from articles, email messages, untagged audio or video data, etc) and semi-structured data (the data that does not reside in fixed fields but uses tags or other markers to capture elements of the data – XML, HTML-tagged text).

Based on the dimension of data source, we distinguish internal data (gathered from a company's sales, customer services, employee records, etc.) and external data (gathered from sources outside a company such as third-party data providers, public social media sites, etc.).

Studies results across the globe shows as that 51% of data is structured, 27% of data is unstructured and 21% of data is semi-structured. [6]

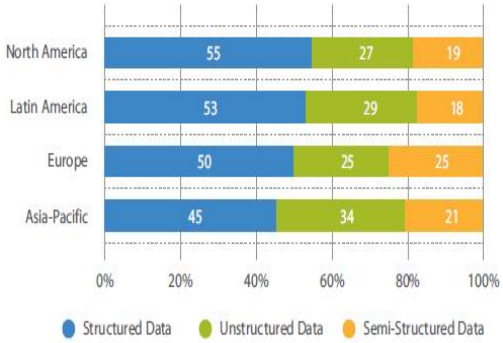


**Fig.3.** Data Structure Graphic [6]

Same reports reveal us that less than a quarter of the data was external.

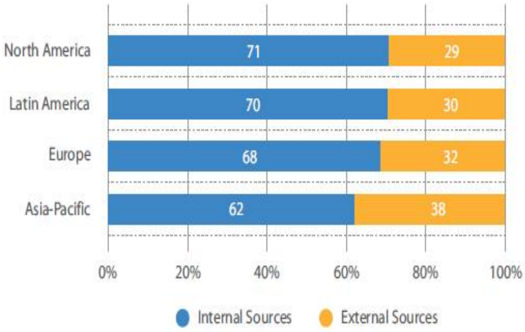


**Fig.4.** Data Sources Graphic [6]

## 3. The Value of Big Data

Using big data can be a key factor for companies in outperforming their competitors. It is estimated that a retailer embracing big data has the potential to increase its operating margin by more than 60 percent. More than that, big data creates new growth opportunities and entirely new categories of companies, such as those that aggregate and analyze industry data. Many of these will be companies that sit in the middle of large information flows where data about products and services, buyers and suppliers, and consumer preferences and intent can be captured and analyzed.

Many companies have access to valuable pools of data generated by their products and services. Networks will even connect physical products, enabling those products to report their own serial numbers, ship dates, number of times used, and so on.

It is important that all companies take big data seriously. In most industries, established competitors and new entrants alike will leverage data-driven strategies to innovate, compete, and capture value.

Based on analysis on the gathered data companies can design products that better match the needs of their customers.

Some executives are seeking data the organization has that might be of value to another organization, and from which the firm might be able to profit. That's the opportunity side. In 2012, about one-quarter of the companies surveyed by experts (27%) were capitalizing on this opportunity: selling their digital data. U.S. companies profited least from such data, with only 22% doing so. In contrast, half the Asia-Pacific companies sell their digital data. About one-quarter of European and Latin American companies sold their digital data in 2012.

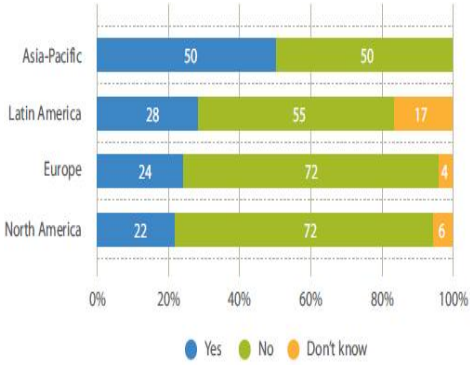


**Fig.5.** Big Data Selling Graphic [6]

For the approximately one-quarter of companies that sell their digital data, the annual revenue from selling such data was not trivial. In 2012, on an average, selling digital data contributed $21.6 million to the revenue of companies.

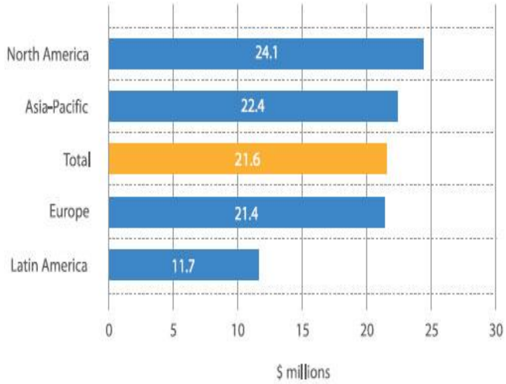


**Fig.6.** Big Data Revenue Graphic [6]

So clearly, some companies are profiting from their data. However, of the 73% of companies that did not sell such data, 22% said they do plan to sell such data by 2015; 55% don't; and 23% did not know. That means by 2015, 43% of companies will sell their digital data (the 27% that already do today, plus the 22% of the 73% that don't today) [6].

## 4. Investing in Big Data

According to researches the investments company made in Big Data were sizable. These investments were measured in two ways: by the median and the average survey respondent. The median spending on Big Data was $10 million, which was 0.14% of revenue (based on median revenue of survey respondents: $6.9 billion). The average survey respondent spending on Big Data was $88 million in 2012, which was 0.5% of average revenue (of $19 billion).

In 2012, 15% of the companies invested at least $100 million apiece on Big Data initiatives. About half of them (7%) invested at least $500 million each. However, on the other end of the spectrum were the 24% of companies that spent relatively little on Big Data – less than $2.5 million each.

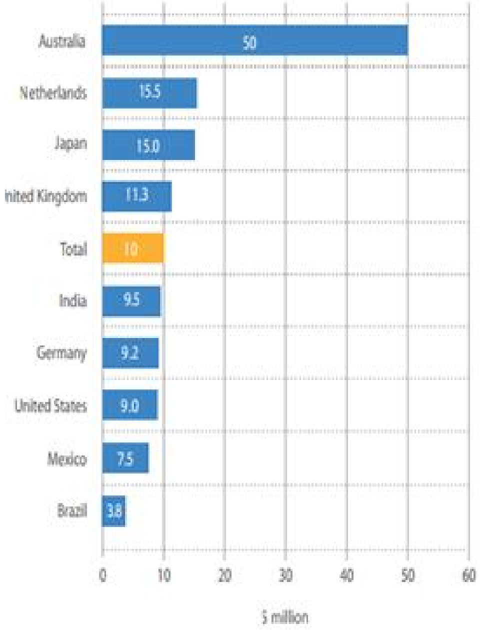


**Fig.7.** Big Data Investments Graphic [6]

According to the same report [6], by the year 2015, companies across the surveyed regions expect to spend 75% more on Big Data, with Australia and U.K. companies projecting the highest spending per company. Median spending across all countries is projected to increase by 75% to $17.5 million.

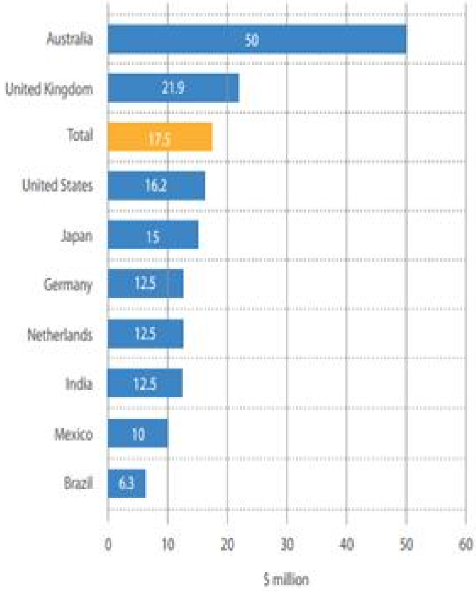


**Fig.8.** Big Data Spending Projecting Graphic [6]

## 5. Big data in manufacturing

The manufacturing sector was an early and intensive user of data to drive quality and efficiency, adopting information technology and automation to design, build, and distribute products since the dawn of the computer era. In the 1990s, manufacturing companies racked up impressive annual productivity gains because of both operational improvements that increased the efficiency of their manufacturing processes and improvements in the quality of products they manufactured.

The manufacturing sector has been the backbone of many developed economies.

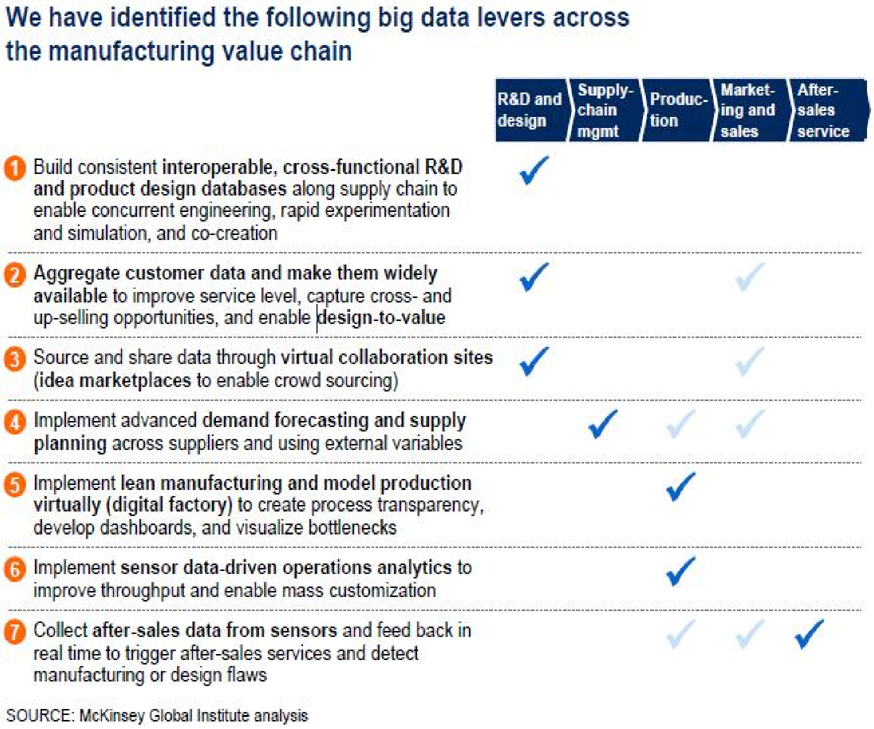
Increasingly global and fragmented manufacturing value chains create new challenges that manufacturers must overcome to sustain productivity growth. In many cases, technological change and globalization have allowed countries to specialize in specific stages of the production process.

To continue achieving high levels of productivity growth, manufacturers will need to leverage large datasets to drive efficiency across the extended enterprise and to design and market higher-quality products. The "raw material" is readily available; manufacturers already have a significant amount of digital data with which to work. Manufacturing stores more data than any other sector—close to 2 exabytes of new data stored in 2010. This sector generates data from a multitude of sources, from instrumented production machinery (process control), to supply chain management systems, to systems that monitor the performance of products that have already been sold (e.g., during a single cross-country flight, a Boeing 737 generates 240 terabytes of data).

And the amount of data generated will continue to grow exponentially. The number of RFID tags sold globally is projected to rise from 12 million in 2011 to 209 billion in 2021. IT systems installed along the value chain to monitor the extended enterprise are creating additional stores of increasingly complex data, which currently tends to reside only in the IT system where it is generated.

Manufacturers will also begin to combine data from different systems including, for example, computer-aided design, computer-aided engineering, computer-aided manufacturing, collaborative product development management, and digital manufacturing, and across organizational boundaries in, for instance, end-to-end supply chain data.

According to McKinsey Global Institute [7] analysis the following big data levers across the manufacturing value chain have been identified:



We have identified the following big data levers across the manufacturing value chain

| | R&D and design | Supply-chain mgmt | Produc-tion | Market-ing and sales | After-sales service |
|---|---|---|---|---|---|
| 1 Build consistent interoperable, cross-functional R&D and product design databases along supply chain to enable concurrent engineering, rapid experimentation and simulation, and co-creation | ✓ | | | | |
| 2 Aggregate customer data and make them widely available to improve service level, capture cross- and up-selling opportunities, and enable design-to-value | ✓ | | | ✓ | |
| 3 Source and share data through virtual collaboration sites (idea marketplaces to enable crowd sourcing) | ✓ | | | ✓ | |
| 4 Implement advanced demand forecasting and supply planning across suppliers and using external variables | | ✓ | ✓ | ✓ | |
| 5 Implement lean manufacturing and model production virtually (digital factory) to create process transparency, develop dashboards, and visualize bottlenecks | | | ✓ | | |
| 6 Implement sensor data-driven operations analytics to improve throughput and enable mass customization | | | ✓ | | |
| 7 Collect after-sales data from sensors and feed back in real time to trigger after-sales services and detect manufacturing or design flaws | | | ✓ | ✓ | ✓ |

SOURCE: McKinsey Global Institute analysis

**Fig.9.** Big Data Levers [7]

- Research and development and product design

The use of big data offers further opportunities to accelerate product development, help designers home in on the most important and valuable features based on concrete customer inputs as well as designs that minimize production costs, and harness consumer insights to reduce development costs through approaches including open innovation.

a. Product lifecycle management. Over decades, manufacturing companies have implemented IT systems to manage the product lifecycle including computer aided-design, engineering, manufacturing, and product development management tools, and digital manufacturing. However, the large datasets generated by these systems have tended to remain trapped within their respective systems. Manufacturers could capture a significant big data opportunity to create more value by instituting product lifecycle management (PLM) platforms that can integrate datasets from multiple systems to enable effective and consistent collaboration.

b. Design to value. While obtaining customer input through market research has traditionally been a part of the product design process, many manufacturers have yet to systematically extract crucial insights from the increasing volume of customer data to refine existing designs and help develop specifications for new models and variants. Best-in-class manufacturers conduct conjoint analyses to determine how much customers are willing to pay for certain features and to understand which features are most important for success in the market.

c. Open innovation. To drive innovation and develop products that address emerging customer needs, manufacturers are relying increasingly on outside inputs through innovative channels. With the advent of Web 2.0, some manufacturers are inviting external stakeholders to submit ideas for innovations or even collaborate on product development via Web-based platforms. Consumer goods companies such as Kraft and Procter and

Gamble invite ideas from their consumers as well as collaborate with external experts, including academics and industry researchers, to develop new products.

- Supply chain

Manufacturers, especially those producing fast-moving consumer goods, have significant additional opportunities to improve demand forecasting and supply chain planning. The volatility of demand has been a critical issue for manufacturers. Their retailing customers have pushed hard for increased flexibility and responsiveness from suppliers, given the diverging and ever-changing preferences of consumers. Other trends, such as the increasing use of promotions and tactical pricing, have only magnified volatility issues facing suppliers.

Manufacturers can improve their demand forecasting and supply planning by the improved use of their own data. But as we've seen in other domains, far more value can be unlocked when companies are able to integrate data from other sources including data from retailers, such as promotion data (e.g., items, prices, sales), launch data (e.g., specific items to be listed/delisted, ramp-up/ramp-down plans), and inventory data (e.g., stock levels per warehouse, sales per store). By taking into account data from across the value chain (potentially through collaborative supply chain management and planning), manufacturers can smooth spiky order patterns. The benefits of doing so will ripple through the value chain, helping manufacturers to use cash more effectively and to deliver a higher level of service.

- Production

Big data are driving additional efficiency in the production process with the application of simulation techniques to the already large volume of data that production generates. The increasing deployment of the "Internet of Things" is also allowing manufacturers to use real-time data from sensors to track parts, monitor machinery, and guide actual operations. [8]

a. Digital factory. Taking inputs from product development and historical production data (e.g., order data, machine performance), manufacturers can apply advanced computational methods to create a digital model of the entire manufacturing process.

b. Sensor-driven operations. The proliferation of Internet of Things applications allows manufacturers to optimize operations by embedding real-time, highly granular data from networked sensors in the supply chain and production processes. These data allows ubiquitous process control and optimization to reduce waste and maximize yield or throughput. They even allow for innovations in manufacturing that have not been possible thus far, including nano-manufacturing.

- Marketing and sales/after-sales support

As we have described, manufacturing companies are using data from customer interactions not only to improve marketing and sales but also to inform product development decisions. Increasingly, it is economically feasible to embed sensors in products that can "phone home," generating data about actual product usage and performance. Manufacturers can now obtain real-time input on emerging defects and adjust the production process immediately.

There are also many opportunities to leverage large datasets in the marketing, sales, and after-sales service activities. As we can observe in many sectors, opportunities range from the segmentation of customers to applying analytics in order to improve the effectiveness of sales forces. An increasingly important application for manufacturers is using sensor data from products once they are in use to improve service offerings.

## 6. Manufacturing/operations – Benefits and challenges

Manufacturing and production managers believe the greatest opportunities of Big Data for their function are to detect product defects and boost quality, and to improve supply planning. Better detection of defects in the manufacturing/production processes is next on the list.

A $2 billion industrial manufacturer said that analyzing sales trends to keep its manufacturing efficient was the main focus of its Big Data investments. The company's products are largely engineered to order. Understanding the behavior of repeat customers is critical to delivering in a timely and profitable manner. Most of its profitability analysis is to make sure that the company has good contracts in place. The company says its adoption of analytics has facilitated its shift to lean manufacturing, and has helped it determine which products and processes should be scrapped.

They see far less opportunity in using Big Data for mass customization, simulating new manufacturing processes, and increasing energy efficiency.

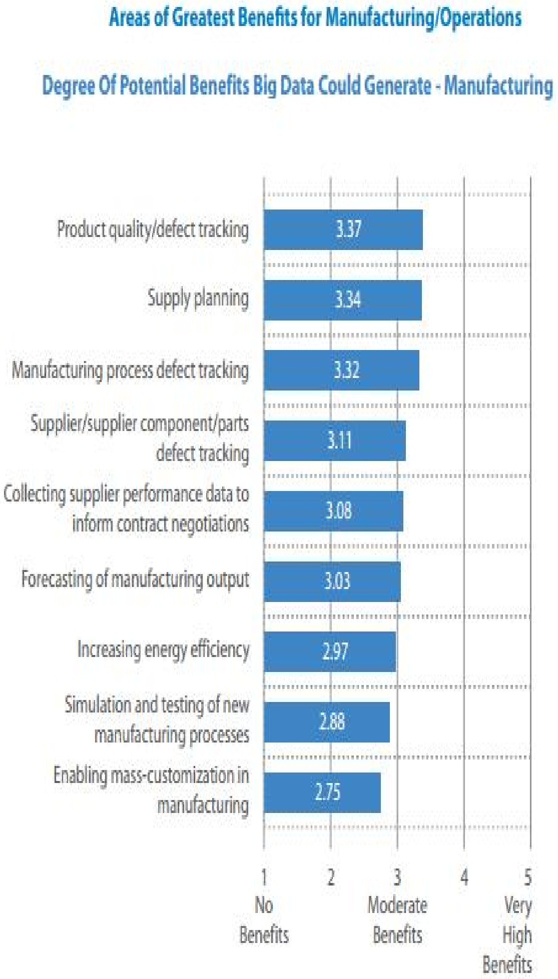


**Fig.10.** Greatest Benefits Areas for Manufacturing [7]

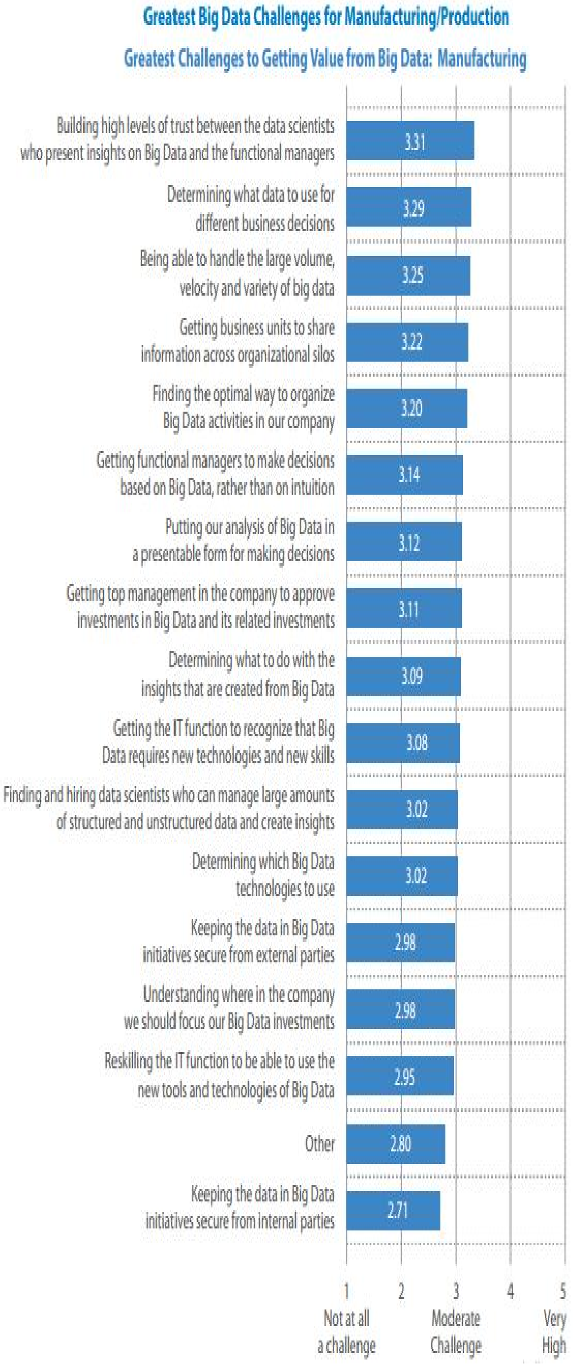


**Fig.11.** Big Data Challenges for Manufacturing [7]

Many of the levers also require access to data from different players in the value chain. To optimize production planning, data from various tiers of suppliers will be necessary. Demand planning will require customer data from retailers.

Manufacturing companies will also need to build the capabilities needed to manage big data. Despite the fact that the sector has been dealing with large datasets for two decades, the rising volume of data from new sources along the supply chain and from end markets requires a new level of storage and computing power and deep analytical expertise if manufacturers are to harvest relevant information and insights. There is a shortage of talent with the right experience for managing this level of complexity. Manufacturers will need not only to recruit new talent but also to remove organizational obstacles that today prevent such individuals from making maximum contributions.

Finally, where big data applications touch consumers and other end users, there are privacy issues. One of the most promising ideas is using product sensor data to create finely targeted after-sales services or cross-selling. But wielding this lever will be possible only if consumers don't object to suppliers monitoring how they use their products. Manufacturers must therefore address privacy concerns proactively, in collaboration with policy makers, and communicate with end users about choices and data transparency.

## 7. Conclusions

Big data allows organizations to create highly specific segmentations and to tailor products and services precisely to meet those needs. This approach is well-known in marketing and risk management but can be revolutionary elsewhere

Big data enables companies to create new products and services, enhance existing ones, and invent entirely new business models. Manufacturers are using data obtained from the use of actual products to improve the development of the next generation of products and to create innovative after-sales service offerings.

Manufacturers have tremendous potential to generate value from the use of large datasets, integrating data across the extended enterprise and applying advanced analytical techniques to raise their productivity both by increasing efficiency and improving the quality of their products. In emerging markets, manufacturers can begin to build competitive advantage that goes beyond their (thus far) relatively low labor costs. In developed markets, manufacturers can use big data to reduce costs and deliver greater innovation in products and services.

For manufacturers, opportunities enabled by big data can drive productivity gains both through improving efficiency and the quality of products. Efficiency gains arise across the value chain, from reducing unnecessary iterations in product development cycles to optimizing the assembly process. The real output value of products is increased by improving their quality and making products that better match customers' needs.

Data have become an important factor of production today—on a par with physical assets and human capital—and the increasing intensity with which enterprises are gathering information alongside the rise of multimedia, social media, and the Internet of Things will continue to fuel exponential growth in data for the foreseeable future. Big data have significant potential to create value for both businesses and consumers.

**References**

[1] Big Data Analytics - TWDI Research

[2] Peter Lyman and Hal Varian, "How much information?", 2003, the Library of Congress Web site

[3] John F. Gantz, David Reinsel, Christopher Chute, Wolfgang Schlichting, John McArthur, Stephen Minton, Irida Xheneti, Anna Toncheva, and Alex Manfrediz, "The expanding digital universe," IDC white paper, sponsored by EMC, March 2007

[4] Nasscom – Crisil GR&A Analysis

[5] The Emerging Big Returns on Big Data - Tata Consultancy Services

[6] McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity – June 2011

[7] "The Internet of Things," McKinsey Quarterly, March 2010.

**Bogdan NEDELCU** graduated Computer Science at Politehnica University of Bucharest in 2011. In 2013, he graduated the master program "Engineering and Business Management Systems" at Politehnica University of Bucharest. At present he is studying for the doctor's degree at the Academy of Economic Studies from Bucharest.

# 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]

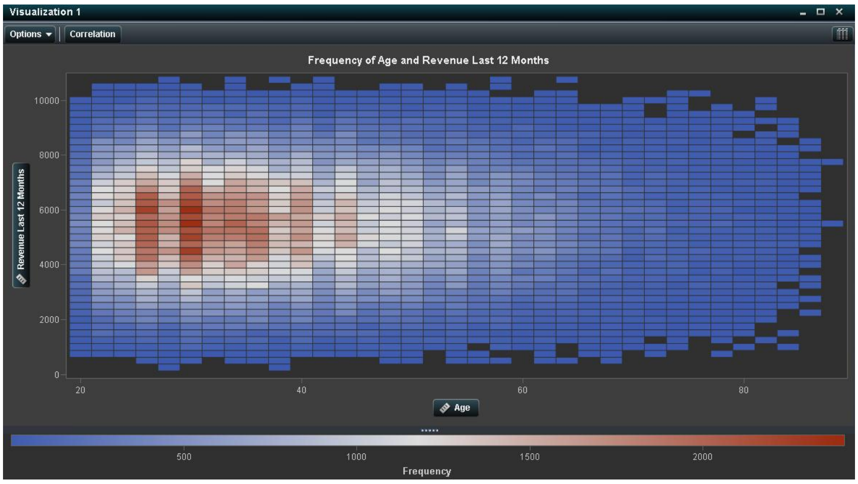


**Fig. 1.** Data Visualization: Correlation between Age & Revenue (in last 12 months)
*Source:* [*www.sas.com*](http://www.sas.com)

### 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.

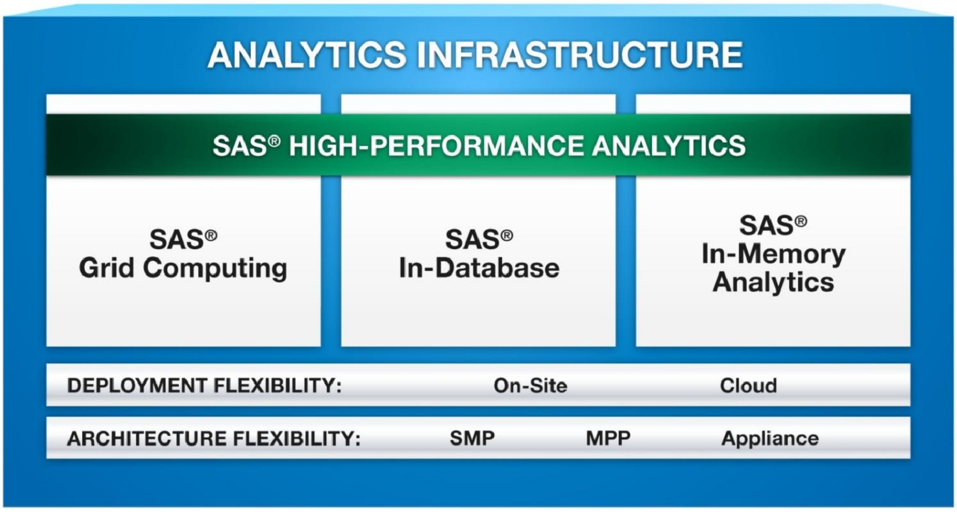

ANALYTICS INFRASTRUCTURE
SAS® HIGH-PERFORMANCE ANALYTICS
SAS® Grid Computing
SAS® In-Database
SAS® In-Memory Analytics
DEPLOYMENT FLEXIBILITY: On-Site Cloud
ARCHITECTURE FLEXIBILITY: SMP MPP Appliance

**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.

# Big Data Challenges

Alexandru Adrian TOLE
Romanian – American University, Bucharest, Romania
adrian.tole@yahoo.com

*The amount of data that is traveling across the internet today, not only that is large, but is complex as well. Companies, institutions, healthcare system etc., all of them use piles of data which are further used for creating reports in order to ensure continuity regarding the services that they have to offer. The process behind the results that these entities requests represents a challenge for software developers and companies that provide IT infrastructure. The challenge is how to manipulate an impressive volume of data that has to be securely delivered through the internet and reach its destination intact. This paper treats the challenges that Big Data creates.*

***Keywords***: *Big Data, 3V's, OLAP, security, privacy, sharing, value, infrastructure, technological solutions*

## 1 Introduction

Economic entities and not only, had developed over the years new and more complex methods that allows them to see market evolution, their position on the market, the efficiency of offering their services and/or products etc. For being able to accomplish that, a huge volume of data is needed in order to be mined so that can generate valuable insights.

Every year the data transmitted over the internet is growing exponentially. By the end of 2016, Cisco estimates that the annual global data traffic will reach 6.6 zettabytes[1]. The challenge will be not only to "speed up" the internet connections, but also to develop software systems that will be able to handle large data requests in optimal time.

To have a better understanding of what Big Data means, the table below represents a comparison between traditional data and Big Data (**Table 1.** Understanding Big Data).

**Table 1.** Understanding Big Data

| *Traditional Data* | *Big Data* |
|---|---|
| Documents | Photos |
| Finances | Audio and Video |
| Stock Records | 3D Models |
| Personnel files | Simulations |
| | Location data |

This example provides information about the volume and the variety of Big Data.

It is difficult to work with complex information on standard database systems or on personal computers. Usually it takes parallel software systems and infrastructure that can handle the process of sorting the amount of information that, for example, meteorologists need to analyze.

The request for more complex information is getting higher every year. Streaming information in real-time is becoming a challenge that must be overcome by those companies that provides such services, in order to maintain their position on the market.

By collecting data in a digital form, companies take their development to a new level. Analyzing digital data can speed the process of planning and also can reveal patterns that can be further used in order to improve strategies. Receiving information in real-time about customer needs is useful for seeing market trends and forecasting.

The expression "Big Data" also resides in the way that information is handled. For processing large quantities of data that is extremely complex and various there needs to be a set of tools that are able to navigate through it and sort it. The methods of sorting data differ from one type of data to another. Regarding Big Data, where the type of data is not singular, sorting is a multi-level process.

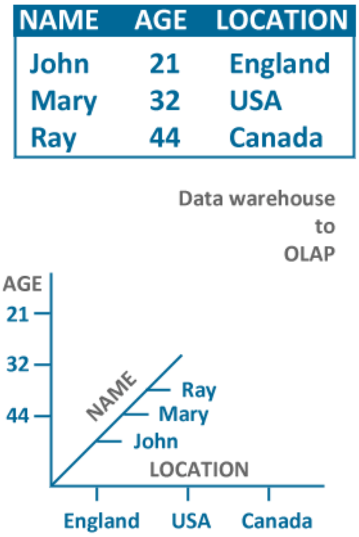
Big Data can be used for predictive analytics, an element that many companies rely on when it comes to see where they are heading. For example, a telecommunication company can use data stored from length of call, average text messages sent, average bill amount to see which customers are likely to discard their services.

## 2. Volume, Velocity, Variety

The “3V’s”, how Doug Laney calls them in his article *3-D Data Management: Controlling Data Volume, Velocity and Variety*, published in 2001, represents key elements that are considered vital regarding the characteristics of Big Data systems.

The first characteristic of Big Data, which is “**Volume**”, refers to the quantity of data that is being manipulated and analyzed in order to obtain the desired results. It represents a challenge because in order to manipulate and analyze a big volume of data requires a lot of resources that will eventually materialize in displaying the requested results. For example a computer system is limited by current technology regarding the speed of processing operations. The size of the data that is being processed can be unlimited, but the speed of processing operations is constant. To achieve higher processing speeds more computer power is needed and so, the infrastructure must be developed, but at higher costs.

By trying to compress huge volumes of data and then analyze it, is a tedious process which will ultimately prove more ineffective. To compress data it takes time, almost the same amount of time to decompress it in order to analyze it so it can be displayed, by doing this, displaying the results will be highly delayed. One of the methods of mining through large amount of data is with OLAP solutions (Online Analytical Processing) (**Fig.1.** Data warehouse -> OLAP). An OLAP solution consists of tools and multidimensional databases that allow users to easily navigate and extract data from different points of view. Therefore, it identifies relations between elements in the database so it can be reached in a more intuitive way. An example of how OLAP systems are rearranging the data imported from a data warehouse is below. For obtaining results various OLAP tools are used in order for the data to be mined and analyzed.



| NAME | AGE | LOCATION |
|---|---|---|
| John | 21 | England |
| Mary | 32 | USA |
| Ray | 44 | Canada |

**Fig. 1.** Data warehouse -> OLAP

“**Velocity**” is all about the speed that data travels from point A, which can be an end user interface or a server, to point B, which can have the same characteristics as point A is described. This is a key issue as well due to high requests that end users have for streamed data over numerous devices (laptops, mobile phones, tablets etc.). For companies this is a challenge that most of them can’t keep up to. Usually data transfer is done at less than the capacity of the systems. Transfer rates are limited but requests are unlimited, so streaming data in real-time or close to real-time is a big challenge. The only solution at this point is to shrink the data that is being sent. A good example is Twitter. Interaction on Twitter consists of text, which can be easily compressed at high rates. But, as in the case of “Volume” challenge, this operation is still time-consuming and there will still be delay in sending-receiving data. The only solution to this right now is to invest in infrastructure.

"**Variety**" is the third characteristic of Big Data. It represents the type of data that is stored, analyzed and used. The type of data stored and analyzed varies and it can consist of location coordinates, video files, data sent from browsers, simulations etc. The challenge is how to sort all this data so it can be "readable" by all users that access it and does not create ambiguous results. The mechanics of sorting has two key variables at the beginning: the system that transmits data and the system that receives it and interpret it so that can be later displayed (**Fig. 2.** Send-Receive).

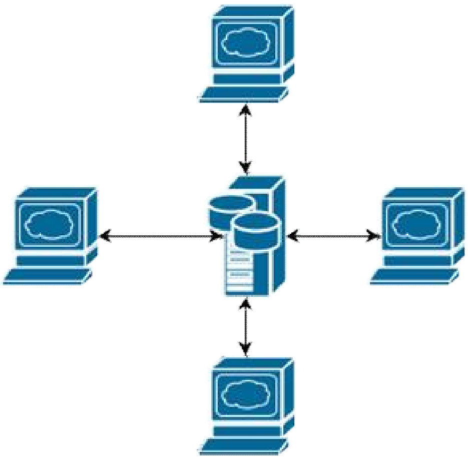


**Fig. 2.** Send-Receive

The issue of these two key aspects is that they might not be compatible regarding the content of the data transferred between them. For example, a browser can send data that consists of user's location, favorite search terms and so on. Meanwhile, the Big Data system receives all this information unsorted, so it's difficult to for it to understand whether this user is from "London" or from "orange". To avoid this "mess" created in Big Data solutions, all systems that send data should be standardized so that can send data in a logical array that, afterwards, it can be easily analyzed and displayed in a proper manner.

After Laney's "3V's" another two "V's" where added as key aspects of Big Data systems.

The fourth "V" is "**Value**" and is all about the quality of data that is stored and the further use of it. Large quantity of data is being stored from mobile phones call records to TCP/IP logs. The question is if all together can have any commercial value. There is no point in storing large amount of that if it can't be properly managed and the outcome can't offer insights for a good development.

"**Veracity**" is the fifth characteristic of Big Data and came from the idea that the possible consistency of data is good enough for Big Data. For example, if A is sending an email to B, B will have the exact content that A sent it, if else, the email service will not be reliable and people will not use it. In Big Data, if there is a loss regarding the data stored from one geo-location, is not an issue, because there a hundreds more that can cover that information.

Current technologies software technologies try to overcome the challenges that "V's" raises. One of these is Apache Hadoop, which is open source software that its main goal is to handle large amounts of data in a reasonable time. What Hadoop does is dividing data across a multiple systems infrastructure in order to be processed. Also, Hadoop creates a map of the content that is scattered so it can be easily found and accessed.

## 3. Useless to useful

The quantity of data that is being stored is not always one hundred percent useful. On the other hand the data stored is, in most cases, not already sorted and it represents piles of data that can consist of location information, web traffic log, financial data etc. To become useful, specialists must sort so that can be later analyzed and can be of any value (**Fig. 3.** Sort - Analyze). IT specialists say that they spend more time trying to "clean up" the data than they are analyzing it. Sorting and cleaning up data is a challenge that is hardly overcome. To do that, companies usually hires trained people that are able to manipulate the type of data that executive employees or higher management will further use. This process is time consuming and the costs are proportional to the volume of data.

A lot of companies try to sort and analyze the data that they stored with their own employees that have minimum skills or don't have them at all. The lack of skills in sorting Big Data will most certainly conclude into faulty results and/or truncated data that cannot serve its purpose.

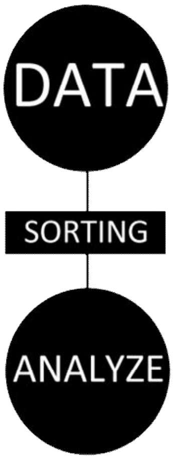


**Fig.3.** Sort - Analyze

A solution to eliminate the "need" for specialists is to implement software solutions which do not require special skills to understand how to put it at work. To be able to do that there is one more obstacle to overcome: quality of data. To achieve this, the architecture of the source that collects the data must be able to already sort it logical. For this to be accomplished, the data that is collected must be received in an understandable manner for the software that sorts it. These are the obstacles that will not be easy overcome, because there is no such thing as the idea of "controlled environment" when it comes to describing the World Wide Web. This problem can only be solved if the sets of data come, for example, from the financial department to the higher management of a company. In this case, data is already manipulated and is easy to understand and analyze to create projections.

A set of data, in Big Data environment, can be processed with OLAP tools. By doing so, there are connections between information that can be made. This set of tools have the purpose to rearrange the data provided into "cubes", which represents an IT architectural design that has the meaning of creating sets of data assembled into a logical and easy way to access it. By doing this, specialists achieved a higher speed, therefore a smaller waiting time, in processing large amount of data. The usefulness of the data that is being processed in an OLAP environment can still be questionable because all the data that was provided is being analyzed and sorted.

**4. Data privacy. Data security.**

This has many implications and it concerns individuals and companies as well. Individuals have the right, according to International Telecommunications Union, to control the information that may be disclosed regarding them. Information posted by users on their online profiles is likely to be used in creating a "users profile" so that can be further used by companies to develop their marketing strategies and to extend their services. Individual's privacy is still a delicate problem that can only be solved with drastic solutions. Allowing persons to choose whether they post or not information about them is a more secure way to achieve privacy, but will also cause software to "malfunction". For example in a social network, if a person is allowed to choose whether he/she wants to complete the fields regarding personal information and, in the same time, allow them to choose if the social network can store information about their IP address, location etc., this could be a possible threat to everyone else that is using the same social network.

For companies the privacy issue is more related to the sensitive data that they work with. Whether is financial data, clients list, perspective projects, all represents valuable data that may or may not be disclosed. Companies have multiple choices regarding where to store their information. They can either store it on cloud systems (**Fig. 4.** Cloud computing), "in-house" systems or a hybrid solution.

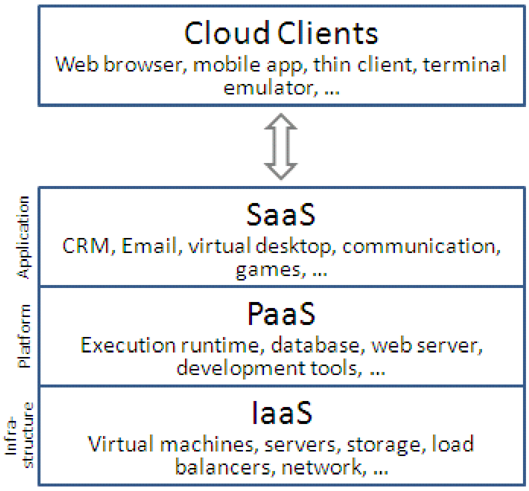

Cloud Clients
Web browser, mobile app, thin client, terminal emulator, ...
Application
SaaS
CRM, Email, virtual desktop, communication, games, ...
Platform
PaaS
Execution runtime, database, web server, development tools, ...
Infra-structure
IaaS
Virtual machines, servers, storage, load balancers, network, ...

**Fig. 4.** Cloud computing

By storing data on cloud systems is more convenient for companies in terms of cost. Also, a cloud system is not only characterized by storage space, but as well for the speed of processing requested operations. The data security still remains a contentious issue in this case.

To solve that, some companies choose to build their own infrastructure for storing and manipulate the data that they have. For smaller companies this can be a solution, but, in most cases, to implement such a system the costs are high. Also, to maintain this type of system it requires trained personnel and the more the company grows, the more it will be needed an add-on to the infrastructure. After all, this solution will prove redundant. The only gain of this solution is privacy.

Manipulating data, collecting it and store it in a proper manner that is in the advantage of the beneficiary and as well for the user that provides the data, will remain an important issue to be solved by IT security specialists. One solution for this matter, besides keeping all the data stored on an "in-house" Big Data system, is to encrypt it (**Fig. 5.** Encryption). By encrypting the data with a personal key it makes it unreadable for persons that don't have the clearance to see it. The downside of using encryption is that you have to use the same software that encrypted it to read it and analyze it, or, in worst case scenario, if you want to make it available for every software that is on the market the process implies more steps which take time. First step is to encrypt it using special encryption software, after that, each time the data is used for manipulation or analysis it must be decrypted and after work is finished, encrypt it again.

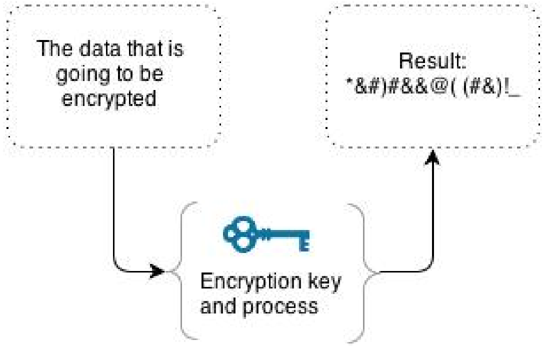

The data that is going to be encrypted
Result: *&#)#&&@( (#&)!_
Encryption key and process

**Fig. 5.** Encryption

To achieve performance from this process there has to be an OLAP system that is capable of doing encryption at the same time with reading data. By doing this the process will be much faster and data can be managed almost in real-time.

## 5. Sharing data

Sharing the information proves to be one of the most valuable characteristics of development. Information about almost anything can be found by simply doing a "Google search". Every person and company has at their disposal large amount of information that can use it to serve their purposes. Everything is available only if everyone shares it. Regarding persons, there is a difference between what is personal and what can be made public. The issue of what is personal and what is public mostly resides in the point of view of the services that they use.

Regarding companies, this is a challenge that most refuse to overcome. The reason that companies don't want to share their own "Big Data" warehouse is more related to competitiveness and sensitive data that they have. Otherwise, if this line is crossed, each company will have more data that they can analyze so that more accurate results can be obtained. With better results, comes better planning. If companies share the information that they hold about current market situation and/or possible clients and strategies to approach them, the grade of development will be drastically reduced and they will start focusing on how to hold to their current clients.

Sharing "Big data" at a level where each entity will show all the information that they hold is impossible. The framework of displaying data should be wider. A more transparent representation of current information that a company holds will be in the advantage of everyone. By doing this, the type of information and the way it is structured can help further development of software systems that can be standardized and can work with all types of data imported from various sources.

**6. Infrastructure faults**

Storing and analyzing large volumes of data that is crucial for a company to work requires a vast and complex hardware infrastructure. If more and complex data is stored, more hardware systems will be needed.

A hardware system can only be reliable over a certain period of time. Intensive use and, rarely, production faults will most certainly result in a system malfunction. Companies can't afford to lose data that they gathered in the past years, neither to lose their clients. For avoiding such catastrophic events they use a backup system that does the simple operation of storing all data. By doing this, companies obtain continuity, even if they are drawn back temporary. The challenge is to maintain the level of services that they provide when, for example, a server malfunction occurs right when a client is uploading files on it. To achieve continuity, hardware systems are backed by software solutions that respond in order to maintain fluency by redirecting traffic to another system. When a fault occurs, usually a user is not affected and he/she continues work without even noticing that something has happened. (**Fig. 6.** System failure)

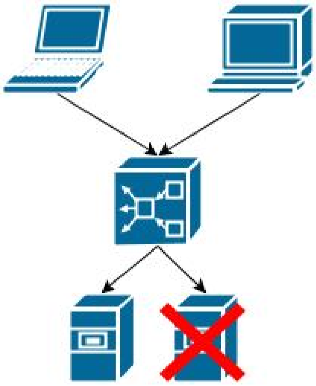


**Fig. 6.** System failure

The flow of data must not be interrupted in order to obtain accurate information. For example, Google is sending one search request to multiple servers, rather than sending it to only one. By doing this, the response time is shortened and also there is no inconsistency in the data that users sends – receives.

System failure affects the process of storing data and is making more difficult to work with. There can be created a permanent connection between the device, that is sending data, and the system that is receiving it, as a solution to this problem. By creating a loop, the "sender" will make sure that the "receiver" has no gaps regarding the data that should be stored. This loop should work as long as the system that is receiving data tells the system that sends it to stop because the data that is stored is identical to the one sent. So, is a simple comparison process that can prevent loosing data. This process can also slow down the whole process. To avoid this from happening, for any content that is transmitted, the sender must generate a "key". This key is then transferred to the receiver to compare it with the key that it generated regarding the data that was received. If both keys are identical than the “send-receive” process was successfully completed. For better understanding, this solution is similar with the MD5 Hash that is generated over a compressed content. But, in this case, the keys are compared automatically.

Loosing data is not always a hardware problem. Software can as well malfunction and cause irreparable and more dangerous data loss. If one hard drive fails, there is usually another one to back it up, so there is no harm done to data, but when software fails due to programming “bug” or a flaw in the design, data is lost forever. To overcome this problem, programmers developed series of tools that will reduce the impact of a software failure. A simple example is Microsoft Word, which saves from time to time the work that a user is doing in order to prevent the loss of it in case of hardware or software failure. This is the basic idea of preventing complete data loss.

## 7. Technologies for Big Data

Once realized the amplitude of information that is crossing the internet, specialists started to question how to handle this amount of data. To obtain good insights and mine this information they had to develop tools capable of creating the expected results. A common implementation that handles Big Data is **MapReduce** (**Fig. 7.** MapReduce).

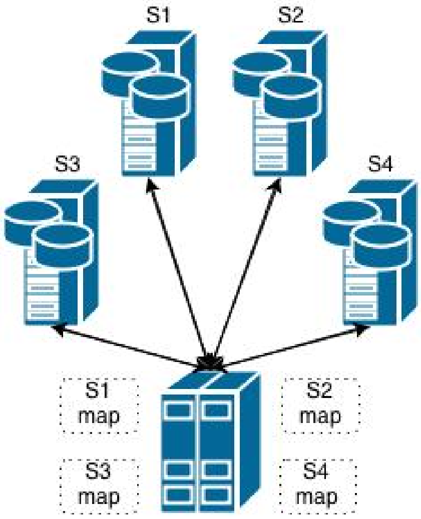


**Fig. 7.** MapReduce

This is more of a technique that programmers use when they are confronted with large amount of data.

MapReduce consists of two things: mapping and reducing. By mapping a certain dataset is restructured into a different set of values. Reducing is a process that takes several “mapped” outputs and forms a smaller set of tuples.

The most popular technology that is able to mine and sort data is **Hadoop**. Being open source software, Hadoop is the most implemented solution for handling Big Data. It has enough flexibility to work with multiple data sources, or even assemble multiple systems to be able to do large scale processing. Hadoop is used by large companies such as Facebook and Google. Hadoop also use HDFS (**Hadoop Distributed File System**) that has the role to split data into smaller blocks and distribute it throughout the cluster. In order to assist Hadoop, Facebook developed a software system called **Hive**. Hive is basically a “SQL-like” bridge that connects with Hadoop in order to allow conventional applications to run queries. The advantage is that is simple to use and understand. It combines the simplicity and utility of a standard relational database with the complexity of a Hadoop system. The downside of using Hive is latency. Because it is built on Hadoop, Hive can have high latencies on executed queries, compared to IBM’s DB2. Large companies use Hadoop as a starting point in order to deploy other solutions.

DB2 is a fast and solid data manipulating system. It has feature that reduces the cost of administration by doing an automated process that increases storage efficiency and improves performance.

Oracle, on the other hand, comes with a complete system solution for companies (**Fig. 8.** Oracle solution).

It starts from the basic ideas of Big Data sources, which can be traditional data generated by ERP systems, sensor data and social data, defined by the feedback that the company receives from customers and other sources. The solution given by Oracle is to create a system from top to bottom, based on NoSQL. A NoSQL database is capable of handling various types of data that traditional relational databases are unable to handle and lose data consistency. NoSQL derives from “Not only SQL”, which means that it allows regular SQL queries to be executed. Oracle’s solution is presented in 4 steps: ACQUIRE > ORGANIZE > ANALYZE > DECIDE [2]. All the steps combine different solutions like HDFS, NoSQL, Oracle Big Data connectors and Cloudera CDH.

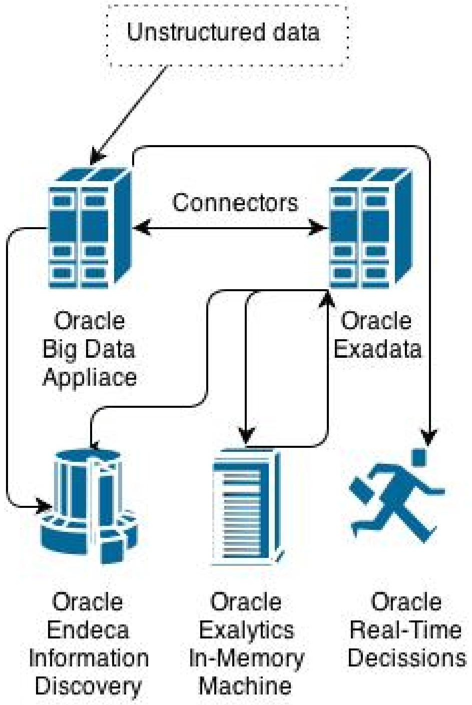

Unstructured data
Connectors
Oracle Big Data Appliace
Oracle Exadata
Oracle Endeca Information Discovery
Oracle Exalytics In-Memory Machine
Oracle Real-Time Decissions

**Fig. 8.** Oracle solution

CDH or *Cloudera’s Distribution Including Apache Hadoop*, offers batch processing (uses MapReduce and HDFS), interactive SQL (allows users to execute SQL queries) and interactive search [3]. All these key features that Cloudera offers are solutions that allow users to navigate through clusters and retrieve data that they need.

SAS offers multiple solutions to overcome Big Data mining and analysis. It also tries to cover all that is necessary for a company to create value from stored data. One solution is SAS DataFlux which is a data management solution that can provide users the right tools for integrating data, mastering data and data quality. It also allows access and use of data across company and also provides a unified set of policies in order to maintain data quality. SAS also provides high-performance analytics solution that is providing the company good insights from analyzing data in a structured, easy to read, report. This is basically one of the main goals when working with Big Data, to get best insights from quality data. Also SAS provides analytics solution that is based on a drag-and-drop system which can provide easy to understand and customized reports and charts.

SAS is more oriented in providing software solutions to help companies benefit from data that they have stored.

The problem of handling Big Data doesn’t always resides in analysis and mining software solutions. It has a great impact over hardware systems and their capability of processing. The two of them, software and hardware solutions, create a complete Big Data system that can be viable and will produce the expected outcome. In order to handle large and complex data sets, a solution must be divided according to job process. For example, a basic data storage and mining solution should have a system that will store brief information about the data that exists on clusters. By doing this, the data mining process is drastically reduced because the role of this system is to orientate the user where to look and what to look for. Also, this system should be able to evenly “spread” the data among clusters. By performing this, clusters can be monitored so there will be no overloading and, therefore, slowing down the outcome.

Another solution to treat Big Data is to design a system that is capable of making differences between various types of data. Searching through one type of data is easier than to search through different types of information. As an example, searching through a full text file is easier to find answers than searching through a text file that has images as well that can provide answers to questions. This is a “divide et impera” technique. By doing this, one cluster with special software design will handle video file, another will handle 3-D models, another plain text files etc. (**Fig. 9.** Divide et impera)

This system will allow data fragmentation which will be faster to process. The software solution that can handle data fragmentation can be achieved through MapReduce.

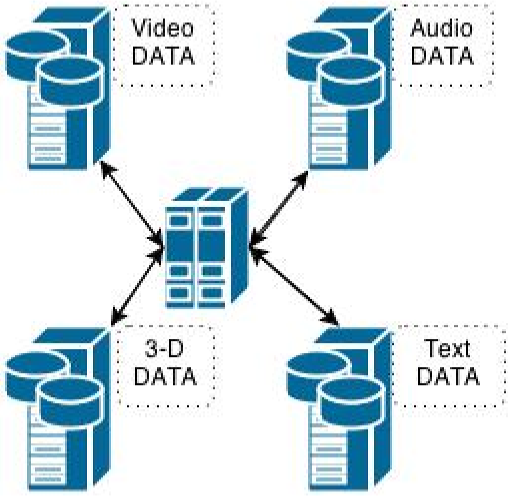


**Fig. 9.** Divide et impera

Besides the basic function of dividing data, MapReduce can be configured to recognize the type of data that is mapping. Special designed software for each cluster will have the job to sort data by various key elements which are set by the user. Clusters can also analyze in a basic manner the information stored so that the “mainframe” will act as a control center for the system. This will help analysts achieve fast and accurate results and will also allow real-time updates about ongoing information.

Regarding Big Data “Volume” is a characteristic, as well as a challenge. Trying to deal in a fast manner with large data sets will be difficult for a system. Trying to lower the volume might help solve this problem. For a Big Data solution is faster to process 1GB of data instead of 1TB. When it comes to Big Data, valuable information should be the subject of analysis. Spending time on “cleaning” information can lead to a result that is no longer valid or is too late to be applied. To speed up this process, an automated data “clean-up” process should be implemented. In most cases, not all the information collected is needed for the analysis. To do this, a data filtering solution can be created. For example, a mainframe can decide which data is needed and which is not. The one that is needed will be transferred to a main cluster that provides the information needed for immediate analysis. The rest of the information will be transferred to a secondary cluster that will hold only data that can be later analyzed or even deleted.

So, “looping”, “divide et impera” and “filtering”, these can all form a Big Data solution that can be helpful. This solution covers the data loss that can occur from a hardware malfunction or a software error. It will also manage a data distribution among clusters by data type and previously set aspects that will ensure better and faster analysis of the information stored. Least but not last, will provide an automatic filtering process that will facilitate the evaluation of valuable data. To achieve this, the solution must have a coordination center, a processing system and special designed software for each cluster system.

The only challenge that needs a new approach is the “Velocity” issue. In order to obtain higher processing and transferring speed, the volume of data that is manipulated must be reduced. This cannot be possible without slowing the analysis process and, therefore, “Volume” stays untouched.

## 8. Conclusions

Building a viable solution for large and complex data is a challenge that companies in this field are continuously learning and implementing new ways to handle it. One of the biggest problems regarding Big Data is the infrastructure’s high costs. Hardware equipment is very expensive for most of the companies, even if Cloud solutions are available. Each Big data system requires massive processing power and stable and complex network configurations that are made by specialists. Besides hardware infrastructure, software solutions tend to have high costs if the beneficiary doesn't opt for open source software. Even if they chose open source, to configure there is still needed specialists with the required skills to do that. The downside of open source is that maintenance and support is not provided as is the case of paid software. So, all that is necessary to maintain a Big Data solution working correctly needs, in most cases, an outside maintenance team.

Software solutions are limited by hardware capabilities. Hardware can only be as fast as current technologies can offer. A software solution just sends the tasks in order to be processed. The software architecture tries to compensate the lack of processing speed by sorting and ordering the requests, so that processes can be optimized in order to achieve best performance.

To sort through data, so that valuable information will be extracted for further use, requires human analysis skills. A computer program can only do what is programmed to do, it cannot see grey areas and cannot learn or adapt to new types of information unless is programmed to handle it. Therefore, human capabilities are used to sort data with a set of tools which speed up the process. All this will only increase the time that results will be displayed and so, the analysis of the results, in order to evaluate current position or forecast, will decrease the beneficiary's time for taking measures or plan properly.

## References

[1] Global data center traffic – Cisco Forecast Overview - http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.html

[2] Oracle Big Data strategy guide, http://www.oracle.com/us/technologies/big-data/big-data-strategy-guide-1536569.pdf

[3] Cloudera's 100% Open Source Distribution of Hadoop, http://www.cloudera.com/content/cloudera/en/products/cdh.html

**Alexandru Adrian TOLE** (born 1986 in Romania) graduated from the Faculty of Domestic and International Commercial and Financial Banking Relations of the Romanian – American University in 2009. He also graduated the Scientific Master in Finance, Banking, Insurances. He works at the Ministry for Information Society. He is pursuing a Ph. D. in the area of Business Intelligence systems.

# Personalized e-learning software systems. Extending the solution to assist visually impaired users

Diana BUTUCEA
University of Economic Studies, Bucharest, Romania
dianabutucea@gmail.com

*Discussing the subject of e-learning in the context of the latest updates of technology nowadays represents quite a challenge when the topic must be addressed to special classes of computer users. The paper will present a theoretical framework for visually impaired persons, followed by a technical implementation of the concept in relation with the e-learning context. The solution proposes an analytical approach over the computer aided learning mechanism, defining the concept of personalized learning and providing an example of implementation for a software system that, subsequently, offers support and assistance for visually impaired computer users. The technique implements a specially designed software library, integrates it in an e-learning software system and combines the power of a web-based solution with the support and guidance offered by a text-to-speech integration, resulting into a reliable e-learning software implementation. The paper also focuses on the theoretical aspects of the problem, and will present its conclusions at the end.*

***Keywords:*** *e-learning, personalized learning, text-to-speech, JAWS, integration*

## 1 Introduction

The interfaces and locations where computerization entered the learning process have changed dramatically since the advent of the web. In the last decade, classic research has been replaced by web based research, especially because libraries have changed from the printed format and their physical spaces for storage, into electronic learning resources and virtual laboratories. At the same time, the increase of computerization field has major implications on how students perceived their need to access the information (location, rating content, use or creation of information). In this context, the online field, search engines updates and the exponential growth of the web 2.0 and its technologies proposed, allowed this irreversible change of the current context information. [9]

Undoubtedly, the most dramatic rise in training and development over the past 20 years has been the increased use of technology, explained by the convergence of technologies used to deliver content. [1]

Adopting advanced forms of disseminating information and new information technology solutions proposed, was a main stay in the extensive development of e-learning. Due to the popularization and wide applicability of internet technologies, it was possible to develop virtual learning environments which broke temporal or spatial barriers of the research. From the point of view of information technology development process, each forward step made by internet is an impact prerequisite and a positive effect for distance education model [11], bringing with each evolving value to the field.

Since e-learning systems are becoming more available, many instructors have begun to use these systems in their teaching process. Their desire to try to use such systems, however, doesn't guarantee that they will continue to use them on long term. Previous studies are showing that continued use of the information system is determined by its perceived usefulness. The studies made by Roca [10] and Chiu [3] indicate a major factor in the continued use of the system, which represents perceived usefulness, made from the actual experience of use. Therefore, the design of an information system should consider all the useful features for its users. However, these features cannot be established or updated before they use the system and provide comments upon it. [13]

Another considered aspect is the personalized learning. This occurs when e-learning systems make deliberate efforts to adapt to the educational experiences that fit the needs, goals, skills and interests of its students. Researchers have recently begun to investigate various techniques to help teachers improve e-learning systems. [7]

Introducing these as a starting point, I have developed an e-learning model that offers a secure working-base for e-learning, considering the technology and functionality included.

What this article aims, is to present practical solution in order to assist visually impaired users, taking into consideration financial resources and current technology standards.

## 2. Solutions for visually impaired persons

As the use of virtual learning environments and other computer-based educational resources is increasing every day, the concern about the inclusion of any course in these systems is crucial. If educational instruments are not developed properly, the use of such systems can become an additional factor in the exclusion of students with disabilities from educational process. However, it is important to consider how the use of the computer has increased to all more opportunities in order to make their lives easier. The development of assistive technologies has provided great opportunities for people with disabilities to transform their way of life in a productive, efficient and result oriented way.

Although in recent years many advances have been made regarding assistive technologies, it was also found a number of shortcomings, somehow inherent in the interaction with common technologies. For example, graphical resources of the images can be automatically translated into text, in order to be read by the screen reader. The presence of elements of interaction that can be achieved only through pointing devices is also a barrier for people with visual impairment.

Promoting inclusion in regular educational fields showed a genuine concern for both teachers and government officials. Therefore, promoting social inclusion, regardless of disability, is an important issue to be addressed in the context of e-learning environments. Making e-learning systems accessible to all was a challenge. Indeed, many e-learning systems fail to adhere to web accessibility guidelines. [4] The challenge of designing these systems more affordable is more serious if we take into consideration the synchronized interactive technologies and multimedia resources.

Many reports of the researchers are showing the efforts to promote inclusion in e-learning field. One example is the University Notebook Model. [8] One of the aims of the project is to develop specific applications, which will support the production of wide e-learning systems. The Center of Studies for the Blind and Partially Blind Students, involved in this project, has devoted special attention to questions about the blind or visually impaired persons. The Center investigates, for example, issues to be considered when preparing a course, issues of the documents used in higher education, the use of multimedia, the availability of documents and others.

Other resources that can improve e-learning applications from the point of view of blind people are adopting and integrating a suitable screen reader and voice recognition systems. Wald [14] analyzed the way automatic speech recognition can support universal access to communication and learning, by making text synchronized with speaking. This implies support for the blind persons, for those with visual impairments or dyslexia in order to read and search for information.

Traditional desktop screen readers, such as JAWS1 and Virtual Vision2, are already part of life for the blind people. Many efforts have been made and are made to implement them. For example, Chen [2] presented the Audio Browser software, which is designed with a standard Personal Digital Assistant (PDA), having a number of features that are making the interface more accessible and easier to use. It is used a touch screen, buttons to enter data and non-speech audio. The response for the user is communicated through "speech", thus allowing easy navigation through stored information and also for accessing the system controls.

IWB (Interactive Whiteboards) became a common resource in the study classes around the world. Slay [12] shows the reports of the benefits of this technology of some teachers, such as effectiveness, flexibility, versatility, opportunity to access multimedia content, support multiple needs in a single lesson, and the ability to handle the teaching that class requires it, allowing teachers to maintain control over the training group, which would be more difficult to achieve with a computer. [4]

In the process of testing visually impaired persons, the study made by Hochheiser [5] presents in detail a probabilistic strategy approach of how these people use web interfaces. The sample studied by the author shows that, by defining a tree structure of web pages, visually impaired persons will have almost a normal experience in using software testing platform.

In Romania, computer-assisted learning for people with disabilities is at an early stage of development. There are initiatives in this regard, especially since both the Romanian legislation and the European Union encourages the development and promotes assisted software, using e-learning. [6]

## 3. E-learning application

For the ease of development, in order to take as many out-of-the-box features that are compatible with a wide range of devices and browsers, as well as better accessibility, the Twitter Bootstrap framework has been used in developing the graphic interface.

The reason for this choice is given by the fact that the Twitter Bootstrap provides a wide variety of components and tools for the development of a consistent interface, which operates on both mobile devices and desktop units, with higher resolution. Therefore, the arrangement of elements on the page can dynamically change, depending on the resolution using the method behind the media queries and the responsive design.

The user`s interface can be separated into modules accessible to unauthenticated users, that are considered by the system as "guests", modules available to authenticated users as students and modules for teachers or administrators. The Global Diagram is shown in Figure 1.

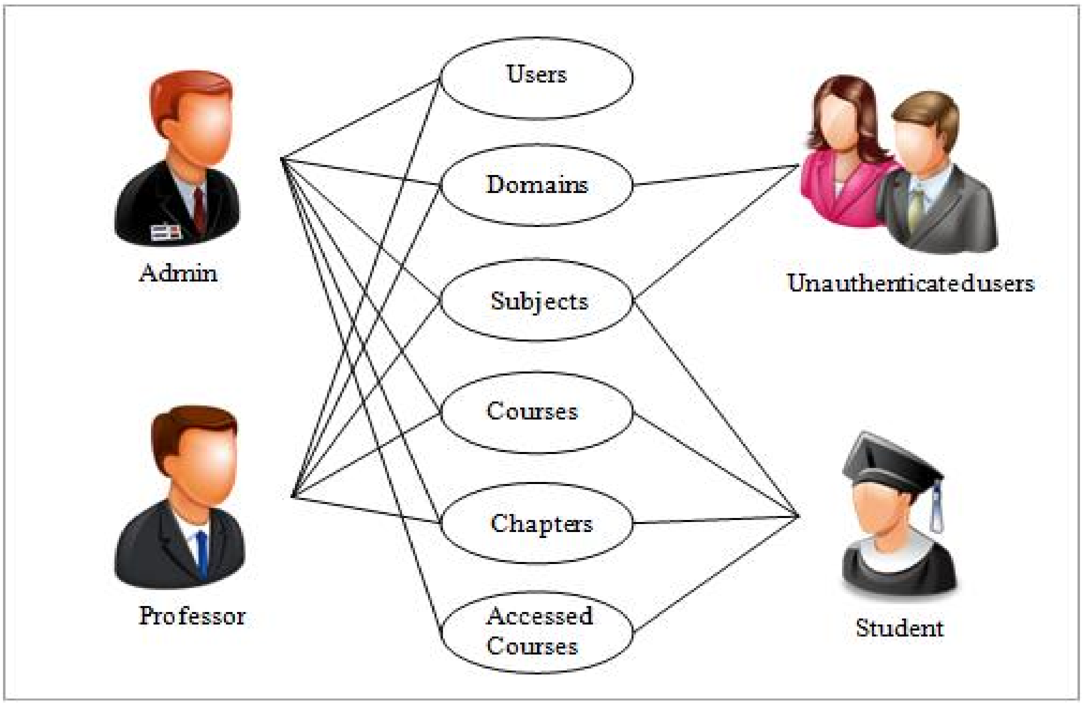


**Fig. 1.** Global Diagram

The Admin user (administrator) in the context of the application is a teacher with a set of extended rights, so that an Admin can only be changed by a user with equal rights and has access (such as teachers) to all the functionality of the application.

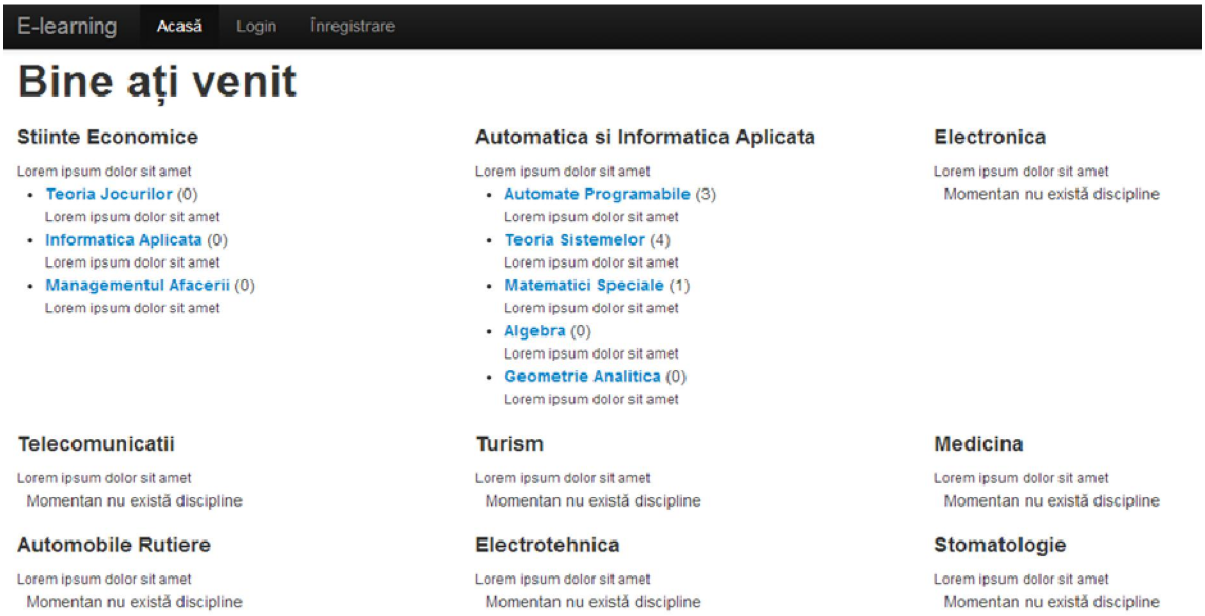
At the moment of the first access by a new user, he will have access to the initial screen that contains a low number of information about the system and the possibility of creating an access account, these being the pages for unauthenticated users.

The Main Page is shown in Figure 2. It exposes information about the number of courses available under each subject belonging to all domains registered, but without the possibility of being accessed. The chose of this interface has the main argument the wish to attract as users only those students who are strictly interested in the available material.

In the module accessible to the students, the user which successfully authenticates is greeted by a page as shown in Figure 3, with access to its own domain of study (chosen at registration) and to the latest accessed courses. The student can view all the information exposed by the teacher (in the domain) and can edit his own profile - changing the password is the only element that can be edited for the moment as a measure taken for avoiding identity theft situations.

Another functionality implemented is draining the list of accessed courses. This option was introduced in order to allow the user to decide on the amount of information displayed, depending on the importance given.

The page with the greatest importance, to which have access both advanced users who have rights, but especially students (to whom it is particularly intended) is the Course detailed Page (Figure 4).



Bine ați venit

Stiinte Economice

Lorem ipsum dolor sit amet

- Teoria Jocurilor (0)
  Lorem ipsum dolor sit amet
- Informatica Aplicata (0)
  Lorem ipsum dolor sit amet
- Managementul Afacerii (0)
  Lorem ipsum dolor sit amet

Automatica si Informatica Aplicata

Lorem ipsum dolor sit amet

- Automate Programabile (3)
  Lorem ipsum dolor sit amet
- Teoria Sistemelor (4)
  Lorem ipsum dolor sit amet
- Matematici Speciale (1)
  Lorem ipsum dolor sit amet
- Algebra (0)
  Lorem ipsum dolor sit amet
- Geometrie Analitica (0)
  Lorem ipsum dolor sit amet

Electronica

Lorem ipsum dolor sit amet
Momentan nu există discipline

Telecomunicatii

Lorem ipsum dolor sit amet
Momentan nu există discipline

Turism

Lorem ipsum dolor sit amet
Momentan nu există discipline

Medicina

Lorem ipsum dolor sit amet
Momentan nu există discipline

Automobile Rutiere

Lorem ipsum dolor sit amet
Momentan nu există discipline

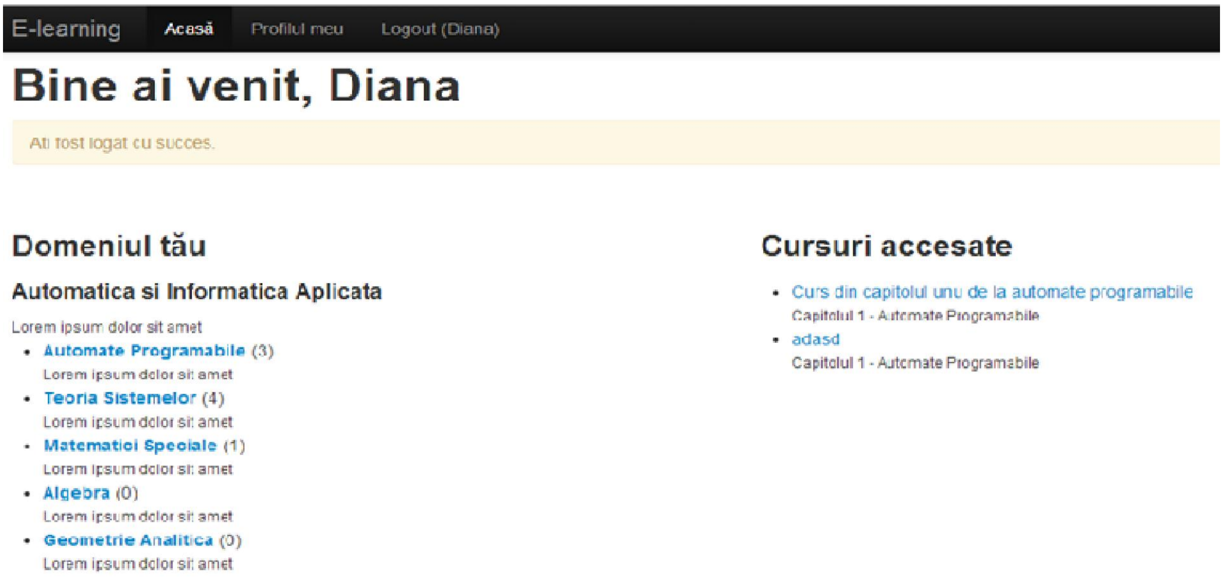
Electrotehnica

Lorem ipsum dolor sit amet
Momentan nu există discipline

Stomatologie

Lorem ipsum dolor sit amet
Momentan nu există discipline

**Fig. 2.** Main Page - unauthenticated users



Bine ai venit, Diana

Ati fost logat cu succes.

Domeniul tău

Automatica si Informatica Aplicata

Lorem ipsum dolor sit amet

- Automate Programabile (3)
  Lorem ipsum dolor sit amet
- Teoria Sistemelor (4)
  Lorem ipsum dolor sit amet
- Matematici Speciale (1)
  Lorem ipsum dolor sit amet
- Algebra (0)
  Lorem ipsum dolor sit amet
- Geometrie Analitica (0)
  Lorem ipsum dolor sit amet

Cursuri accesate

- Curs din capitolul unu de la automate programabile
  Capitolul 1 - Automate Programabile
- adasd
  Capitolul 1 - Automate Programabile

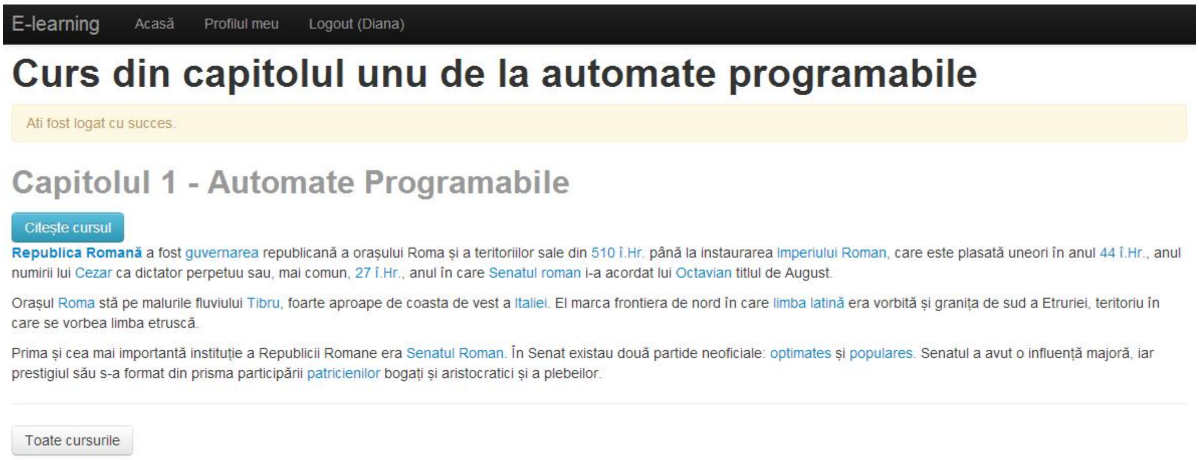
**Fig. 3.** Main Page – Student module

## 3. The proposed solution

Given the above, it was started an investigation process of the situation, by contacting the staff of Centre of functional rehabilitation for visually impaired in Braşov. This center aims to specific social and professional inclusion of visually impaired adults at risk of social marginalization and increase their quality of life through the provision of comprehensive social services. In matters of informatics aspects, beneficiaries of the center are guided by a specialist to use the Braille display, keyboard and the working stations on which is installed a speech synthesis application.

In terms of technology, the issue raised by the person responsible for training those who want to learn how to use the computer, was related to the application of speech synthesis. The one used in the rehabilitation center is JAWS1. From personal observations and explanations received, it could easily be concluded that the voice provided by the application and the fact that it does not include Romanian language, make it extremely difficult to understand the information which is "read". Newer versions of JAWS program contain substantial improvements, but the need to purchase a separate license for each working station is a considerable financial effort, especially in the socio-economic conditions of our country. Therefore, at least for the moment, upgrading and updating to the new version of JAWS software is not possible.

In order to assist people with visual disabilities, we included in the course presentation page (Figure 4) the functionality accessed through the "Read the course" button.



Curs din capitolul unu de la automate programabile

Ati fost logat cu succes.

Capitolul 1 - Automate Programabile

**Republica Romană** a fost guvernarea republicană a orașului Roma și a teritoriilor sale din 510 î.Hr. până la instaurarea Imperiului Roman, care este plasată uneori în anul 44 î.Hr., anul numirii lui Cezar ca dictator perpetuu sau, mai comun, 27 î.Hr., anul în care Senatul roman i-a acordat lui Octavian titlul de August.

Orașul Roma stă pe malurile fluviului Tibru, foarte aproape de coasta de vest a Italiei. El marca frontiera de nord în care limba latină era vorbită și granița de sud a Etruriei, teritoriu în care se vorbea limba etruscă.

Prima și cea mai importantă instituție a Republicii Romane era Senatul Roman. În Senat existau două partide neoficiale: optimates și populares. Senatul a avut o influență majoră, iar prestigiul său s-a format din prisma participării patricienilor bogați și aristocratici și a plebeilor.

**Fig. 4.** Course detailed Page

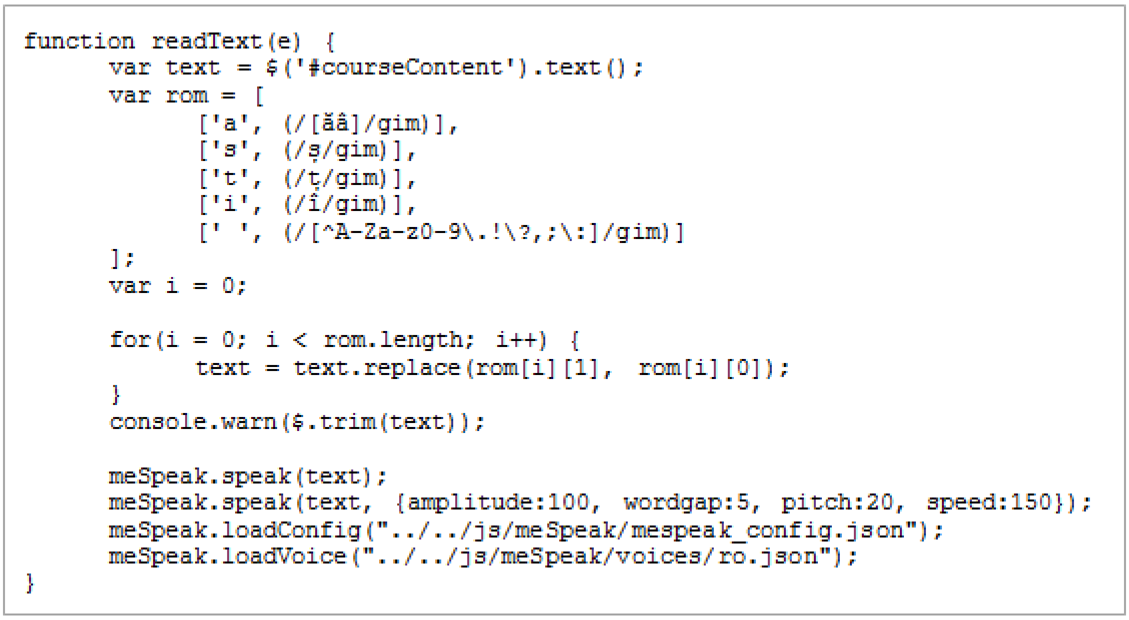
It invokes a JavaScript library called *meSpeak.js*. The application is a conversion of the popular open source Text-to-Speech algorithm, called eSpeak. The conversion is performed automatically by the *emscripten* tool that converts the C++ source code into JavaScript code compatible with modern versions of web browsers. The code that invokes this functionality is based on completion of a string given as argument (Figure 5).

Because meSpeak does not correctly interpret the accented characters (it omits them completely), it has been implemented a transliteration of special characters before being sent to the script in order to be read. Also other special characters (for example: . ! ? , ; :) cause interruption of reading, which is why I chose their elimination from the text before reading, using the function *gim* (*Global* - replacement occurs throughout the whole text, not only at their first appearance, *case Insensitive*, *Multiline*).

In the implementation process, a particular attention was paid to the voice amplitude parameters adjustment (amplitude), the pause between words measured in units of millisecond ($10^{-3}$ sec word gap), the pitch of the voice, the speed of speech (the number of words read per minute). Those have been adjusted by trial and error, in order to obtain a variant of speech as near as possible of the normal rhythm.

The use of this voice synthesis system has proven to be a sufficient and satisfactory choice to implement targeted functionality, allowing easy synthesis of Romanian language phrases with high level of voice quality, easily understood by the human operator.

The priority of development a new system of speech was considered to be significantly lower compared with the importance of such a system within the platform developed to support people with visual impairments



```
function readText(e) {
        var text = $('#courseContent').text();
        var rom = [
                ['a', (/[ăâ]/gim)],
                ['s', (/ș/gim)],
                ['t', (/ț/gim)],
                ['i', (/î/gim)],
                [' ', (/[^A-Za-z0-9\.!\?,;\:]/gim)]
        ];
        var i = 0;

        for(i = 0; i < rom.length; i++) {
                text = text.replace(rom[i][1], rom[i][0]);
        }
        console.warn($.trim(text));

        meSpeak.speak(text);
        meSpeak.speak(text, {amplitude:100, wordgap:5, pitch:20, speed:150});
        meSpeak.loadConfig("../../js/meSpeak/mespeak_config.json");
        meSpeak.loadVoice("../../js/meSpeak/voices/ro.json");
}
```

**Fig. 5.** *emscripten* functionality

## 4. Conclusions

Computer-assisted learning processes are a step forward to the global education model. Their effectiveness is determined by several factors, including that the final user can specify affinity towards this type of system, how the model implementation was made in educational institutions (directly dependent on the degree of national computerization of a country), funding volumes obtained for such implementations etc.

Therefore, the article presents at a detailed level, the current state of knowledge. The learning model proposed by the technology and functionality included, offers a secure working-base for e-learning. The solution is stable and easy to maintain in terms of resources and presents extensive capabilities, in comparison with other already existing in the consumer market.

An important element of personal contribution is the model which meets people with visual disabilities, solution that proves to be extremely useful to them. Such an application has proven to be necessary to facilitate the learning to these people and to make this process more enjoyable and more efficient. It is noted, thus, the developing of an integrated software system specifically dedicated to learning and continuous training of any type of user, the only condition being, as in any field, that the person is willing to be trained.

**Acknowledgments**
This work was cofinanced from the European Social Fund through Sectorial Operational Programme Human Resources Development 2007-2013, project number POSDRU/107/1.5/S/77213 „Ph.D. for a career in interdisciplinary economic research at the European standards”.

**References**

[1] Brown, K.G., „A field study of employee e-lea*rning activity and outcomes”, Human Resource Development Quarterly*, 16, 2005, 465-480

[2] Chen, X., Tremaine, M., Lutz, R., Chung, J.W., Lacsina, P., „Audiobrowser: A mobile browsable information access for the visually impaired”, *Universal Access in the Information Society*, 5, 2006, 4-22

[3] Chiu, C.M., Hsu, M.H., Sun, S.Y., Lin, T.C., Sun, P.C., „Usability, quality, value and elearning continuance decisions”, *Computers & Education*, 45, 2005, 399-416

[4] Freire, A.P., Linhalis, F., Bianchini, S.L., Fortes, R., Pimentel, M., „Revealing the whiteboard to blind students: An inclusive approach to provide mediation in synchronous e-learning activities”, *Computers & Education*, 54, 2010, 866-876

[5] Hochheiser, H., Lazar, J., „Revisiting breadth vs. depth in menu structures for blind users of screen readers”, *Interacting with Computers - Elsevier*, 22, 2010, 389-398

[6] Isaila, N., Nicolau, I., „Promoting computer assisted learning for persons with disabilities”, *Procedia Social and Behavioral Sciences*, 2, 2010, 4497-4501

[7] Klasnja-Milicevic, A., Vesin, B., Ivanovic, M., Budimac Z., „E-Learning personalization based on hybrid recommendation strategy and learning style identification”, *Computers & Education*, 56, 2011, 885-899

[8] Klaus, J., „Living, teaching and learning at any time and at any place. E-learning opportunities and barriers for visually impaired students.”, In *Proceedings of the 9th international conference on computers helping people with special needs*, 2004, 151-156

[9] McClure, R., „Writing Research Writing: The Semantic Web and the Future of the Research Project”, *Computers and Composition*, 28, 2011, 315–326

[10] Roca, J.C., Chiu, C.M., Martinez, F.J., „Understanding e-learning continuance intention: an extension of the technology acceptance model”, *International Journal of Human-Computer Studies*, 64, 2006, 683-696

[11] Shi, Y., Wang, M., Qiao, Z., Mao, L., „Effect of Semantic Web technologies on Distance Education”, *Procedia Engineering*, 15, 2011, 4295-4299

[12] Slay, H., Siebörger, I., Hodgkinson-Williams, C., „Interactive whiteboards: Real beauty or just lipstick?”, *Computers & Education*, 51, 2008, 1321-1341

[13] Sun, P.C., Cheng, H.K., Finger, G., „Critical functionalities of a successful e-learning system – An analysis from instructors' cognitive structure toward system usage”, *Decision Support Systems*, 48, 2009, 293-302

[14] Wald, M., Bain, K., „Universal access to communication and learning: The role of automatic speech recognition”, *Universal Access in the Information Society*, 6, 2008, 435-447

**Diana BUTUCEA** graduated in 2008 from the Economic International Relations Department of the Academy of Economic Studies in Bucharest, in 2009 from the Faculty of Automatics and Applied Informatics of the Transylvania University of Brasov and in 2010 from the Economic Informatics masters at the Academy of Economic Studies in Bucharest. Her parallel interests, in economy and software engineering, are now merging into her studies and researches since she is PhD student at the Academy of Economic Studies, Bucharest, studying integrated software systems, web technologies and e-learning.

# Comments on "Problem Decomposition Method to Compute an Optimal Cover for a Set of Functional Dependencies"

Xiaoning PENG, Zhijun XIAO
Department of Computer Engineering, Huaihua University, China
pxnxzj@gmail.com

Maier, in his seminal paper [1], presents that the problem of finding an optimal cover is NP-complete. Cotelea's paper [2] proposes a problem decomposition method for finding an optimal cover. The basic idea of the method is: (1) a big and intractable problem (optimal cover problem) is broken down into some smaller problems; (2) particular solutions of these smaller problems are combined to construct the initial problem solution. More specifically, the main idea of the method is: (1) a relational schema $R$ with a set $F$ of functional dependencies is partitioned in some subschema; (2) each subschema will be found determinants with the fewest attributes (including repeated); (3) substituting the groups of attributes in $F$ that are determinants in some subschema with the shortest determinants, Cotelea states that this happens only for equivalence classes of functional dependencies containing the determinants in the left or right sides as subsets.

The idea embodied in this method is classic in the algorithm theory area, and Cotelea proposes many valuable results. However, are these sufficient for finding all optimal covers? (Seemingly unlikely)

Please see a counterexample as the following.

**Example 1.** Given $G=\{AD\to E, A\to BC, B\to A, C\to A\}$ that is a minimum and reduced set of functional dependencies, please find an optimal cover for $G$. For example, $H=\{AD\to E, A\to B, B\to C, C\to A\}$ is an optimal cover for $G$.

The contribution graph of $G$ is shown in Figure 1.

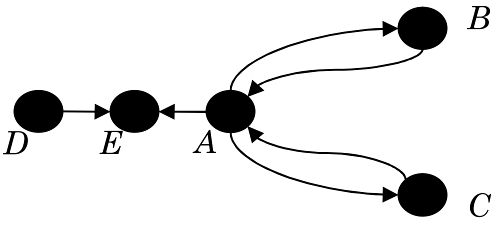


**Fig. 1.** A contribution graph for $G$

$G$ is divided into two equivalence classes $G=G_1\cup G_2$, where $G_1=\{AD\to E\}$, and $G_2=\{A\to BC, B\to A, C\to A\}$. Note that $G_1$ and $G_2$ satisfy the strict partial order. The equivalence class $G_1$ precedes the equivalence class $G_2$.

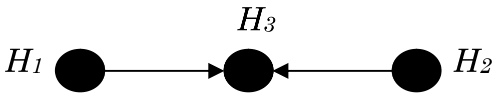


**Fig. 2.** Condensed graph of the graph from Figure 1

Figure 2 shows that $H_1$, $H_2$ and $H_3$ satisfy the strict partial order. The set of vertices of the graph in Figure 1 can be divided into three equivalence classes of attributes $S_1=\{D\}$, $S_2=\{A, B, C\}$ and $S_3=\{E\}$, and are reduced to the following sequence of non redundant equivalent classes of attributes $T_1=\{D\}$, $T_2=\{A, B, C\}$.

The set $G$ of functional dependencies, below, is projected on the sets of attributes $T_1$ and $T_2$, resulting in the following sets of functional dependencies:

$\pi_{T1}(G)=\Phi$,

$\pi_{T2}(G)=\{A\to BC, B\to A, C\to A\}$.

Thus, for the non redundant classes of attributes, there were obtained the following sets of determinants $\{D\}$, $\{A, B, C\}$, respectively.

Now the groups of attributes that are determinants and part of dependencies in $G$ are substituted by those with the smallest length. Substitutions occur in the equivalence classes of dependencies which precede corresponding class that has generated the determinant. Then, we can obtain an optimal cover $F=\{AD\rightarrow E, A\rightarrow BC, B\rightarrow A, C\rightarrow A\}$ for $G$. However, $H$ rather than $F$ is the optimal cover for $G$.

Example 1 tells us that the idea embodied in the Cotelea's paper is not sufficient for finding all optimal covers, in spite of Cotelea presents a correct example in his paper. The optimal cover problem is NP-complete, and it seems hard to find a deterministic algorithm for finding an optimal cover.

## References

[1] D. Maier, "Minimum Covers in the Relational Database Model", Journal of the ACM, 1980, V.27, N 4, pp.664-674.

[2] Vitalie Cotelea, "Problem Decomposition Method to Compute an Optimal Cover for a Set of Functional Dependencies", Database Systems Journal, 2011, V.2, N 4, pp.17-30.