CEPIS UPGRADE is the European Journal for the Informatics Professional, published bi-monthly at <http://cepis.org/upgrade>

Publisher
CEPIS UPGRADE is published by CEPIS (Council of European Professional Informatics Societies, <http://www.cepis.org/>), in cooperation with the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática,* <http://www.ati.es/>) and its journal Novática

CEPIS UPGRADE monographs are published jointly with Novática, that publishes them in Spanish (full version printed; summary, abstracts and some articles online)

CEPIS UPGRADE was created in October 2000 by CEPIS and was first published by Novática and INFORMATIK/INFORMATIQUE, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies)

CEPIS UPGRADE is the anchor point for UPENET (UPGRADE European NETwork), the network of CEPIS member societies' publications, that currently includes the following ones:
- inforewiew, magazine from the Serbian CEPIS society JISA
- Informatica, journal from the Slovenian CEPIS society SDI
- Informatik-Spektrum, journal published by Springer Verlag on behalf of the CEPIS societies GI, Germany, and SI, Switzerland
- ITNOW, magazine published by Oxford University Press on behalf of the British CEPIS society BCS
- Mondo Digitale, digital journal from the Italian CEPIS society AICA
- Novática, journal from the Spanish CEPIS society ATI
- OCG Journal, journal from the Austrian CEPIS society OCG
- Pliroforiki, journal from the Cyprus CEPIS society CCS
- Tölvumál, journal from the Icelandic CEPIS society ISIP

Editorial correspondence: Llorenç Pagés-Casas <pages@ati.es>
Advertising correspondence: <info@cepis.org>

Subscriptions
If you wish to subscribe to CEPIS UPGRADE please send an email to info@cepis.org with 'Subscribe to UPGRADE' as the subject of the email or follow the link 'Subscribe to UPGRADE' at <http://www.cepis.org/upgrade>

ISSN 1684-5285

Monograph of next issue (October 2011)
"Green ICT"

(The full schedule of **CEPIS UP**GRADE is available at our website)

CEPIS UPGRADE
The European Journal for the Informatics Professional
http://cepis.org/upgrade
Vol. XII, issue No. 3, July 2011

Monograph
Business Intelligence
(published jointly with Novática*)
Guest Editors: *Jorge Fernández-González and Mouhib Alnoukari*

UPENET (UPGRADE European NETwork)

CEPIS NEWS

* This monograph will be also published in Spanish (full version printed; summary, abstracts, and some articles online) by **Novática**, journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*) at <http://www.ati.es/novatica/>.

Presentation

Business Intelligence: Improving Decision-Making in Organizations

Jorge Fernández-González and Mouhib Alnoukari

1 A Concept hard to define

When we talk about Business Intelligence (BI) it seems that we all clearly understand the concept. Nothing could be further from the truth. BI is a concept difficult to define. Its small nuances and large applications make people understand different things. So the question is: What is Business Intelligence?

BI is a somewhat ambiguous term encompassing a number of different acronyms, tools, and disciplines: OLAP, Data Warehousing, Data Marts, Data Mining, Executive Information Systems, Decision Support Systems, Neural Networks, Expert Systems, Balanced Scorecards, and many others. It is impossible to give an exact definition of all the terms used in Business Intelligence. Some authors have gone as far as calling it a jungle.

The multifaceted and diverse fauna inhabiting this jungle have three characteristics in common.

The first is that they provide information for controlling the business processes, regardless of where the information is stored.

Obviously, BI forms part of a company's information system, which is what controls the proper functioning of the processes performed in the company.

In a classical organization, processes are affected by external perturbations, such as changes in the market, replacement products, new legislation, etc., which must be controlled and corrected. And we all know that over time systems tend toward disorganization and chaos. This is why the measurement of performance indicators and their comparison against the organizations' objectives is the best way to find out if something is going wrong in our organization.

Processes generate and consume information as they are being performed. Part of that information (what we call operational information) is consumed in the short term, but a large proportion is stored in various transactional systems (ERP, CRM, SCM, etc.) until it can be used for tactical (medium-term) and/or strategic (long-term) decision-making.

Grouping this information and putting it at the disposal of the process control system in a timely manner, regardless of which operational system it may have originated in, will help us optimize our processes, whether they are of an

The Guest Editors

Jorge Fernández-González graduated as an Informatics Engineer from the *Facultad de Informática de Barcelona* (UPC), Spain, and is currently pursuing his doctorate in Software, specializing in Information Systems, at the same university. He divides his professional time between three activities. First and foremost he works as an information systems professional as Director of Business Intelligence Consulting at Abast Solutions, a company operating nationwide. Here he has worked in several different areas of consulting in the company's ERP, CRM, and R&D departments while helping with the implementation of tailored solutions. The second of his activities is university lecturing. He is currently lecturing in the LSI department (Department of Languages and Informatics Systems) of UPC (*Universitat Politècnica de Catalunya*) and he is responsible for the subject "Information Systems for Organizations" offered by the *Facultad de Informática de Barcelona*. He has also been a collaborating lecturer at UOC (*Universitat Oberta de Catalunya*), a lecturer for master and postgraduate studies at the *Fundación Politécnica*, and delivers lectures as a guest lecturer at business schools such as ESADE and EAE. He combines the above two activities with his work as a disseminator. He forms part of the editorial team of the journal *Gestión del Rendimiento* (Performance Management), he writes articles for the journal DATA.TI (formerly Datamation), he delivers conferences and seminars, and he writes in various Internet portals and thematic blogs, including his own blog <http://sistemasdecisionales.blogspot.com> dedicated to decisional systems. <jorge.fdez.glez@gmail.com>

Mouhib Alnoukari received his PhD degree from the Arab Academy for Banking and Financial Sciences (Damascus, Syria). He is currently working as an ICT faculty member at the Arab International University, Damascus (Syria) and the Arab Academy for Banking and Financial Sciences, Damascus (Syria). Dr. Alnoukari has (co-)chaired many ICT symposiums and conferences in Syria including the 1st National Symposium of Business Intelligence in Syria (BISY 2010) and MENA ICT Week 2011. His research interests are in the areas of Business Intelligence, Data Mining, Data Warehousing, Agile Methodology, Software Engineering, and Databases in which he has published more than 20 journal and conferences papers. He has also (co-)edited more than 20 ICT books both in Arabic and English languages, including: "Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications" to be published by IGI Global, September 2011. He has also participated in various reference book chapters such as "Handbook of Research on Discrete Event Simulation Environments- Technologies And Applications" by Evon M. O. Abu-Taieh and Asim A. El Sheikh (eds). IGI Global. (2009), and "Business Information Systems: Concepts, Methodologies, Tools and Applications", edited by Information Resources Management Association, USA, IGI Global, 2010. <m-noukari@aeu.ac.sy>

operational, tactical, or strategic nature. Obviously the level of aggregation and standardization of heterogeneous data sources will be higher for processes of a decisional nature, and it is precisely this decisional nature that gives a new dimension to the definition of Business Intelligence: ***decision-making support is the second and most important of the three characteristics*** that all components of Business Intelligence have in common.

BI does not only present information but it makes it possible for that information to be managed and browsed to enable us to analyse causes. Analysis is fundamental to decision-making. Decisions are not made on the basis of a single source of information. Various sources of information are weighed up, interrelated; you might say that the information is "alive". The analysis ability of information is what enables us to make better business decisions.

We cannot make business decisions if we do not talk the language of business. Regardless of where the information is stored and how it may have been transformed or aggregated, the important thing is to deliver this information to business users in a language that they understand, are comfortable with, and which needs no interpretation for them to understand it. And ***this is the third characteristic of BI: information oriented towards the language of business users***. In this way their work is made easier and the decision-making required to improve processes and gain a competitive edge in the market is speeded up.

We might therefore define Business Intelligence as the system which provides us with the information required to control processes, and the information used by business users for the purpose of decision-making.

Perhaps the most important characteristic of BI is that it is focused on enabling business users to make decisions with semantically appropriate information. We are not talking about either data or IT; we are talking about business and information users.

2 What is in this Monograph?

At this point we started this **UP**GRADE monograph by imagining a scenario where a business analyst is looking at the information contained in a business report. What would happen f our analyst had misunderstood one graph?

Because analysts' brains have different ways of understanding or discarding the information displayed it is very important how we show it. For this reason we open this issue with an article by ***Josep-Lluis Cano-Giner***, from ESADE Business School (Spain), entitled "*Business Information Visualization*".

But once we have submitted this information we need to have tools that allow us to use it efficiently. That's what ***R. Dario Bernabeu*** and ***Mariano A. García-Mattío*** from the eGlu Business Intelligence Group (Argentina) talk about in their article "*BI Usability: Evolution and Tendencies*".

On the other hand, the mature form of each organization and its adoption of different BI tools change according to multiple factors. ***Paul Hawking*** from Victoria University (Australia) describes a case study on company's trials and tribulations in regard to their Business Intelligence implementations. His article is entitled "*Towards Business Intelligence Maturity*".

To achieve a good level of BI Maturity we need to choose a good BI platform. Choosing the right tools depends on the specific needs and goals that an organization is trying to optimize, along with the nature of its data and analysis requirements. ***Mahmoud Alnahlawi***, a software architect from Palo Alto (California, USA), shares his knowledge with us in the article "*Business Intelligence Solutions: Choosing the Best Solution for your Organization*".

But not only a good set of tools can assure our success. We must think about a strategic plan, too. In his article "*Strategic BI for NGOs*", ***Diego Arenas-Contreras*** from the company Formulisa (Chile) explains to us how to plan and apply a Business Intelligence (BI) strategy to a nonprofit organization starting from the understanding of organizational processes and the identification of information needs, relevant available data and proprietary information to meet the information requirements that an organization has.

Nothing of what we have said so far can come true if we are working with erroneous or inconsistent data. Decision-making is based on the information we obtain from business data and all decision-making involves accepting a certain degree of risk, but the truth is that it is not always possible to have complete and hard data available. ***Oscar Alonso-Llombart***, from the company Penteo (Spain), share his experiences with us in the article "*Data Governance, what? how? why?*".

In the same way we need to extract, transform, and load data into our data warehouses. ***Veit Köppen***, from the Otto-von-Guericke University (Magdeburg, Germany), ***Björn Brüggemann***, from the company Capgemini (Germany), and ***Bettina Berendt***, from the *Katholieke Universiteit Leuven* (Belgium) tell us about the ETL Process in their article "*Designing Data Integration: The ETL Pattern Approach*".

All BI projects need a good methodological approach to succeed. In this way one of the guest editors of this monograph, ***Mouhib Alnoukari*** from the Arab International University (Damascus, Syria), shares with our readers the knowledge and experiences gained while preparing a book which he has authored on the use of agile methodologies for building Business Intelligence applications with an article entitled "*Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications*".

Finally, the monograph closes with the article "*Social Networks for Business Intelligence*" by ***Marie-Aude Aufaure*** and ***Etienne Cuvelier*** from the MAS Laboratory at the *Ecole Centrale Paris* (France), which explains the integration of social networks in enterprises and public administrations from the business intelligence point of view.

To close this presentation, let us express our most sincere thanks to all the authors for their valuable contribution. We also would like to express our gratitude to the Chief Editor of **UP**GRADE ***Llorenç Pagés-Casas***, for giving us the opportunity to prepare this monograph and for his support during the preparation process.

Business Information Visualization

Josep-Lluís Cano-Giner

Managers have more and more data available and less and less time to access it, as they need to make decisions quickly. Its correct representation can become a key element for facilitating decision making. The paper starts with a review of the history and importance of information visualization. We also provide an example of how this visualization can be improved, and we conclude with an account of new needs that are arising in this field, as reflected by both organizations and their managers.

Keywords: Business Intelligence, Graphical Representation, Information Visualization, Management Information Systems.

Author

Josep-Lluís Cano-Giner has been a Professor of the Department of Information Systems Management at ESADE since 1989. He has a Degree in Business Sciences and a Master in Business Administration and Management from ESADE. He also has a Degree in Business Administration and Management from the Technical University of Catalonia (UPC). He obtained his Diploma of Advanced Studies (DEA) at UPC. He is currently engaged in research in the field of factors affecting the increase in use of management information systems by managers. As a business consultant, he has ample experience in strategic planning of information systems and solution selection, especially in business intelligence. His academic publications include papers and business cases such as "Government intelligence in Catalan public universities", "Aventis", "Aguirre&Newman", "Bodegas Torres", and "Raventós i Blanc at a crossroads". He is also author of the book *Business intelligence: Competir con información* (Competition through Information). <josepllluis.cano@esade.edu>

1 Introduction

The applications used by organizations are generating ever larger amounts of information. This information is handled in real or almost real time. Knowledge creation and decision making are the two main reasons why organizations store information, along with operations support and the need to fulfil legal obligations. Both reasons depend on the criteria of individuals, who have to use the visualization of the presented information to extract key aspects that will enable them to recognize hidden patterns or trends. The visualization thus becomes the interface between computers and people's minds. The cognitive capacities of humans have limitations; by visualization we mean the process of transforming data, information and knowledge into a representation to be used in a way that shows an affinity with the cognitive capacities of human beings.

2 Examples from the History of Information Visualization

Several authors have investigated the history of information visualization, most notably Tufte [1]. Here we will look at three examples provided by three different authors corresponding to three different representations, with the aim of showing both the advantages and the disadvantages of using information visualization.

In 1786, the Scottish engineer William Playfair realized that economic transactions could easily be represented graphically. Furthermore, in his opinion, representation using time series and bar charts simplified understanding and retention. The author published his *Commercial and Political Atlas*, in which he described England's foreign trade. It also included, for the first time, a new type of graph: the pie chart. In Figure 1 we reproduce a graph of the exports and imports between England and Denmark plus Norway [2]. In it we clearly see the moment at which the sign of the balance of trade between the two countries changes, together with the growth of the balance in England's favour.

One of the most famous examples of information presentation was provided by Charles Minard, a French civil engineer who used visualization to tell the story of Napoleon's tragic march on Moscow[1] in 1812. In the diagram of Figure 2, he used a coloured bar, the width of which indicated the size of the army (originally 422,000 troops), to show how the forces gradually dwindled as they approached

[1] See <http://www.math.yorku.ca/SCS/Gallery/images/minard.gif>.

> "Correct representation of data can become a key element for facilitating decision making in organizations"

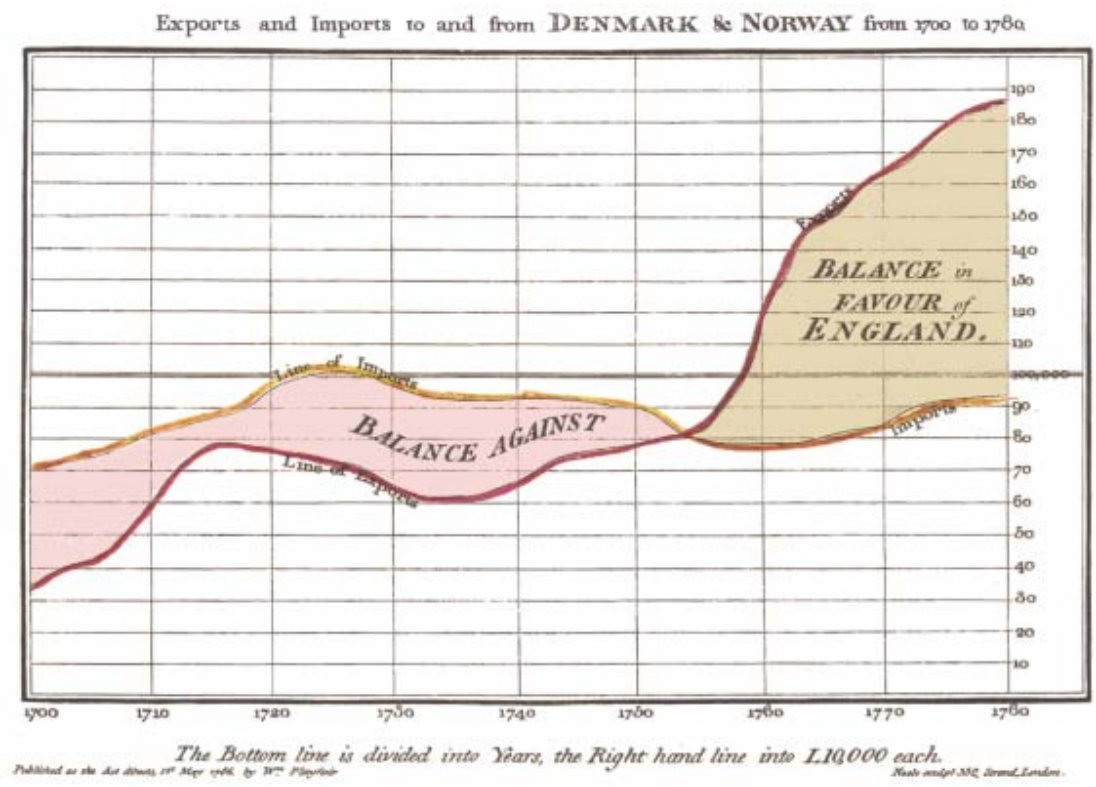

Figure1: Exports and Imports between Denmark and Norway from 1700 to 1780. [Source: W. Playfair.]

the Russian capital. In turn, another bar, this time black, indicated those who returned from Moscow (only 10,000 troops got home). At the foot of the diagram we find the outdoor temperatures, which were the soldiers' greatest problem. In the middle of the diagram, we see the widening of the black bar, due to the incorporation of stragglers who had tried to advance on the left flank, and also the dramatic narrowing when they had to cross a river with icy water. At the end of the retreat, we can compare the width of the two bars: the coloured bar representing those who set out, and the black one, those who returned. A simple diagram shows us the course of history in a very powerful manner. Robert Spence [3] wonders if we could listen to Tchaikovsky's 1812 overture and view the diagram at the same time[2].

[2] The reader can perform this exercise by accessing <http://www.youtube.com/watch?v=k-vQKZFF-9s&feature=related> [last accessed 5.1.2011]. This overture, Op. 49, was composed to commemorate the victorious Russian resistance in 1812 against the advance of Napoleon Bonaparte's Grande Armée. The overture was premiered in Moscow on 20 August 1882. The work is recognized by its triumphant finale, which includes a salvo of cannon fire and the pealing of bells.

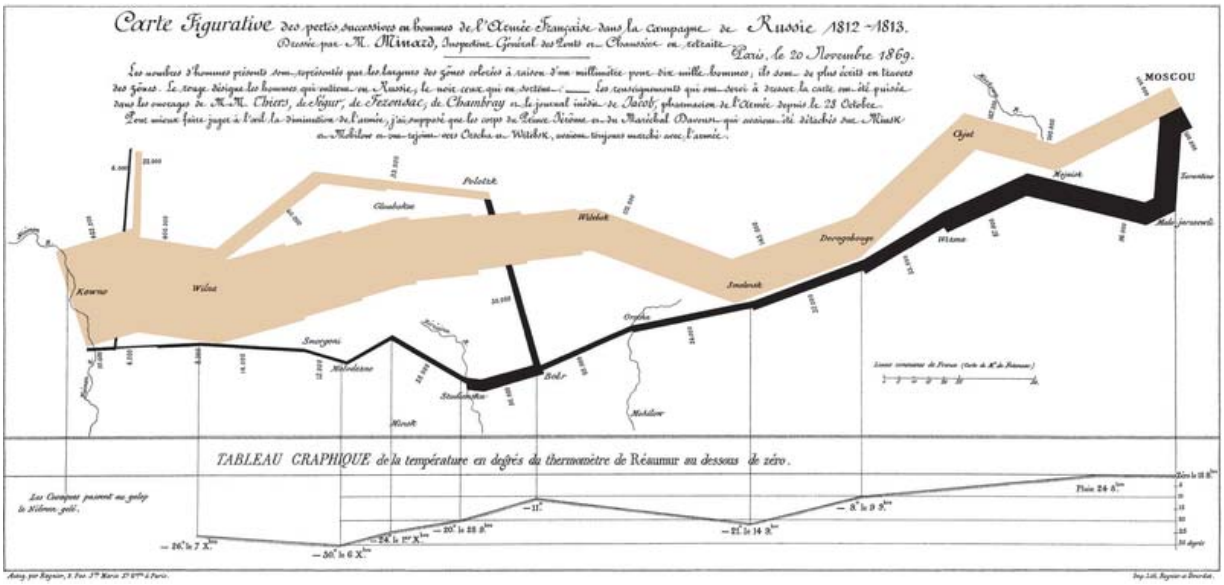

Figure 2: Map of Napoleon's Forces in the Russian Campaign. [Source: Charles Minard, 1861.]

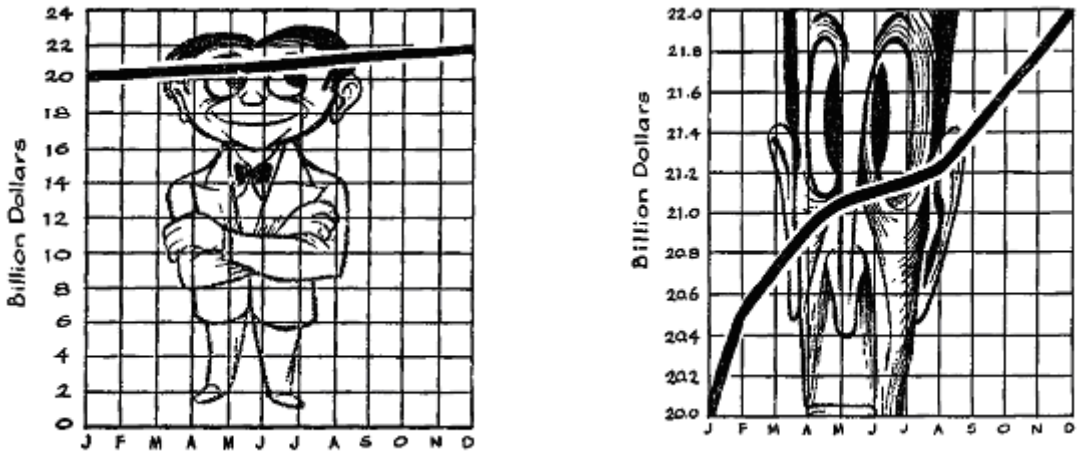

Figure 3: Graph showing how a Figure changes when the Scale of the Axis changes. [Source: Darrell Huff.]

In 1954, Darrell Huff published *How to Lie with Statistics* [4], in which he showed how the graphical representation of statistics can be manipulated to support different, sometimes conflicting, interests. Obviously his great contribution was to show us how to do it right. In Figure 3 we provide an example of the representation of a line graph that is very useful for depicting trends or forecasts. The X-axis (the horizontal one) indicates the months of the year, while the Y-axis (the vertical one) displays the volume; for example, of sales in billion dollars. In the left-hand graph the information is represented correctly: the Y-axis starts at 0 and the distances between the values of the two axes are equivalent. On the other hand, in the right-hand graph the Y-axis starts at 20, with the result that the expression of the character who appears superimposed on the graph changes to one of astonishment at the results obtained.

In reviewing three of the historical examples of information visualization we have highlighted, in the first case, the importance of representing information to grasp what is happening; in the second, that a good representation of information enables us to understand a situation better; and in the third, that if the representation of the information is manipulated, intentionally or otherwise, it can bring us to interpret the facts wrongly. If the representation is right we will be able to make the right decisions, but what will happen if someone manipulates the representation to other ends?

> “Information visualization is defined as ‘*The use of computer-supported, interactive, visual representations of abstract data to amplify cognition*’ ”

EVOLUTION OF THE RECORD FOR THE 100 METRES

Athlete	*Date*	*Time*
Asafa Powell (JAM)	14-6-2005	9,77s
Tim Montgomery (EE UU)	14-9-2002	9,78s
Maurice Greene (EE UU)	16-6-1999	9,79s
Donovan Bailey (CAN)	27-7-1996	9,84s
Leroy Burrell (EE UU)	6-7-1994	9,85s
Carl Lewis (EE UU)	25-8-1991	9,86s
Leroy Burrell (EE UU)	14-6-1991	9,90s
Carl Lewis (EE UU)	24-9-1988	9,92s
Calvin Smith (EE UU)	3-7-1983	9,93s
Jim Hines (EE UU)	14-10-1968	9,95s

Figure 4: Graph of the Evolution of the World Record for the 100 Metre Race. [Source: El País, 15/06/2005.]

3 Visualizing Information

Information visualization is used in fields as varied as medicine, engineering, statistics, business and even sport. We have chosen the last of these to illustrate the difficulties that are – or may be – encountered in presenting information graphically, through a graph (see Figure 4) published in a well-known Spanish newspaper[3]. The graph is reproduced below as published, showing the world record for the 100 metre race through history: the athletes, their nationality, the date the new record was set, and the times. The graph also includes a grey bar together with a picture of a sprinter. The reader may be surprised to find that the

[3] The article "Huracán Powell" appeared in El País on June 15, 2005.

EVOLUTION OF THE RECORD FOR THE 100 METRES

Athlete	Time	Date	Finish
Asafa Powell (JAM)	9,77s	14-6-2005	1
Tim Montgomery (EE UU)	9,78s	14-9-2002	2
Maurice Greene (EE UU)	9,79s	16-6-1999	3
Donovan Bailey (CAN)	9,84s	27-7-1996	4
Leroy Burrell (EE UU)	9,85s	6-7-1994	5
Carl Lewis (EE UU)	9,86s	25-8-1991	6
Leroy Burrell (EE UU)	9,90s	14-6-1991	7
Carl Lewis (EE UU)	9,92s	24-9-1988	8
Calvin Smith (EE UU)	9,93s	3-7-1983	9
Jim Hines (EE UU)	9,95s	14-10-1968	10

Figure 5: Proposed Graph of the Evolution of the World Record for the 100 Metre Race. [Source: Author.]

> “In 1786, the Scottish engineer William Playfair realized that economic transactions could easily be represented graphically”

longest bar belongs to the fastest time. I was surprised too.

After analysing the graph for a while, I realized that what the author was trying to represent was what the finish line would have looked like if the race had been run by the 10 sprinters who held the world record for the 100 metres (Carl Lewis and Leroy Burrell broke the record twice, so they would be running in two lanes).

Working on the above graph, we could propose some improvements, for instance adding the finish line to the graph and changing the position of the times to make them easier to read and understand. We propose a possible solution in Figure 5.

Assuming that the proposed representation makes the information easier to interpret, we could now go on to ask ourselves whether the gap represented between the sprinters corresponds to the real gap. Using the data provided, we can calculate that the gap between the first record and the last would be 1.81 metres, i.e., in the 9.77 seconds that Asafa Powell took to run 100 metres, Jim Hines would have run 98.19 metres. So, does the gap between the sprinter and the finish line correspond to 1.81 metres? To answer this question we represent the distances by making use of spreadsheet graphics[4]. The result is shown in Figure 6.

[4] In Figures 6 to 10, we have omitted the presentation of values in order to facilitate understanding of the effect that we want to show with the graphic.

Figure 6: Graphic of the Evolution of the World Record for the 100 Metre Race, starting at 0 Metres. [Source: Author.]

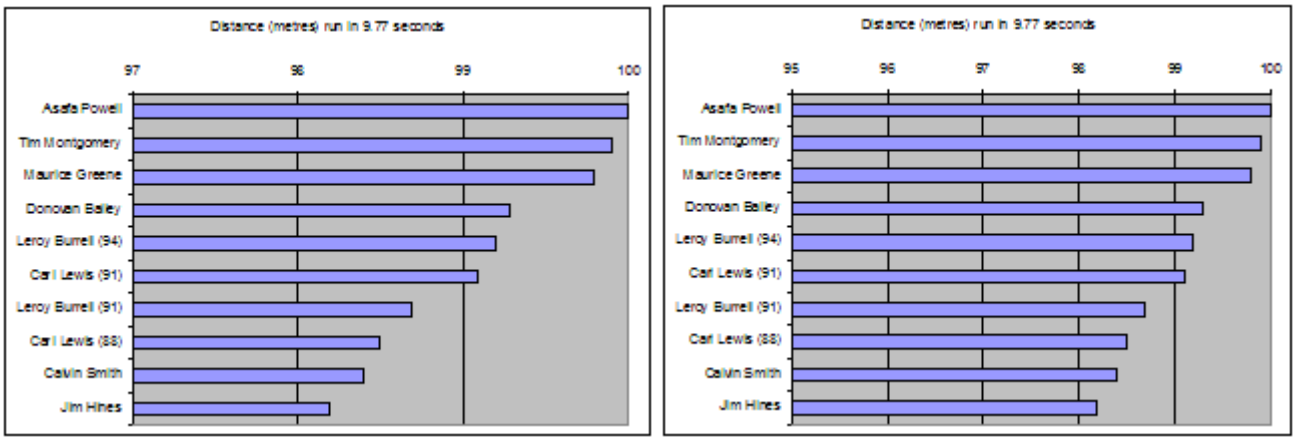

Figure 7: Graphic of the Evolution of the World Record for the 100 Metre Race, starting at 97 and 95 Metres. [Source: Author.]

It is apparent at a glance that the gaps between the sprinters have shrunk in relation to the original graph. What has happened? In the original graph the values of the X-axis are not shown, so we might wonder what value they start at.

In Figure 7 we present two graphics. The left-hand graphic starts at 97 metres, and the right-hand one at 95 metres.

If we change the lower limit of the X-axis when visualizing the graphics, the person who analyses them may interpret them differently, and this may bring them to make a different decision.

We might also ask ourselves whether distance run is really the best variable to represent the differences between the various world records for the 100 metre dash. Alternatively, we could focus on speed, i.e., metres per second, to show the difference between sprinters.

In Figure 8 we show the graphic of the speeds obtained in the 10 world records we are contemplating. In this case

> “In 1954, Darrell Huff published *How to Lie with Statistics*, in which he showed how the graphical representation of statistics can be manipulated to support different, sometimes conflicting, interests”

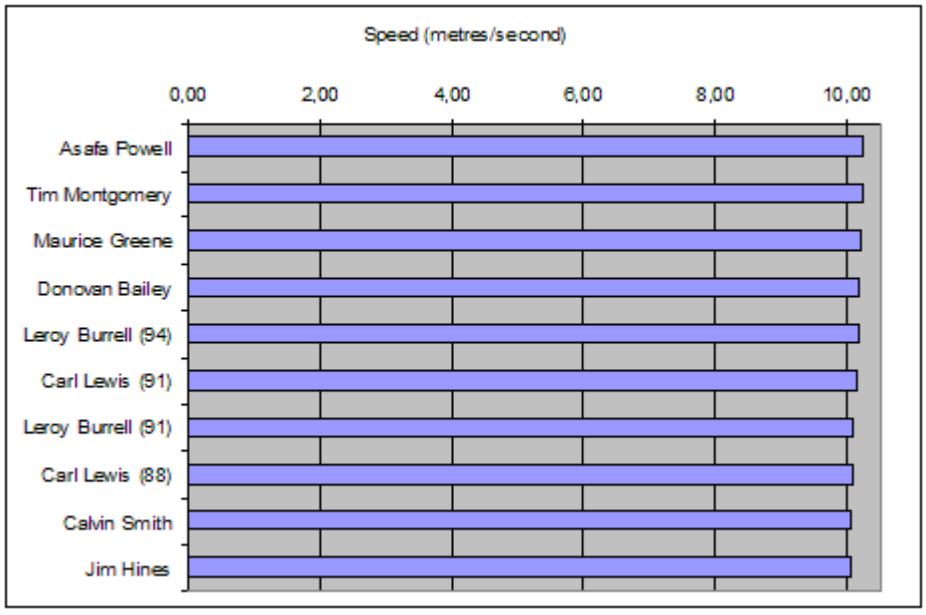

Figure 8: Graphic of Speed of World Records for the 100 Metre Race. [Source: Author.]

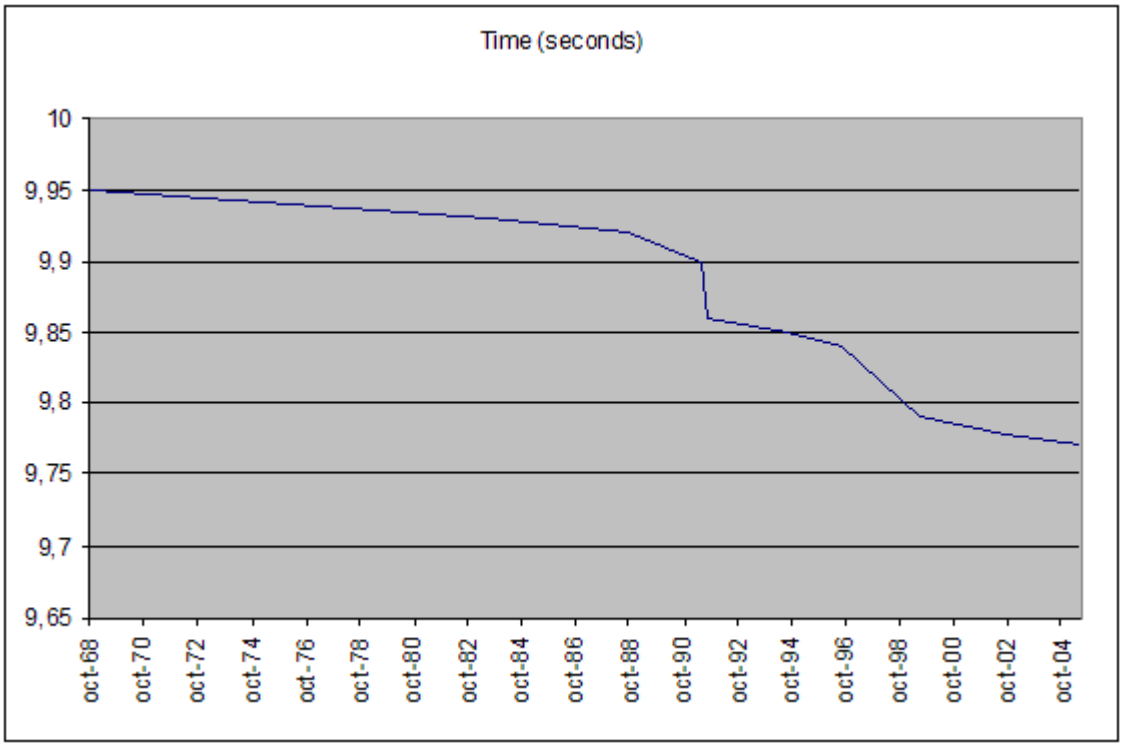

Figure 9: Graphic of the Evolution of the World Record for the 100 Metre Race, starting at 9.65 Seconds. [Source: Author.]

we only present the graphic in which the X-axis starts at 0, since if it started at some other value we would find the same as in the previous case. The difference in metres per second is only 0.19 between the first and the last.

Let us go a little further in the analysis. Returning to a historical perspective, we could ask how the 10 fastest 100 metre world records have evolved over time[5].To do this,

[5] It is very important to stress that we are not asking about how the 100 metre world record has evolved throughout history. To answer this question we would need to have all the world records, and it is not the purpose of this paper to analyse them.

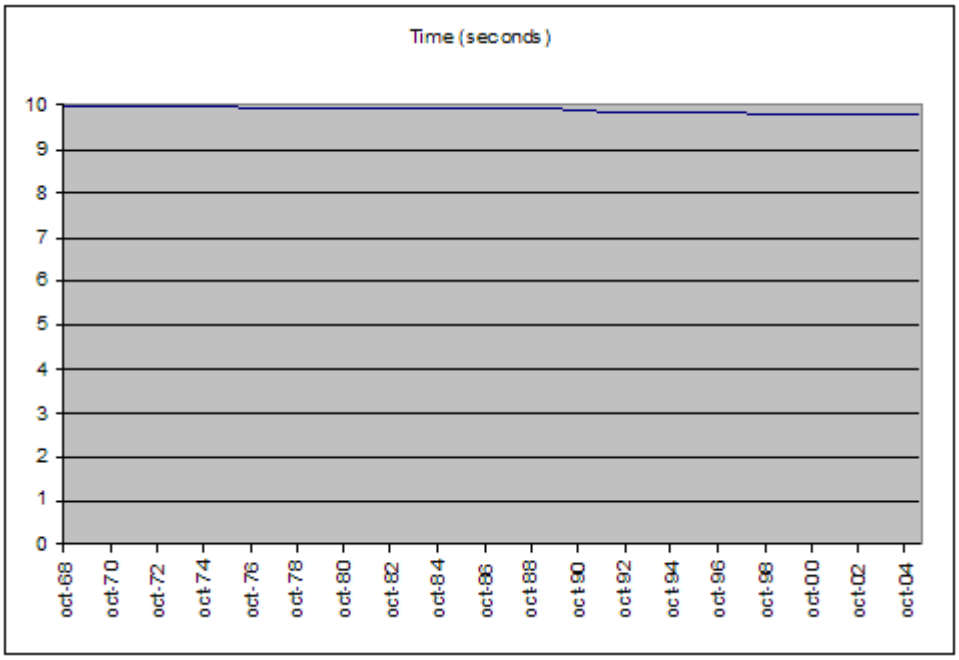

Figure 10: Graphic of the Evolution of the World Record for the 100 Metre Race, starting at 0 Seconds. [Source: Author.]

Date	Record	Athlete	Nationality	Time (seconds)	Time lag %	Speed (m/s)	Distance in metres run in 9.77s
14/06/2005	1	Asafa Powell	JAM	9,77	0,00%	10,24	100,00
14/09/2002	2	Tim Montgomery	EEUU	9,78	0,10%	10,22	99,90
16/06/1999	3	Maurice Greene	EEUU	9,79	0,20%	10,21	99,80
27/07/1996	4	Donovan Bailey	CAN	9,84	0,72%	10,16	99,29
06/07/1994	5	Leroy Burrell (94)	EEUU	9,85	0,82%	10,15	99,19
25/08/1991	6	Carl Lewis (91)	EEUU	9,86	0,92%	10,14	99,09
14/06/1991	7	Leroy Burrell (91)	EEUU	9,90	1,33%	10,10	98,69
24/09/1988	8	Carl Lewis (88)	EEUU	9,92	1,54%	10,08	98,49
03/07/1983	9	Calvin Smith	EEUU	9,93	1,64%	10,07	98,39
14/10/1968	10	Jim Hines	EEUU	9,95	1,84%	10,05	98,19

Table 1: Evolution of the World Record for the 100 Metre Race. [Source: Author.]

> “If the representation is right
> we will be able to make
> the right decisions,
> but what will happen
> if someone manipulates
> the representation to other ends?”

	Objectives of graphics	Original graph of the world record for the 100 metre race
1	Tufte suggests that we should simply show the data. Sometimes graphic designers tend to show aggregations of data instead of the data itself.	The sprinters' countries are given next to their names. If they are made to appear in a separate column the information will be easier to read and the preponderance of US sprinters will be better reflected.
2	He suggests that we ensure that the user is thinking about the substance of the graphic and not the graphic itself.	In order to interpret the graph it was necessary to recognize that the author was attempting to represent the finish line of a hypothetical race between the last 10 record-holding sprinters. The ranking of each record should be indicated.
3	Avoid all unnecessary decorations.	The representation of the sprinters is not necessary.
4	Compress as much information as possible into as small a space as possible.	Metres run or the speed of each record could have been included.
5	Graphics should be designed to encourage the user to make comparisons between different pieces of data.	The times are not aligned, which makes them difficult to compare.
6	Graphics should provide views of the data at many levels of detail.	As we are dealing with a single non-interactive graph, this does no apply.

Table 2: Tufte's Principles compared to the Original Graph of the World Record for the 100 Metre Race. [Source: Author.]

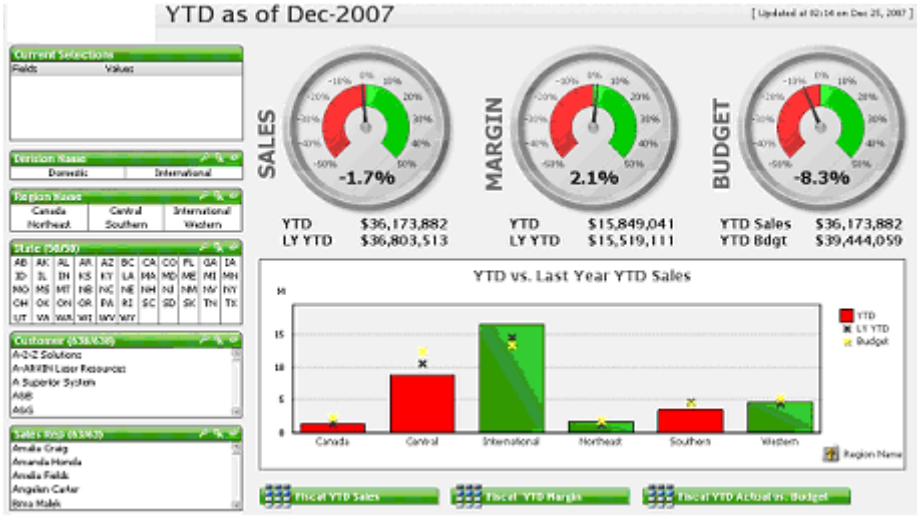

Figure 11: QlikView Dashboard Example.

we can use a graphic indicating the values over time. The default graphic offered by the spreadsheet is shown in Figure 9.

Again, a cursory glance might lead us to a wrong conclusion: that the record has been slashed, as the slope is steep. But if we start the Y-axis at a value of 0, the graphic changes significantly (see Figure 10).

It might look as if no times have been plotted in the above graphic. However, if we look carefully we find that there is a line between the values 9 and 10 seconds, with a slight downward slope over about 37 years.

In other words, it has taken nearly 37 years to reduce the time by 10 hundredths of a second, which means about 5 thousandths of a second each year. The second graphic seems to reflect this situation better than the first. In the extreme, two different representations of the same data can lead us to two conclusions that may even contradict each other.

Through graphical representation, we have enabled those people who were interested in the phenomenon under analysis to grasp what has actually happened in the last 10 world records for the 100 metres, the little difference there is between them, and the difficulty involved in improving on them. Perhaps it would have been better to use another sort of graphical representation: a table with the values and the calculations used to represent them graphically. Three new columns have been added to Table 1: the percentage time lag behind the fastest record, speed in metres per second, and the distance the other sprinters would have run in the time it took Asafa Powell to reach the finish line.

If the conclusion is that the best way to understand this situation is to use the table with values, then this is the best representation. This decision probably comes down to personal choice, and at the same time is influenced by each person's knowledge of the results of the 100 metre event. That is to say, it depends not only on who represents the information, but also on who visualizes it. We might ask ourselves whether it always makes sense to establish an invariable format for reporting over the years, as most organizations do, on the grounds that by not changing the format they facilitate interpretation, or whether we should change it in order to achieve an improvement in the visualization if we want to find a better representation of changes that have taken place in the data.

4 Information Visualization

According to Card et al. [5], information visualization is defined as "*The use of computer-supported, interactive, visual representations of abstract data to amplify cognition*"

The authors differentiate it from scientific visualization, which is usually based on physical data.

The same authors have carried out a literature review[6], at

	Close	Max	Min
AT&T	40,28	41,34	33,30
Boeing	98,15	100,59	84,79
Citigroup	53,98	55,20	48,27
Exxon Mobil	85,94	85,94	69,56
General Electric	38,12	38,12	34,09
General Motors	34,66	36,20	28,85
Intel	24,24	24,24	18,76
Microsoft	30,49	31,11	26,63

Figure 12: Table of Changes in Market Prices. [Source: <http://www.edwardtufte.com>.]

> "Information visualization is used in fields as varied as medicine, engineering, statistics, business and even sport"

> “We might ask ourselves whether it always makes sense to establish an invariable format for reporting over the years, or whether we should change it in order to achieve an improvement in the visualization”

the same time justifying how visualization amplifies cognition, or in other words, the concept of cognition (from the Latin *cognoscere*, "to know") refers to human beings' faculty of processing information through perception, acquired knowledge and the subjective characteristics that allow them to evaluate it.

Note that the proposed definition includes the term *interactive*, as nowadays representations are usually based on the use of computers, thus allowing interaction between the user and the computer application.

Those readers who wish to go deeper into the definition of *visualization*, the various technologies that support it or the relationship between cognitive theories and problem solving tasks, as well as visual representations, are referred to the paper by Tegarden [6].

In this paper, Tegarden summarizes the six objectives that any graphic should meet according to Tufte [1][7]. In Table 2, these objectives are related to the original graph of records for the 100 metres that we have used as an example.

5 Business Information Visualization

Managers need information to make decisions, and they need it to be presented in such a way as to facilitate its interpretation. To this end, organizations usually develop business intelligence projects. One of the key aspects of these projects is the correct representation of data. In the present paper we will not deal with the basic concepts of data representation. For this, the bibliographical references given below provide an ample in-depth account of this issue. Nevertheless, we will give attention to new trends and needs in visualization. One of the authors who stand out most in this field is Stephen Few [8]. In his view, the information visualization of the future will have to cope with new needs, as discussed below.

- *Dashboards and scorecards*: Managers need to be able to access data that will enable them to analyse the situation in a short space of time, in such a way that once a problem has been detected, a few clicks with the mouse are enough to get down to the right level of detail to grasp what is happening and take corrective measures. Scorecards represent perspectives of strategic areas, objectives, measures, and stoplight indicators, whereas in dashboards the information presented can vary considerably and usually includes graphical representations. In dashboards the complexity of the information visualization increases, as they can be used to present interrelated data or graphics. In most cases it is also necessary to present a large amount of information in a very limited space (see Figure 11).
- *Geospatial visualization*: When data is land-based, in situ visualization is becoming increasingly necessary. From the well-known geographic information systems (GIS) to Google Earth and various types of web services, we find media that can enable us to relate sales or expenses, for example, to geographical variables.
- *Animated scatter plots*: In some cases we need to compare two magnitudes, e.g., investment in advertising and sales over time. To do this, it is necessary to make use of a new type of representation that includes animations to convey the passing of time. One of the best examples of animated data representation is <http://www.GapMinder.org>, where we can see, for example, the relationship between income per person in different countries and their child mortality rate over time.
- *Treemaps*: One example of this type of graphical representation is the trading volume on the New York Stock Exchange, aggregated by industry in such a way as to allow the viewer to compare prices and see how they have changed from the day before (available at <http://www.SmartMoney. com>).
- *Sparklines*: This type of graphical representation is characterised by its small size and high data density. It is common practice to use several at once in order to represent different information that can sometimes be complementary. The term *sparkline* was proposed by Edward Tufte, who described them[7] as "small, high-resolution, simple, word-sized graphics". An example of a sparkline is shown in Figure 12.

[6] Table 1.3 on page 16 of the book cited in reference [5].

[7] More examples are available at <http://www. edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001OR&topic_id=1>.

> “Organizations develop business intelligence projects; one of the key aspects of these projects is the correct representation of data”

> “In any business intelligence project we should ensure that the graphical representation is the most appropriate one, and therefore we need specialists to this end”

■ *Representing relationships*: Sometimes we need to represent relationships among entities, such as among websites. Each entity acts as a node in a network, and has links with others. One example is Vizster showing people networks, (see <http://hci.stanford.edu/jheer/projects/vizster/early_design/>).

6 Conclusion

Information visualization needs have changed over the ages. The time available to managers to make decisions has become shorter and shorter, and new needs have arisen. As a result, researchers have proposed – and will continue to propose – new solutions to meet them. Correct representation of data should facilitate its interpretation and shorten the time managers have to spend on it, and this stands as the main purpose of information visualization.

If the information visualization is not the right one it may cause managers to make the wrong decisions. In this paper we have presented several examples in which the visualization of information in a "manipulated" fashion could lead to wrong interpretations, with the attendant increased risk of a mistake being made by management. In any business intelligence project we should ensure that the graphical representation is the most appropriate one, and therefore we need specialists who can guarantee this with the minimum information visualization. Without the right representation we will not provide the value that is expected, and we will be hard pressed to retain management’s interest in using this solution.

References

[1] E.R. Tufte. The Visual Display of Quantitative Information. Cheshire: Graphics Press, 1983.

[2] W. Playfair, H. Wainer, I.Spence. The Commercial and Political Atlas and Statistical Breviary. New York: Cambridge University Press, 2005.

[3] R. Spence. Information Visualization. Essex: ACM Press, 2001.

[4] D. Huff. How to Lie with Statistics. Penguin; New Ed edition, 1991.

[5] S.K. Card, J.D. Mackinlay, B. Shneiderman. Readings in Information Visualization: Using Vision to Think. San Francisco: Morgan Kaufmann Publishers, 1999.

[6] D. P. Tegarden. "Business Information Visualization". Comunications of AIS, vol.1, art. 4, enero, 1999.

[7] E.R. Tufte. Envisioning Information. Cheshire: Graphics Press, 1990.

[8] S. Few. Data Visualization, Past, Present, and Future. Percentual Edge, 2007.

Recommended Bibliography

■ J. Best. Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists. Berkeley and Los Angeles: University of California Press, 2001.

■ J. Best. More Damned Lies and Statistics: How Numbers Confuse Public Issues. Berkeley and Los Angeles: University of California Press, 2004.

■ W.S. Cleveland. The Elements of Graphing Data. Summit: Hobart Press, 1994.

■ S. Few. Show Me the Numbers: Designing Tables and Graphs to Enlighten. Oakland: Analytics Press, 2004.

■ S. Few. Information Dashboard Design: The Effective Visual Communication of Data. Sebastopol, CA: O’Reilly Media, Inc., 2006.

■ S. Few. Now You See It: Simple Visualization Techniques for Quantitative Analysis. Oakland: Analytics Press, 2009.

■ J.G. Koomey. Turning Numbers into Knowledge: Mastering the Art of Problem Solving. Oakland: Analytics Press, 2001.

■ D. Niederman, D. Boyum. What the Numbers Say: A Field Guide to Mastering Our Numerical World. New York: Broadway Books, 2003.

■ N.B. Robbins. Creating More Effective Graphs. Hoboken: John Wiley and Sons, Inc., 2005.

■ T. Segaran, J. Hammerbacher. Beautiful Data. O’Reilly Media, 2009.

■ E.R. Tufte. Visual Explanations. Cheshire: Graphics Press, 1997.

■ E.R. Tufte. Beautiful Evidence. Cheshire: Graphics Press, 2005.

■ C. Ware. Information Visualization: Perception for Design, second edition. San Francisco: Morgan Kaufmann Publishers, 2004.

■ D.W. Wong. The Wall Street Journal Guide to Information Graphics: The Dos and Don’ts of Presenting Data, Facts, and Figures. New York: W. W. Norton & Company, 2010.

BI Usability: Evolution and Tendencies

R. Dario Bernabeu and Mariano A. García-Mattío

This article initially introduces us to the concepts of usability and Business Intelligence (BI), in order to later define BI Usability. First, a historical graphic is presented with the most important highlights, which are the antecedents for what are now known as BI systems. Later, these highlights are systematised, decade by decade, taking into account evolution and innovation in BI information systems, and also highlighting usability in each time period. Finally, the way BI Usability was developed through time is described, and usability tendencies are identified.

Keywords: BI, Business Intelligence, Historic Evolution, Usability.

1 Introduction

What is known today by the name Business Intelligence (BI) has an origin and evolution that should be looked at in order to introduce the concept that will be the subject of this article: "BI Usability".

One of the principal goals of BI is that users find the information they need to make decisions in due time and proper form. The form includes, among other things, the format in which the information is presented and the level of interaction expected to obtain the desired result. The previous points make up the term "BI Usability".

Usability can be defined as software's ease of use, in which factors such as the familiarity of the design, comfort, attractiveness, level of interaction permitted, response time, etc., also come into play.

Various definitions of usability have been selected to complement the concept[1] :

- The ISO/IEC 9126 defines usability as "*the software's capacity to be understood, learned, used, and to be attractive to the user in specific use conditions*." At the same time, the ISO establishes four basic principles on which usability is based: ease of learning, ease of use, flexibility, and robustness.
- Jakob Nielsen, the father of usability, defines it as "*a quality attribute that evaluates how easy user interfaces are to use*."
- Janice (Ginny) Redish, independent consultant, defines what an interface should allow users to do: "*find what*

[1] The following definitions were taken from Wikipedia.

Authors

R. Dario Bernabeu is a Systems Engineer by the IUA (*Instituto Universitario Aeronáutico* - Aeronautical Universitary Institute) of Córdoba, Argentina. Cofounder of eGluBI <http://www.eglubi.com.ar>, he specialises in development and implementation of OSBI solutions (Open Source Business Intelligence). project management, analysis of requirements/needs, deployment and configuration of BI solutions, design of data integration processes, Data Warehouse modelling, design of Multidimensional Cubes and Business Models, development of ad hoc reports, advanced reports, interactive analysis, dashboards, etc. He is a teacher, researcher, geek, and open-source software enthusiast, and his most notable publication is "Data Warehousing: Research and Concept Systematisation – HEFESTO: Methodology for the Construction of a DW". Coordinator of the social network Open BI Network, <http://www.redopenbi.com>, he makes many contributions to various forums, wikis, blogs, etc. <DarioSistemas@gmail.com>

Mariano A. García-Mattío is a Systems Engineer by the IUA (*Instituto Universitario Aeronáutico* - Aeronautical Universitary Institute) of Córdoba, Argentina, and Specialist in Distributed Systems and Services by the FaMAF - UNC (*Facultad de Matemática, Astronomía y Física de la Universidad Nacional de Córdoba* - Faculty of Mathematics Astronomy and Physics of National University of Córdoba). He is an Associate Professor at the IUA. He is a teacher in charge of Applied Databases at the UCC (Catholic University of Córdoba**)**. Co-director of the research project on New Information and Communication Technologies at the UCC. Member of the "Virtual Laboratories" research project at the IUA. Co-founder of eGluBI <www.eglubi.com.ar>. Coordinator of the social network Open BI Network. He specialises in JAVA SE and Java EE technologies, administration and design of databases and OSBI. <magm3333@gmail.com>

> "One of the principal goals of BI is that users find the information they need to make decisions in due time and proper form"

> *BI Usability* refers to the design of software dedicated to BI that includes an interface that is friendly, intuitive, and easy-to-use. *Usability* can be defined as software's ease of use

they need, understand what they find, and act appropriately, within the limits of time and effort that they consider adequate for the task."

Business Intelligence could be defined as a concept that integrates, on one hand, storage, and on the other hand, processing of large quantities of data, with the principal goal of transforming it into knowledge and decisions in real time, through simple analysis and exploration. This knowledge should be timely, relevant, useful, and should be adapted to the organisational context[2].

In the framework of these conceptual approximations to usability and BI, it is possible to propose a conceptualisation of BI Usability. *BI Usability* refers to the design of software dedicated to BI that includes an interface that is friendly, intuitive, and easy-to-use (and easy to learn to use); an interface that allows for the creation of new contents (interactive analysis, reporting, dashboards), as well as content navigation, with an emphasis on the presentation of these contents, all in a visual and interactive manner, so the user feels comfortable with his tool and takes full advantage of his data.

2 Historical Evolution of Usability in BI

In the following we list the principal highlights that occurred which are antecedents to the shape that BI systems have taken today regarding usability. Figure 1 details this historical development.

In the following, the impact of usability in each of these stages will be shown.

2.1 1960s

BI information systems: In the nineteen-sixties systems were based on files with an almost total hardware dependence. They were principally oriented towards data storage and treatment, but sequential storage systems (tapes) largely impeded the possibility of managing information[3]. The emergence of direct access, together with the creation of the first hard drives, marked a milestone, after which

[2] Datawarehousing: Research and Concept Systematisation – HEFESTO: Methodology for the Construction of a Data Warehouse – Version 2.0 - Bernabeu, R. Dario.

[3] *Information* is a set of organised, ordered, and processed data that constitutes a message that changes the knowledge state of the subject or system that receives it. Source: Wikipedia.

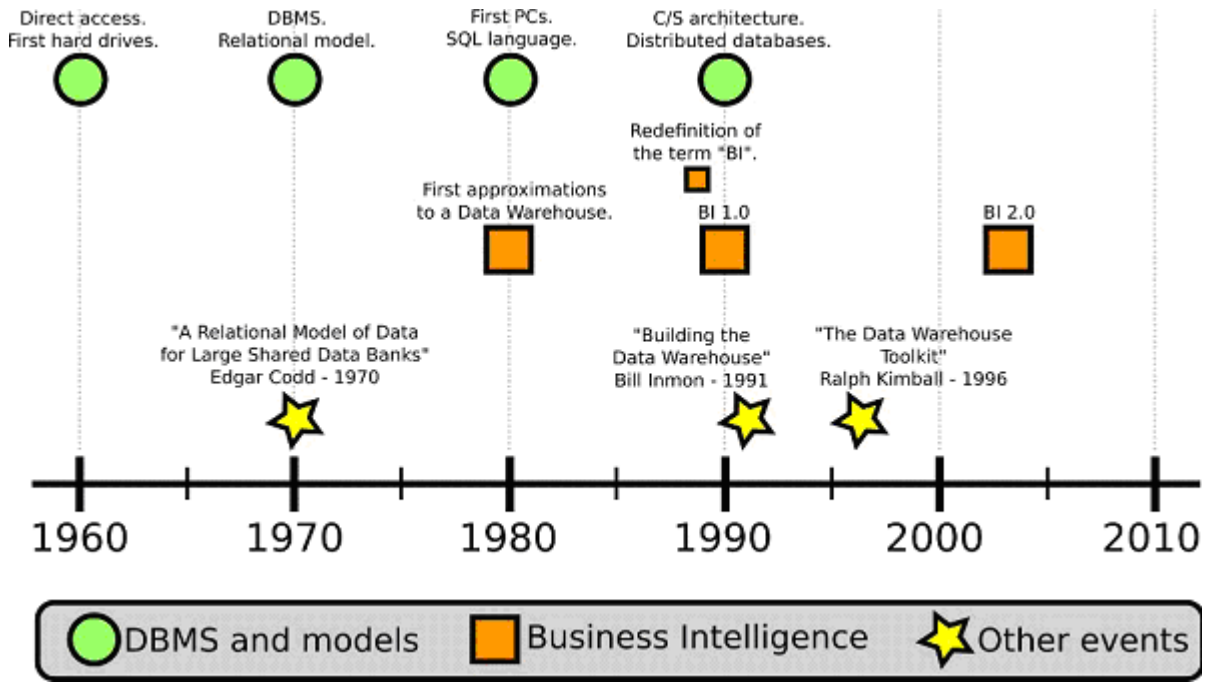

Figure 1: Main Historical Milestones in BI Usability.

software and hardware helped process data to obtain information.

BI Usability in this period: In this time period, the interaction of information systems with users was very precarious. It consisted of consoles that displayed textually a series of options that the user had to select, and generally presented as many screens as options available, and after choosing these options the user obtained printed information summaries and/or specific detailed lists. Based on the definitions presented earlier, there is no doubt that in this period one cannot speak of BI *per se*.

2.2 1970s

BI information systems: In the nineteen-seventies the tendency was marked by the emergence of database management systems (DBMS) and the relational model that was presented in 1969 by Edgar Codd (formally published in 1970). In this decade it is possible to visualise a leap in the evolution of databases, as until then these were mainly based in network models, hierarchies, or simply structured files, whose predominant characteristic was inflexibility and physical relations between entities.

BI Usability in this period: While databases received a great impulse from the relational model, only at the end of the decade were the first versions of systems that supported them created. At the same time, substantial improvements were produced in the responses to requirements of data and information. Interaction with the user improved notably and included interactive text interfaces. This allowed for improvements in the presentation of information per screen due to the possibility of scrolling. Despite all of this, reports continued being static and highly oriented towards transactional information.

2.3 1980s

BI information systems: In the nineteen-eighties, with the appearance on the scene of personal computers, the use of DBMS became more popular, and in 1986 the SQL language was standardised.

The first approximations to the idea of a "Data Warehouse" also appeared, a concept later defined by Bill Inmon and Ralph Kimball in 1992.

In 1989 Howard Dresner redefined the term Business Intelligence, which had first been used in 1958 by Hans P. Luhn.

> “Over the years, BI 1.0 applications saw the solution of important issues such as massive storage, response speed, modularity, etc”

BI Usability in this period: Initially, the providers of the first Data Warehouses place emphasis only on the hardware and in the capacity of their DBMS, and delegated the creation of the GUI to the developers/programmers in each company. In those years, the people in charge of designing and implementing the DW ran into many inconveniences and difficulties, as these people were used to working with transactional/operational systems (OLTP), relational modelling, and, fundamentally, to facing projects of this nature.

This attachment to traditional systems lead to the failure of a high percentage (some say as high as 80%) of the projects of this period, due to not understanding that the development and implementation of a DW cannot be compared to that of an OLTP, and much less is it viable to attempt to adapt methodologies and models, given that tools designed specifically for this new concept should be employed. With respect to interactivity, the improvement was notable. The programming languages allowed for creating friendlier and more user-oriented graphic and textual user interfaces. Reports were more personalisable and parameterisable, and the first information graphics (pie graph, bar graph, etc.) saw the light of day.

Spreadsheets require a special mention, as they radically changed the interaction between the end user and information, granting the possibility of maintaining and interacting with one’s own data. But the possibilities that spreadsheets offered produced as a result piles of redundant and unorganised data, due to their not being designed to manage databases. Later, these piles were dragged along and great efforts were required to process them, organise them, and convert them into a dataset that could be used effectively.

2.4 1990s

BI information systems: In the nineteen-nineties we find organisations/businesses full of PCs, personal DBMSs, spread sheets, etc. that make up a set of heterogeneous data and decentralised and unconnected information. The architecture known as client/server (C/S) allowed for the appearance of a new paradigm in application functioning and communication. DBMSs were one of the categories that most took advantage of this architecture, giving rise to distributed databases, improving intercommunication in organisations/businesses and making databases more consistent and useful. However, there also existed a number of inherited formats (spreadsheets, plaintext files, etc.) for which the contribution of the C/S architecture was not significant, although the idea of standardising processes of data integration was[4].

BI Usability in this period: The diverse publications of Bill Inmon and Ralph Kimball, where they detail how to build and design a DW, as well as defining a conceptual framework for the topic, helped to clarify concepts, and,

[4] *Data Integration* is the set of techniques and sub-processes that extracts data from different origins, manipulates it, integrates it, transforms it, dumps it into another data source, etc.

> “In the first years of the 21st century, the advent of Ajax and the technologies mentioned in the previous paragraph mark what is today known as BI 2.0 ”

more than anything, define a reference point, based on which DW and BI applications would be built.

At this point the first software applications oriented to DW appear, such as: IBM OLAP Server, Cognos, Business Object, SAS, Microstrategy, Oracle, etc. These tools are known as BI 1.0 applications, and their most important or notable features are the following:

- Limited with respect to analysing large volumes of data in an acceptable time, as the physical storages structures were not optimised for this purpose. Nor were there tools to improve DW performance such as: multidimensional clusters, self-maintained aggregate tables, buffers with multidimensional structures, etc.
- Limited with respect to the possible sources of data.
- No general consensus regarding the design of GUIs for administration and navigation.

In summary, flexibility was not generally a virtue of these tools, although they fulfilled the basic tasks inherent to DW, and, more importantly, were DW-oriented.

In the first years of the 21st century, around 2003**,** BI 2.0 comes forth with the development of software dedicated to BI that begins to incorporate new functionalities, characteristics, and technologies, such as: interactivity, web browsers, JS, Ajax, JSON, flexibility, intuitive end-user-oriented GUIs, web services, etc. These dedicated software packages are known as BI suites.

3 Events

The following describes the changes and events that occurred to give shape to BI Usability.

Over the years, BI 1.0 applications saw the solution of important issues such as massive storage, response speed, modularity, etc. This was possible largely due to hardware advances such as parallelism, multi-processing, etc., and more robust software architectures and implementations such as OLEDB, JDBC, middlewares, frameworks, etc. In this context and with the passage of time, the development of BI applications gained experience and matured.

As has happened in many cases in the history of computer science, once issues that limited growth were solved, other issues that had previously been avoided or left aside were given priority. One of these issues was usability. More thought was given to the importance of BI applications being more attractive, intuitive, and easy to use so that users felt comfortable and could take better advantage of their data.

Until this point, the great majority of BI applications were desktop applications, as users were used to the response speed and user interface components (also called widgets) of this kind of application, although some offered a very limited web interface. The limitations of the web interfaces were characterised by the period in which they were developed, "before the Web 2.0" in which pages were loaded completely for each requirement, the bandwidth was consumed by basic requirements, and the widgets were very basic and could not compete with the desktop versions. In short, they were not pleasant or familiar to users.

4 BI Usability Now and Future Tendencies

In recent years, the appearance of Ajax[5], and with it the maturation and/or creation of technologies that allow for representing and transporting data in an efficient and standardised manner, facilitates the creation of attractive and powerful GUIs and the interaction between data and GUIs (JSON, web services, frameworks, JavaScript, flash, CSS, etc.) and has changed the web development paradigm, moving from lightweight clients to fat clients with high capacity for processing, interaction, and visualisation. The applications developed with this technique are executed on the client side (fat client), which only requires of the server what is specifically needed (not the entire page, like before), which can be done asynchronously, ensuring that the user never loses interactivity with the application.

The advent of Ajax and the technologies mentioned in the previous paragraph mark what is today known as BI 2.0, and tips the balance to the side of web development[6]. BI 2.0 applications focus on design and presentation of que-

[5] Ajax, acronym for Asynchronous JavaScript And XML, is a web-development technique for creating interactive applications or RIA (Rich Internet Applications). Source: Wikipedia.

[6] The term *Web 2.0* (2004 – present) is commonly associated with social phenomenon, based on the interaction achieved through different web applications that facilitates sharing information, interoperability, user-centered design, and collaboration in the World Wide Web. Examples of Web 2.0 are web communities, web services, web applications, blogs, wikis, educational platforms, resource-sharing environments, and social networks. Source: Wikipedia.

> “Software dedicated to BI begins to incorporate new functionalities, characteristics, and technologies ”

> “ OSBI (Open Source Business Intelligence) has given SMEs the possibility of implementing BI solutions, which they were previously denied because of the high cost of the tools ”

ries, reports, OLAP analysis, etc. through interactive graphics, flash and JavaScript objects, personalised and highly parameterisable dashboards, etc. All of this places emphasis on the graphic interface and user interactivity.

Currently the development and growth of BI applications and related technologies has not abated. On the contrary, it continually receives new impulses. Some of the developments that could eventually mark new tendencies for change, or even establish new milestones, are:

- Non-SQL databases.
- Cloud Computing.
- SQL Streaming.
- In-Memory OLAP.
- Mobile technologies.
- GPS-based technologies.
- Touch technologies.
- Voice recognition.
- Fast prototyping.

When speaking of developing tendencies, ignoring OSBI (Open Source Business Intelligence) would definitely be an error. In recent years the growth and contribution of the free software and open source community has been large and important, with the resulting achievement that OSBI applications can compete with private applications, and, in some cases, mark tendencies and impulse very innovative ideas. As is habitual, and part of the free software and open source culture, there exist a great number of high-quality fora, blogs, wikis, and social networks (for example Open BI Network[7]), etc. Millions of users around the world use these media as their primary form of making contributions and sharing knowledge, allowing diverse projects to grow continually and become more robust, making this software a viable and safe option.

OSBI has given small and medium-sized businesses the possibility of implementing BI solutions, which they were previously denied because of the high cost of the tools. This has greatly increased the demand for this kind of solution. This increased demand is not only with regard to quantity, but also an increase in the functional and non-functional requirements, such as: better interfaces, greater interactivity, faster responses, information delivery in diverse forms and through diverse channels, etc.

5 Conclusions and Future Work

In conclusion, we can remark that the evolution towards BI 2.0, and particularly the emphasis on usability, has developed in step with the processes of advancement and maturation of the tools and applications, the hardware utilised, user expectations of a current BI application (interactivity, familiarity, intuitive GUI, wizards, etc.) and the OSBI world.

We can assume then that this course will be maintained in similar directions as technology changes. We consider that it is not crazy to imagine a scenario in a not-so-distant future in which a supermarket stock manager approaches a shelf and orders his mobile device through a voice command while pointing the camera so that it takes in the image of various products, and immediately sees on the screen, superimposed on the image of the shelf being captured, products of the same type grouped together shown in semi-transparent colors, with the intensity of the color indicating the stock. The manager could click on one of these areas, which could show an alert icon to indicate the need for immediate restocking.

The example can be enriched by imagining that in addition to these tools, the manager’s mobile screen shows a graphic history of the stock where tendencies and seasonality (among other things) can be rapidly analysed. This information would allow the manager to make a stocking decision and, through his mobile device, use a voice command "make order", which would lead to a screen with a pre-written message for the most convenient supplier (based on distance, price, efficiency, etc.) with the order information and nothing more than the voice command "send" necessary to complete the order. The work of this imaginary supplier would be realised from and with his mobile device.

The earlier example is not a science fiction story or a fantasy. We are already on path towards this development. While still immature, the technologies exist, and are some of those mentioned earlier.

Bibliography

- W. H. Inmon. "Building the Data Warehouse", Wiley, 3ª edición, 2002.
- Ralph Kimball. Articles, <http://www.kimballgroup.com/>.
- C.J. Date. "An Introduction to Database Systems", ISBN 0-321-19784-4.
- G.W. Hansen, J.V. Hansen. "Database Management and Design", Prentice Hall.

[7] *Red Open BI* (Open BI Network), the first Spanish-language social network dedicated to OSBI. <http://www.redopenbi.com/>.

- R. Elmasri, S. Navathe. "Fundamentals of Database Systems", Addison Wesley, 4th edition, 2004.
- R. Camps-Paré. "Introducción a las bases de datos" (Introduction to Dababases). <http://dc364.4shared.com/download/iXU2ujDm/IBD_-_Libro_-_Introduccion_a_l.pdf?tsid=20110826-175802-78968c50>.
- D. Costal-Costa. "Introducción al diseño de bases de datos" (Introduction to Database Design). <http://dc261.4shared.com/download/niAOU9mw/Dolors_Costal_Costa_-_Introduc.pdf?tsid=20110826-175443-c0e2096b >.
- D. Costal-Costa. "El modelo relacional y el álgebra relacional" (Relational Model and Relational Algebra). <http://ocw.uoc.edu/computer-science-technology-and-multimedia/bases-de-datos/bases-de-datos/P06_M2109_02148.pdf >.
- A. Tanenbaum, M. Van Steen. "Distributed systems - Principles and Paradigms", Prentice Hall, 2nd edition (October 12, 2006).

Towards Business Intelligence Maturity

Paul Hawking

Although Business Intelligence is seen as priority by many companies, the level of benefits achieved varies significantly from company to company. Researchers have attempted to relate the types of benefits achieved to the company's Business Intelligence maturity. This paper, adopting a case study methodology, investigates one company's trials and tribulations in regards to their Business Intelligence implementations. The paper documents a number of Business Intelligence best practices and maps these practices to a Business Intelligence Maturity Model.

Keywords: Business Intelligence, Maturity Model, SAP.

1 Introduction

Companies today have come to realise the importance of providing accurate, relevant and timely information — information that allows their organisational personnel to engage in effective decision-making practices. Evans and Wurster [1, pp. 72] in their paper on Information Economics stated that "*… information is the glue that holds business together*". Companies have developed and implemented systems to facilitate the collection, processing and dissemination of information. One such system, Enterprise Resource Planning (ERP), has enabled companies to gain efficiencies in their business processes and associated transactions through the high degree of integration of their company-wide business processes and the standardisation of the associated data [2]. ERP systems are an essential element of the corporate information systems infrastructure, allowing a business to be competitive in today's world, as well as providing foundation for future growth [3].

Accenture interviewed 163 executives from large enterprises around the world to identify how companies were using Enterprise Resource Planning (ERP) systems to improve business performance and the specific practices that resulted in sustained value creation [2]. They found that the implementation of an ERP system resulted in sustained value creation; however some corporations realised far more comparable benefits than others. These benefits were directly related to the actions of management in regards to the development and evolution of their ERP system. Davenport et al [2] identified three major evolutionary stages in regards to benefit realisation facilitated by ERP systems. These were:

Integrate: Unification and standardisation of data and processes. Use the ERP systems to better integrate business processes and the associated organizational units.

Optimise: Align business processes to the overall corporate strategy through the utilisation of embedded "best practice" processes with the ERP system.

Informate: Utilising the information generated by the ERP system to transform work practices. This refers to transforming ERP systems data into context rich information, through Business Intelligence, to support effective decision making.

These evolutionary stages are reflective of a company's ERP systems maturity level. The concept of maturity is often used to describe the advancement of both people and organisations. Implicit is that with increasing maturity there are improvements in quantitative or qualitative capabilities. Accordingly the more mature a company is in regard to their ERP system, the more value they realise from the system.

Author

Paul Hawking is a Senior Lecturer at Victoria University, Australia, and is a leading commentator on ERP systems and specifically SAP solutions. His knowledge is well respected both in industry and academia and is accordingly often required to assist companies with their ERP strategies and understanding SAP solutions. He has presented at leading industry and academic conferences around the world. He is an "Expert Blogger" on the SAP Community Network. He was a past Chairperson and committee member of the SAP Australian User Group (SAUG) for 10 years and was responsible for knowledge transfer. He now designs and advises the SAUG on the content for their events. In 2009 and 2011 Paul was voted by the SAP community as one of the Top Ten Most Influential People in SAP for Australia and New Zealand. He trains and advises academics around the world in regards to ERP systems curricula. Accordingly he was recently awarded "Outstanding Academic 2010" by SAP. Paul is the only academic in the world to achieve SAP Mentor status. Paul is also one Australia's best-selling IT authors having written 12 books which are sold throughout the world. The Microsoft Stable book series has been in publication since 1995. The Microsoft Stable was the first book in Australia to have a supporting website and the book series has had more than $4million in sales to date. Paul's areas of teaching and research include ERP system strategy and implementation, and Business Intelligence. He has published more than 100 research publications. More about the author can be found at <http://www.businessandlaw.vu.edu.au/staff/paulhawking/>. <Paul.Hawking@vu.edu.au>

> “This paper investigates one company’s trials and tribulations in regards to their Business Intelligence implementations”

Harris and Davenport [4] conducted a more extensive follow up study in 2006, involving 450 executives from 370 companies, in an attempt to identify the factors that drove value from ERP systems, as well as how companies used these systems to enhance competitiveness and differentiation. One of the key findings from this study was that improved decision making was the most sought after and realised benefit. While most ERP systems were originally justified on the basis of IT or operational cost savings, senior management’s underlying objective was to improve the quality and transparency of information. Top performing companies were able to achieve this by implementing their ERP systems extensively throughout their organisations across a broad range of business functions. This provided an increased level of integration. They also found that top performing companies were more likely to integrate their business processes across organisational boundaries and with suppliers and customers.

Related to the desired benefit of improved decision making, top performing companies aggressively used information and analytics to improve decision making [4]. These findings are supported by Gartner, a leading business analysis firm, who conducted a worldwide survey of 1,500 Chief Information Officers and identified Business Intelligence as the number one technology priority for companies, followed by ERP systems [5]. Gartner predicted that the worldwide revenue for Business Intelligence software would reach $US10.8 billion in 2011 [6]. The increased expenditure on Business Intelligence reflects the level of impact these systems can potentially have on a company’s performance. IDC, another technology analysis firm, found in a survey of 62 companies that there was on average a 401 percent ROI over a three year period [7]. The Data Warehousing Institute identified in 2005 that the use of Business Intelligence in a number of organisations, such as Hewlett Packard and the US Army, had a significantly positive impact on their performance. Hewlett Packard found in 2004 that, due to their Business Intelligence initiative, the value of worker productivity increased by approximately USD$10.6 million, whilst the company’s reporting costs were reduced by some $8.6million. The US Army found that as a result of their Business Intelligence implementation, 10 trained analysts could complete as much work as 200 traditional analysts. In another example of the value of Business Intelligence, Harrah’s, a major hotel and casino owner in America, found that Business Intelligence contributed to their improve business performance which was associated with their $235 million profit in 2002. Harrah’s spent $10million building a 30 terabyte data warehouse [8] and used Business Intelligence to better understand their customers and their gambling habits [9]. The IDC group collected data from forty three companies in North America and Europe that had implemented a Business Intelligence and found that twenty companies achieved a ROI of less than 100 percent, fifteen achieved an ROI between 101 and 1000 percent, whilst eight achieved an ROI greater than 1000 percent [10].

Although Business Intelligence is seen as a priority for many companies to survive in a competitive market there is uncertainty as to the path to follow. Researchers have identified that companies utilise Business Intelligence in different ways, with varying levels of success. A review of the literature indicates that companies often fail to realise expected benefits of Business Intelligence and sometimes consider the project to be a failure in itself [11][[12][13][14][15]. Gartner predicted that more than half of the Global 2000 enterprises would fail to realise the capabilities of Business Intelligence and would lose market share to the companies that did [16]. A survey conducted by Cutter Consortium Report [17] in 142 companies found that 41 percent of the respondents had experienced at least one Business Intelligence project failure and only 15 percent of respondents believed that their Business Intelligence initiative was a major success. Moss and Atre [18] indicated that 60% of Business Intelligence projects failed due to poor planning, poor project management, undelivered business requirements, or, of those that were delivered, many were of poor quality. A number of authors believe that in many Business Intelligence projects the information that is generated is inaccurate or irrelevant to the user’s needs or indeed, delivered too late to be useful [19].

These researchers have attempted to map Business Intelligence usage and best practices to provide a roadmap for companies to move forward and maximise the benefits of their Business Intelligence initiatives. One approach for this roadmap has been the development of Business Intelligence Maturity Models [20][21][22][23][24][25][26]. The purpose of these models is to provide companies with a

> “Companies have developed and implemented systems to facilitate the collection, processing and dissemination of information, such as Enterprise Resource Planning (ERP)”

roadmap to improve the management of their corporate data, as well as to maximise the benefits obtained from Business Intelligence. The Business Intelligence Maturity Models identify practices incorporating different stages which are associated with a company's Business Intelligence progress and growth. Although there are many Business Intelligence Maturity Models, they each differ in the practices and stages characterising different levels of maturity.

2 ASUG Business Intelligence Maturity Model

The Americas SAP User Group (ASUG) is the largest SAP user group in the world with more than 85,000 members from 4,000 companies [27]. SAP is the market share leader in both ERP systems and Business Intelligence [28]. ASUG developed a series of benchmarking studies to assist its members to better understand the implementation and usage of ERP systems and associated solutions such as Business Intelligence. In 2007, ASUG in conjunction with SAP developed a Business Intelligence benchmarking initiative and has had more than 100 companies participate in the initiative [29]. A website was developed to capture the benchmarking information and a series of presentations was conducted to introduce customers to the initiative. The key questions which the study was intended to answer were:

- How do companies leverage Business Intelligence to drive business performance?
- For which business process is Business Intelligence most critical?
- What are the key performance indicators of an effective Business Intelligence environment?
- How much do top performing companies invest in Business Intelligence?
- What are the best practices that companies can adopt to drive effectiveness and efficiency of their Business Intelligence environment? [29]

Key metrics were designed to capture information to answer these questions. The web site was designed to capture enough information from different company's Business Intelligence experiences to enable relevant comparisons. These details were compared to details from other companies as well as industry standards, allowing a range of Business Intelligence benchmarks to be created. Part of the benchmarking derivation process was the mapping of companies to a maturity model. The ASUG Business Intelligence Maturity Model (Table 1) allows Business Intelligence maturity to be classified as per practices related to Application Architecture, Standards and Processes, Governance, and Information and Analytics. Each of these practices is made up of a number of stages which describe different aspects of Business Intelligence maturity.

It would be expected to find many companies in the early levels of Business Intelligence maturity and therefore provide verification for the practices and associated stages. But are the higher levels of maturity reflective of Business Intelligence best practices? Each year Gartner identifies companies for their Business Intelligence Awards of Excellence. It would be reasonable to expect that a company which achieved such an award would be very mature as per the model. This research adopts a case study approach to investigate the Business Intelligence operations of a re-

Stage	1 Information Dictatorship	2 Information Anarchy	3 Information Democracy	4 Information Collaboration
Information and Analytics	Requirements are driven from a limited executive group	KPI's and analytics are identified, but not well used	KPI's and analytics are identified and effectively used	KPI's and analytics are used to manage the full value chain
Governance	IT driven BI	Business driven BI evolving	BI Competency Centre developing	Enterprise wide BI governance with business leadership
Standards and Processes	Do not exist or are not uniform	Evolving effort to formalise	Exist but are not uniform	Uniform, followed and audited
Application Architecture	BI "silos" for each business unit	Some shared BI applications	Consolidating and upgrading	Robust and flexible BI architecture

Table 1: The ASUG Business Intelligence Maturity Model [24].

> “Three major evolutionary stages have been identified in regards to benefit realisation facilitated by ERP systems: optimise, integrate, and informate”

cent Business Intelligence Award of Excellence recipient, (alias CompPack). The Business Intelligence operations are then mapped to the ASUG Business Intelligence Maturity Model to investigate its applicability.

3 Case Study

A case study research methodology was used to examine CompPack and its use of Business Intelligence to support their overall business strategy. The case study focused on a large company involved in the packaging and processing industry. The data collection process included interviews of key personnel, examination of existing documentation and analysis of internal documentation. Yin [30] suggests that a single, in-depth case study is an appropriate research approach under a number of conditions, one of which being that it is a critical case whereby it meets all the necessary conditions for testing a theory.

CompPack is a global food packaging and processing company which has been established since 1929. This private company has 20,000 employees, 50 factories and sales operations in 150 countries. In 2008, CompPack produced 141 billion packages worldwide resulting in total sales of Euro 8,610million.

CompPack decided to implement a SAP ERP system in 1994 to support their business. Similar to many other companies, CompPack‘s ERP system implementation was not as successful as they would have liked. In 1999, CompPack was faced with a number of issues. There were the issues of the impending Y2K and the impact this would have on the company, especially as some of the legacy systems were almost twenty years old. In addition CompPack’s business had grown globally and the ERP system needed to support these new markets and associated operations. It was decided to undertake a Process Globalisation Project supported by SAP solutions.

The SAP implementation which included the implementation of a data warehouse adopted a phased approach based on geographical locations. The first two phases involved geographic locations associated with CompPack’s smaller markets, thus minimising the risks. The third phase involved implementing SAP in Germany and United States, which represented the majority of CompPack’s markets and thus the highest risk. This implementation was not without its problems. The project took 12 months instead of the planned 6 months and incurred a 300 percent budget overrun.

The implementation of the data warehouse was a relatively small component of the overall SAP implementation. The project overruns limited the scope of the of the data warehouse implementation. The data warehouse was designed to be a large repository of business data based on the premise that if data was collected and stored in one location, then the business users would access it for their business needs. This expectation did not occur. A major reason for this was the lack of performance associated with making the data available to the business users. The performance issues were related to the technical design and infrastructure. Data was extracted from the ERP system into the centralised data warehouse. The data was then aggregated and extracted into geographically based data warehouses (data marts) an, in some cases, the data was further extracted to power users’ personal computers. This series of data extractions resulted in delays in performance in delivering relevant data to the intended users. Accordingly there was a lack of confidence in the centralised data warehouse solution.

In 2005, the staff responsible for the data warehouse realised that, after spending 20 million Euros, the current system was not providing the expected benefits and so arranged a meeting with the Chief Financial Officer (CFO) to discuss the various options. The CFO agreed that there needed to be a change of direction and, in 2006, the data warehouse project was stopped and a new Business Intelligence initiative was commenced. The project was referred to as "Business Warehouse" to differentiate it from the previous project

It was decided to reduce the complexity of the current Business Intelligence environment and that the new project would standardise the BI infrastructure across CompPack to SAP’s Business Information Warehouse (SAP BW), including Business Explorer (Bex) web component for the presentation of reports. This reduced the number of extractions required as per the previous implementation and thus improved overall performance in the providing business data to the users.

The Business Warehouse project had two major milestones. The first was to replace a legacy financial consolidation system by getting the global legal financial accounting data into the SAP BW system and ensure its correctness. The second milestone was associated with loading the management accounting data into the BW system as well ensuring that the correct data was available to report on the key performance indicators (KPI’s) of CompPack’s core business processes. This meant that CompPack had evolved, from a having legal financial accounting view of the company, to a management view of the company involving budgets and core business process performance. This availability of key data, via the BW system, resulted in greater support and acceptance by business users. The Business Intelligence team started to develop standardised processes to enable the provision of more and more key

Performance =

Process X People X Tools

Figure 1: Business Performance at CompPack.

information to support the business.

SAP, in conjunction with hardware partners IBM and HP, developed a "bolt on" infrastructure solution to improve the performance of reporting. The Business Intelligence Accelerator (BIA), utilising blade computing technology, has been reported to improve reporting by up to one thousand times faster, according to Lewis. In early 2009, CompPack implemented the BIA to improve their reporting performance. The reporting response time was reduced from an average of twenty seconds down to five seconds. The availability of financial and management data, in conjunction with improved reporting performance, resulted in greater support and acceptance of the BW system by the business users.

As part of the Business Warehouse project, CompPack considered there were three important phases to their Business Intelligence journey. The first phase involved getting the necessary infrastructure and data in place to provide some quick wins, while at the same time providing a foundation for future development. Prior to the implementation of the Business Warehouse project CompPack had a fragmented corporate reporting applications environment. The second phase involved the governance of Business Intelligence in terms of the processes related to the collecting requirements to the development of reports. A standardised reporting template was developed which included charts, data tables, filters and the ability to change the dimensions for analysis. All reports were developed based on this template and thus, once a user was familiar with the functionality and navigation of one report, they could then apply this knowledge to any other report. The only training that was required was in relation to the business content of the report and its applicability. The governance standardisation enabled a best practice approach to ensure a successful Business Intelligence solution. The final phase was to build upon the foundation laid down by the first two phases to extend the coverage and usage of Business Intelligence to support management and the business.

A major factor of the Business Intelligence initiative's success was due to the agreement by senior management as to the role of Business Intelligence within CompPack. There was agreement that, to improve business performance, there needed to be three things in place. There needed to be the right business processes, people needed to be trained how to execute these business processes and, finally, the correct tools needed to be available to support the people and processes (Figure 1). Business Intelligence was considered to be an essential tool to monitor processes and thus measure performance. CompPack developed a strategy map and balanced scorecard, including relevant KPI's, to implement and monitor their strategy.

The monitoring of business processes through the associated KPI's was integral to the company's performance and this was the main priority for Business Intelligence. Another business priority for Business Intelligence was the need for a single version of truth about the business. This included consistent facts about customers, products, suppliers, past performance and future forecasts. CompPack's Process Globalisation Project was the single largest investment in the company's history and Business Intelligence enabled the company to realise many of the benefits from this investment.

As part of the Business Warehouse project, CompPack consulted with Gartner in an attempt to identify "best practice". One recommendation was the establishment of a Business Intelligence Competency Centre (BICC). A BICC is responsible for developing the overall strategic plan and priorities for Business Intelligence. It defines the requirements (including data quality and governance) and helps the organisation to interpret and apply the insight to business decisions [31]. CompPack considered that a BICC was essential if it was to achieve an enterprise view of the data and reporting requirements.

To fully capture the company's requirements CompPack's BICC was comprised of two structures. The first structure consisted of:

Business Information Management (BIM): This consisted of 5 full time senior business analysts who had a good understanding of the business and the capabilities of Business Intelligence.

Global Information Management (GIM): This project team consisted of between 15 to 25 people and provided the technical Business Intelligence expertise. The BIM and GIM worked closely together with common goals.

Global Information Management Service Delivery Team (GIM SDT): This group involved approximately 12 people and were responsible for ensuring the availability and an ongoing support for reports once they were developed.

Global Process Owners/ Global Process Drivers (GPO/ GPD): This group were responsible for key business processes. CompPack decided that these people were the only people who were allowed to request IT related projects. This resulted in IT having a very focussed business role.

The other structure, which was referred to as the "Ex-

"Top performing companies aggressively used information and analytics to improve decision making"

Measure	Score	Comment
Global Enterprise-wide Adoption – the ultimate measure of BI success – % of employees as active BI users	> 10%	More than 10% of employees are active users; expect to reach 15% in 2009. More than 30000 navigations per day. 20% of employees are registered users.
% coverage in BI of business processes and business performance measurements Single source of truth across borders, processes, businesses	100%	Business performance measurements are available for all business processes and all business units. Expanding coverage within processes and units. Used in all Markets and in the centre.
Response time	5 seconds	Worldwide: all management reports in 15 seconds or less, average navigation step below 5 seconds
Reliability, Consistency & Quality	7AM	All managers have fresh data at 7AM their time worldwide. Information is correct and broadening. Adoption makes sure it stays correct.
Easy to use – low training cost	High user adoption	Information portal based on geography, business roles and business processes; standard layouts make it easy to understand and use
Enables next steps – new major business information initiatives	Global Information Projects	Successful major new information projects – brand information back to our customers, worldwide alignment on Sales Forecasting

Table 2: Business Intelligence Value Scorecard.

tended BICC" consisted of the MIS coordinator from each of the business areas that utilise Business Intelligence. Their role was to act as change agents and encourage the adoption and use of the Business Intelligence solution.

The BICC is overseen by a steering committee made up of senior management and their ongoing support is considered essential to the success of the Business Intelligence initiative. A priority of the BICC is not just to gather requirements and develop reports but also the deployment of those reports and the realisation of their value. The process of gathering requirements, developing reports, deployment and report value realisation has been documented to ensure that the process is standardised, repeatable and clearly understood across the company. This has enabled the process to be refined and improved. A timeline for the report development and deployment process was developed and publicised. This facilitated business areas planning and scheduling their reporting requests. Reports are rolled out quarterly.

CompPack's approach to Business Intelligence has enabled them to gain a high level of success in relation to their Business Intelligence initiative. In December 2008 they had approximately 1800 active users representing about 9% of the employees. By June 2009, the number of active users had increased to 2,600 (12.5%). CompPack believes that this level of usage could not be achieved unless the users perceived the Business Intelligence system to be of value.

To ensure that CompPack's approach to Business Intelligence is best practice, they developed a "Business Intelligence Effectiveness Scorecard". This scorecard consists of a number of assessable components including;

Business Case and Vision: 1) single source of truth; 2) business analysis across borders, processes, businesses; 3) analysts move from data gathering to real business analysis; 4) reduce total reporting cost.

Executive Support: CFO provides visible public support.

Alignment to Business Strategy and Business Processes: Only Global Business Process Owners can request Business Intelligence or CPM projects.

Alignment and Working Practices, Business and IT: Business Transformation Process aligns strategy, process

> "Researchers have attempted to map BI usage and best practices to provide a roadmap for companies to move forward and maximise the benefits of their BI initiatives"

> “ The ASUG Business Intelligence Maturity Model allows BI maturity to be classified as per the most relevant practices ”

and organisation. Business owns scope prioritisation and outcomes.

Extended BICC: Central team contains both business and technical expertise. Network from the centre Business Transformation Officers and Market MIS Coordinators provide the link to adoption.

Predictability – Robust and Effective Delivery Methodology: Compliance to IT Project and Service processes as a subset of Business Transformation process.

CompPack believe that their Business Intelligence approach has satisfied the above criteria. However the above scorecard only reinforces that the correct approach has been implemented. A further scorecard, "Business Intelligence Value Scorecard" was developed to quantify the Business Intelligence impact on the business. This scorecard including measures is displayed in Table 2.

CompPack has noticed that, due to their approach to Business Intelligence and the value generated, different areas of the business are placing greater demands on the Business Intelligence group for new initiatives. This increased demand for Business Intelligence is reflected by the last measure in the above scorecard.

Business Intelligence has enabled CompPack to refine their business processes as they move towards a business transformation. Business Intelligence is used to gauge the performance of business processes and thus essential to understanding the impact of business process redesign. Since the introduction of Business Intelligence, CompPack has seen significant improvements in many of their core business processes. For example CompPack focused on reducing the time between the ordering and implementation of their packaging equipment at a customer’s site. Through the revision and refinement of the associated processes they were able to reduce this time from 140 days down to 47 days. The process of taking a customer’s packaging design and manufacturing it was reduced from 15 days to 5 days. Accordingly Business Intelligence is considered essential to business sustainability and growth at CompPack.

4 Business Intelligence Maturity Model Applicability

CompPack’s Business Intelligence implementation and usage would make it be considered a very mature company as per the ASUG Business Intelligence Maturity Model. KPI’s and analytics are used extensively to manage the entire business. The BICC has enabled the company to develop enterprise wide governance and Business Intelligence leadership while at the same time implementing standardised processes and standards to support the Business Intelligence initiative. This standardisation also applies to their Business Intelligence architecture. These Business Intelligence practices are aligned with the highest level of maturity in the ASUG model, Information Collaboration. This level of maturity is further supported by CompPack achieving a Gartner Business Intelligence Award of Excellence in 2009. Table 3 classifies CompPack’s BI practices as per the Information Collaboration stage of the ASUG Business Intelligence Maturity Model.

CompPack have realised that it is important to measure Business Intelligence from two different perspectives. Firstly, and the most common reason for measuring Business Intelligence is to prove its value as an investment. They have been able to quantify the tangible benefits Business Intelligence has provided the company. The second perspective is to measure Business Intelligence activities for the purpose of monitoring and improving the Business Intelligence process. The development of a BICC and a number of scorecards has enabled CompPack to adopt an enterprise wide approach to business Intelligence. Throughout the case study research CompPack continually emphasised that *"…it is not about Business Intelligence but about corporate performance management and Business Intelligence is only one part of the formula*".

5 Conclusion

The ASUG Business Intelligence Maturity Model attempts to classify Business Intelligence usage and best practices into different stages. As Business Intelligence technology evolves and permeates all aspects of business it would be expected that these stages would also evolve to include different practices. The application of the model to Business Intelligence Award of Excellence winner demonstrates the suitability and applicability of the model at the more mature stages. This paper provides an example of a company’s Business Intelligence journey and what could be considered Business Intelligence best practice.

> “ CompPack, a large multinational company in the food packaging and processing industry, decided to implement a SAP ERP system in 1994 to support their business ”

> “A major factor of the BI initiative’s success in CompPack was due to the agreement by senior management as to the role of BI”

Stage	4 Information Collaboration	CompPack Practices
Information and Analytics	KPI’s and analytics are used to manage the full value chain	• Implementation of Strategy map and balanced Scorecard • Globalisation Process Project
Governance	Enterprise wide BI governance with business leadership	• Establishment of an enterprise wide Business Intelligence Competency Centre supported and promoted by senior management.
Standards and processes	Uniform, followed and audited	• The implementation of the Business Intelligence Competency Centre. • Introduction of BI Effectiveness Scorecard
Application Architecture	Robust and flexible BI architecture	• Business Intelligence Accelerator • SAP Business Intelligence • Business Explorer web reporting

Table 3: The ASUG Business Intelligence Maturity Model and CompPack [24].

References

[1] P. Evans and T. Wurster. "Strategy and the new economics of information". Harvard Business Review, September-October, 70-82, 1997.

[2] T. Davenport, J. Harris, and S. Cantrell, S. "The Return of Enterprise Solutions: The Director’s Cut", Accenture, 2003.

[3] D.C. Chou, H.B. Tripuramallu, and A.Y. Chou. "BI and ERP integration". Information Management and Computer Security, 13(5), 340-349, 2005.

[4] J. Harris and T. Davenport. "New Growth From Enterprise Systems". Accenture, 2006.

[5] Gartner. 2008 Gartner Executive Programs CIO Survey, 2008. <http:// www.Gartner.com> [accessed June 2008].

[6] Gartner. Gartner Forecasts Global Business Intelligence Market to Grow 9.7 Percent in 2011, 2009. <http:// www.Gartner.com> [accessed May 2011].

[7] IDC. "Financial Impact of Data Warehousing". International Data Corporation, 2006.

[8] D. Lyons. "Too much information". Forbes, 110-115, 2004.

[9] S.Williams and N. Williams. "The Profit Impact of Business Intelligence". Morgan Kaufmann, New York, 2006

[10] H. Morris. "The Financial Impact of Business Analytics: Build vs. Buy". DM Review (13:1), 40-41, 2003.

[11] T. Chenoweth, K. Corral, H. and Demirkan. "Seven key interventions for data warehouse success". Communications of the ACM, 49(1), 114-119, 2006.

[12] H.-G. Hwang, C.-Y. Ku, D.C. Yen, and C.C. Cheng. "Critical factors influencing the adoption of data warehouse technology: a study of the banking industry in Taiwan". Decision Support Systems, 37(1): 1-21, 2004.

[13] L.K. Johnson. "Strategies for Data Warehousing". MIT Sloan Management Review, (Spring), 45(3), 9, 2004.

[14] S. Atre. "The Top 10 Critical Challenges For Business Intelligence Success". C. C. Publishing: 1-8, 2003. <http://www.computerworld.com/computerworld/records/images/pdf/BusIntellWPonline.pdf> [accessed June 2007].

[15] S. Adelman and L.T. Moss. Data Warehouse Project Management. Addison Wesley, Boston, 2002.
[16] H.J. Dresner, F. Buytendijk, A. Linden, T. Friedman, K.H. Strange, M. Knox, and M. Camm. "The Business Intelligence Competency Center: An Essential Business Strategy". Gartner Research, ID R-15-2248, Stamford, 2002.
[17] Cutter Consortium Report. "Cutter Consortium Report on Corporate Use of BI and Data Warehousing Technologies", 2003. <http://www.dmreview.com/article_sub.cfm?articleid=6437> [accessed August 2008].
[18] T.L. Moss and S. Atre. "Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-support Applications". Addison-Wesley, Boston, 2003.
[19] D.P. Ballou and G.K. Tayi. "Enhancing data quality in data warehouse environments". Communications of the ACM, 42(1), 73-78, 1999.
[20] H. Watson, T. Ariyachandra, and R.J. Matyska., "Data Warehousing Stages Of Growth". Information Systems Management, 18(3): 41-50, 2001.
[21] K. McDonald. "Is SAP the Right Infrastructure for your Enterprise Analytics". Presentation at American SAP User Group Conference, April, Atlanta, 2004.
[22] P. Den Hamer. "De organisatie van Business Intelligence". Den Haag: Academic Service. Cited in C. Hindriks (2007). Towards chain wide Business Intelligence, University of Twente, 2005.
[23] W. Eckerson. Performance Dashboards: Measuring, Monitoring, and Managing Your Business'. Wiley-Interscience, NYC, 2007.
[24] ASUG. "ASUG/SAP Benchmarking Initiative: Business Intelligence/Analytics". presentation at American SAP User Group Conference, May, Atlanta, 2007.
[25] Hewlett-Packard. "The HP Business Intelligence Maturity Model: describing the BI journey". Hewlett-Packard, 2007.
[26] B. Hostmann. "BI Competency Centres: Bringing Intelligence to the Business". Business Performance Management, November 2007.
[27] ASUG 2008 Annual Report. <http://www.asug.com> [accessed August 2009].
[28] SAP. "SAP Named Worldwide Market Share Leader in Business Intelligence, Analytics, Performance Management Software by Top Industry Analyst Firm", 2011. <http://www.sap.com/news-reader/index.epx?articleID=15099> [accessed June 2011].
[29] ASUG. "ASUG and SAP Benchmarking and Best Practices", 2008. <http:// www.asug.com> [accessed June 2009].
[30] R. Yin. Case Study Research, Design and Methods, 2nd edition. Newbury Park, Sage Publications, 1994.
[31] Gartner. BI Competency Centers: From 'Should We?' to 'How Should We? Presentation at the Gartner Symposium, ITExpo, Sydney, 2006.

Business Intelligence Solutions: Choosing the Best solution for your Organization

Mahmoud Alnahlawi

With the increased awareness regarding the importance of Business Intelligence (BI), a wide array of platforms and tools have come to existence to answer companies demand. Choosing the right tools depends on the specific needs and goals that an organization is trying to optimize, along with the nature of its data and analysis requirements. In this paper different aspects and goals of the business intelligence architecture are described. The way how the Architecture Trade-off Alternative Method (ATAM) can be used to evaluate different vendors and platforms is presented too.

Keywords: Architecture, Business Intelligence, Data Warehousing, LATAM, Systems Design, Software Evaluation.

"The world's total production of information amounts to about 250 megabytes for each man, woman, and child on earth. It is clear that we are all drowning in a sea of information. The challenge is to learn to swim in that sea rather than drown in it."
Peter Lyman and Hal R. Varian

Author

Mahmoud Alnahlawi is an experienced software architect based in Palo Alto, California, USA. He built scalable data systems that range from petabyte-scale data warehouses for offline processing to near realtime data pipelines. His area of expertise include data modeling, real-time document indexing, relational and NoSql databases, distributed processing and cloud data discovery. He has played many critical roles in Fortune-500 companies as well as small startups to solve data problems related to web click stream analysis, sponsored and display advertising as well as security log management. <alnahlawi@gmail.com>

1 Introduction

Both the amount of data and its processing are growing at a very fast pace. More so, academia and industry are continuously trying to find out new ways to harness the power of the data and use it to derive meaning insights that drive and direct innovation in different areas. The uptrend of both phenomena have triggered a proliferation of platforms and tools that aim to solve the problem of storing, processing and presenting the data to facilitate the innovation.

Although well accepted architectures of building a robust data warehousing and business intelligence solution have been around for a long time, having vast solutions on the market requires diligent, systematic and thorough analysis of existing products along with their respective tradeoffs.

The first main question to be asked when looking into building a new business intelligence project is: Should the platforms and tools be built it in house or should off-the-shelf products be used? Many organizations underestimate what it takes to build an end-to-end Business Intelligence Solution. It seems as if building everything in house will always be cheaper than adopting external ones. They end up wasting many cycles of time and resources or even worse, cancel the project. A company needs to clearly articulate the gaps in the existing products that prevent it from adopting them, along with detailed plans of how the gaps are going to be closed.

One of the major dimensions is the budget that can be allocated for the project and the cost of the end-to-end solution. Prices of products vary widely as well as the pricing model. Some companies charge per license seat, others per CPU. Some have unlimited usage for an annual fee or a one time payment.

The other major factor is the reporting requirements that the solution needs to address. Is canned reporting sufficient or do analysts need ad-hoc and interactive reporting slicing and dicing the data by different dimensions? What is the skill set of the users of the product? Are they proficient in SQL and Excel or do they need easy and intuitive user interfaces to work with?

The size and type of data to be analyzed also plays a big role in determining the best option. If the data is very large, it is crucial to pick a tool that can support parallel execution for both Extract-Transform-Load (ETL) and reporting. A slow performing system discourages users and results in overall failure of the project. Scalability is also very important. Picking a solution that not only meets the organization's current needs but also can handle projected data growth and increase usage in a timely manner.

In this paper, a high level overview of different areas needed for building a business intelligence solution is first given, followed by an overview of the Architecture Tradeoff Analysis Method (ATAM) developed by the Software En-

> “ Both the amount of data and its processing are growing at a very fast pace ”

gineering Institute at the Carnegie Mellon University, USA. Next, important quality attributes needed for building a solid business intelligence systems are given. Lastly a sample Utility Tree for a business intelligence system is created to help organizations make the proper platform or vendor decision that meets their goals and requirements.

2 Background

The background section of this paper is broken up into two sub-sections. The first describes an over all architecture for building data warehousing and business intelligence solutions; and the second focuses on describing ATAM.

2.1 Data Warehousing and Business Intelligence Architecture

Ralph Kimball is one of the original architects of data warehousing and business intelligence systems. He described a high level data warehousing and business intelligence architecture which contains three main areas: Operational Source Systems, Data Staging Area, and the Data Presentation Area. Below is an overview of each area.

2.1.1 Operational Source Systems

Rather than being a part of the warehousing and business intelligence system, an operational source system is the input to the warehouse. Often in the initial stages of the design and requirement gathering phases go into assessing and understanding source systems. Two major activities are needed with respect to source systems. The first activity is performing gap analysis on the source system to determine whether all requirements can be filled by it. The second is detail data profiling which is needed to detect possible data quality issue and give requirements to the detail ETL design.

2.1.2 Data Staging Area

The data staging area is where most of the development time and QA stages are typically spent. The staging area is where the ETL (i.e. extract, transform, and load) is completed. The best analogy to the staging area is a closed kitchen where only experienced chefs are allowed to enter. They prepare and cook the data before it is served to the dining area- presentation area. In the staging area, data is first extracted from the source systems. Source systems can have a variety of interfaces such as data hosted in relational database systems or log files generated by a web server. The extraction process may also involve a complex collection system which can collect data from thousands of machines in geographically distributed locations. Once the data is finally in the staging area, different types of transformations are applied to it. Such transformations include cleansing, where erroneous data is detected and possibly corrected; integration, in which desperate data sources are joined together to give an end to end perspective on the data; and aggregation, where the data is summarized and grouped in different ways to facilitate analysis. The staging area is usually a very complex and dynamic environment. It is imperative that it is available, reliable and operable.

2.1.3 Data Presentation Area

After the data has been cleansed, transformed and integrated, it is finally loaded into the data presentation area. The presentation area is analogous to the dining area in the restaurant metaphor that was used above. Data in the presentation area will be accessed in many different ways and by different types of users. A good presentation area may and often does contain many sub-systems that are specialized for different types of users. It services product managers and business owners interested in the Key Performance Indicators (KPIs) of their products. It is where the scientists go to mine the data for interesting and insightful trends. The presentation area may need to handle requests that are expected to return in less than a minute, to queries that can run for hours processing very large and detailed data.

Additionally, different from the staging area, the presentation area is not a closed environment, The presentation area needs to be able to handle different access roles and ensures that the data hosted within is protected and only allowed users can get access to protected information.

The presentation area requires other types of data management as well such as Retention Management, Discovery, Online Analytical Processing (OLAP), Reporting and Visualization tools.

2.2 ATAM – Architecture Tradeoff Analysis Method

The ATAM method shows how well an alternative satisfies different business requirements and how business requirements impact each other. The ATAM method requires a well documented component level architecture along with well defined business requirements.

Business requirements are represented in terms of Qual-

> “ There is a a proliferation of platforms and tools that aim to solve the problem of storing, processing and presenting the data to facilitate the innovation ”

> “Should the platforms and tools be built it in house or should off-the-shelf products be used?”

ity Attributes – things that stakeholders of the product care most about. Quality Attributes are represented in what is called an Utility Tree. An Utility Tree is defined as a hierarchical, tree structure with general broad categories at the first level. Each category is then divided into sub-categories. At the lowest level of the tree are the scenarios. Scenarios represent specific requirements of the architecture that has to be detailed, unambiguous and measurable. Describing Quality Attributes in terms of scenarios is essential since they eliminate ambiguity and give concrete requirements for the development team and test cases of the quality assurance team. A scenario consists of six parts: 1. source, 2. stimulus, 3. environment, 4. artifact, 5. response and 6. response measure. Source describes who generated the stimulus, whether it is System A, User X or bug 123. Stimulus is an event or a condition that needs to be handled by the system. Environment is the state of the system during which the stimulus takes place. Artifact is the part of the system that was impacted by the stimulus. Response is desired behavior of the system after or during the stimulus. And finally, response measure is way to test that the desired response actually took place. For example, consider the following scenario under the performance Quality Attribute of a reporting system:

- Source of stimulus = Users
- Stimulus = 100 users login simultaneously
- Environment = new data is being loaded into the reporting system database
- Artifact = read load of the database is increased
- Response = system should handle load gracefully
- Response Measure: Each report should finish and data returned to the requester within five minutes of report request time.

Once the Utility Tree is constructed, the scenarios are prioritized by the architect using feedback from all stakeholders of the project. It is important to include users, developers, testers and system operators in the process of assigning priorities to ensure that all viewpoints are represented. Once prioritization is complete, the architect then documents how well each alternative, such as Vendor A, handles each scenario, such as automatic error recovery or failover. After this is complete, each scenario will have a priority a score for each alternative.

Once the Quality Attributes have been defined, a mapping between the different scenarios and the different architectural decision or alternative is constructed. Essentially, each architectural decision, such as using platform A, is given a a rank for how well it handles the scenario.

Once the prioritization and the assessment phase is done, analysis of the architecture is ready to take place. The analysis phase identifies for each scenario and each alternative a set of sensitivity points, tradeoffs points, risks and non-risks. Sensitivity points are Quality Attributes or scenarios that are impacted by choosing one alternative over another. Tradeoff points that are doing well on by an alternative implies doing poor on another scenario. Risks are tradeoff points that may result in an undesirable behavior based on the scenarios and non-risks are tradeoff points that are deemed safe with respect to scenarios.

Detailed documentation and examples of the Architecture Tradeoff Analysis Method ATAM, as well as alternatives to it such as Cost Benefit Analysis Method(CBAM) and Microsoft’s Lightweight Architecture Alternative Assessment Method (LAAAM) can be found online at the Software Engineering Institute website, <http:// www.sei.cmu.edu>.

3 Choosing the Proper Solution for your Organization

In this section we will go some of the main quality attributes that should be considered to shape your decision and guide you towards the right solution for your organization. You should take the quality attributes listed in this section and come up with sub-categories and scenarios that are applicable to your organization’s need and requirements. Once that is done, we give a sample Utility Tree that can be used for evaluating how well different vendors meet the different scenarios and aid in making the optimal decision.

3.1 Availability

Availability determines how system deals with failures and has a big impact on the architecture of the system and it’s associated cost and time. The first and largest availability question is what count of Business Continuity Plan (BCP) does your warehouse require. Business continuity planning requirements specify how the system should react in the

> “The size and type of data to be analyzed also plays a big role in determining the best option”

event of large outages such as an earthquake destroying an entire data center. The requirements, and therefore the architecture, vary by organization. Some require ZERO downtime especially if the business intelligence solution is used by production customer facing systems. Other analytical or internal facing systems have more relaxed recovery requirements than can be multiple days. Different vendors have built BCP solutions than range from automatic backup to tape to real-time replication of data over TCP networks.

Typically availability is described by nines – 90, 99, 99,9 etc. Companies with high availability requirements target five nine availability goals, meaning the system has to be up and functional 99.999% of the time allowing for only 5.26 minutes of downtime per year. Going to six nines, allows for only 31.5 seconds of downtime per year! For a system to obtain such high availability numbers, there are minimum requirements that need to be met. The solution should have no single points of failure (SPOF) which is a component whose failure result in the failure of the entire system. There should be ability to provide live updates – updates while the system is up and running. The system needs to be fault tolerant, which is the ability of the system to operate gracefully, with possible degradation of service but not loss of it, in the event of failure of one ore more of its components.

3.2 Scalability

Scalability is one of the major differentiators amongst vendors (along with Performance). Scalability measures the ability of the system to handle large amount of work without performance degradation. Scalability can be defined for sub-systems of the business intelligence architecture and can have different measure of requirements. For example, the presentation tools and OLAP solution need to scale for a certain number of concurrent users. It also needs to scale for certain number of predefined reports or aggregations. The storage sub-system of the presentation area need to scale for a given number of bytes, certain number of rows per table and certain number of concurrent queries.

Different ETL and data storage and processing platforms have different solutions for scalability. Its important to assess how the vendor techniques meet the requirements of your organization. There are two different techniques for handling scalability, vertical or horizontal scale. Vertical scale is the ability to add more resources to a single machine such as increasing memory or CPU. Horizontal scale means adding more machines to a distributed system. Horizontal scale allows for using commodity and cheaper machines instead of specialized and expensive ones. Horizontally scalable systems require shared storage with high throughput access to the data. Tradeoffs between horizontal and vertical scaling models involve high-cost-of scale for hardware vs. high number of machines which might be hardware to manage and operate. Also, larger number of machines consume more power and more data center real estate.

Additionally, data bases have a different techniques that facilitate both vertical and horizontal scaling. A common technique that is supported by almost all vendors is partitioning. Vendors may differ however by the maximum number of allowed partitions. Also, they may offer different partitioning schemes such as range or hash partitioning. Databases have also different threading implementations that allow them to handle vertical scale differently.

3.3 Performance

Performance is extremely important to the success of the business intelligence project. Yet, performance is a very vague and ambiguous term. It relates to many aspects of the system. Scenarios are most helpful for performance requirements. Make sure to specify exact user cases and what is the expected and acceptable response from the system.

One measure of performance is latency – the total time taken by the system from when a request is made until the response is received by the requester. Latency cuts across all aspects of the presentation area. For example, a user of the system logs in to the reporting portal and runs a report. There is latency between the machine of the user and the server hosting the application portal. The application server then typically issues a query against the database system hosting the data. The database server has a given latency for responding to the query which is made up of many smaller latencies, such as the latency to read a block from disk, network latency between different machines in a distributed system or latency by the CPU to add two integers. Scenarios are defined at the perceived performance level which is the visible latency to the user independent of all internal latencies of the system. The architect ensures that the proposed solutions meets the latency scenarios empirically by building different prototypes or proof of concepts.

Throughput is another aspect of performance and it's measured in things per second. It states how many operations, requests, records or queries per second a system can handle. The Transaction Processing Performance Counsel defines a set of performance benchmarks that are vendor independent and publishes performance number of various platforms. It is important to understand the different benchmarks and how vendors being considered for the Business

> “The ATAM method shows how well an alternative satisfies different business requirements and how business requirements impact each other”

> “An Utility Tree is defined as a hierarchical, tree structure with general broad categories at the first level”

Intelligence solution performance on the benchmark.

Different vendors also have varying methods and techniques for enhancing performance of queries. There are different indexing techniques such as b-tree index or bitmap index which is very suitable for a dimensional model design typically used for implementing data warehouse and decision support systems. Other techniques involve passing hints in the SQL statement that tells the query engine the degree of parallelism to use for executing a given query or the join algorithm that is most suitable for the data. Partitioning is also a tool for increasing performance of the data warehouse and is used to prune data and only scan partitions that satisfy the criteria of the query dramatically decreasing the amount of I/O the system has to do.

Some vendors store the data in column oriented fashion, columnar databases, that are essentially a way of vertically partitioning the data. Columnar oriented databases increase performance by only scanning columns that are needed for the execution of the query, either projected or included in the where clause of the query. They also have the advantage of better compressing the data given that columns contain similar values in closer proximity to each other which results in better compression.

Lastly, some vendors rely on proprietor hardware to enhance query performance. Some relational operations are pushed down to the hardware layer resulting in much better performance. Such techniques include pushing filters to hardware so that disk controllers only return data that satisfy a where clause.

3.4 Operability

After a business architecture solution has been built and push to production it lives there for a long time. Most data warehouses have teams dedicated service engineering and database administration teams working tirelessly to ensure that the system is meeting its availability and performance requirements. A well designed system is one that treats operability features as first class citizens. Everything has to be automated, monitored, self-healing and self configuring.

The staging area of the solution is where most of the data processing happens. Therefore, great attention needs to be paid for the operability of the stating area. A very important component to the operability of the staging area is a work-flow management system. Work-flow managers allow developers to express ETL processing as an acyclic direct graph where nodes are processing jobs and edges are dependencies. Advantages of such modeling has enormous benefits to the operability of the ETL system. They typically come with a graphical user interface that allows the operator to inspect the progress, or lack of progress, of the ETL pipeline. They also ensure that processing jobs run in the correct order and that in the case of failures only subset of the pipeline is re-executed.

Another important aspect is alerting ability. Alerting should involve complex event processing that ensures that the right amount of alerts are being sent. Over alerting results in thrashing of operator and possibly loss of important alerts. Different vendors allow for different types of alerting such as paging, e-mailing, integration with ticketing systems or graphical user interfaces. They also allow for different severity levels of alerts such as Info or Critical. They monitor the application, the platform, the different services and the hardware of the end to end solution.

Operability of the presentation area is just as important as operability of staging area. There are numerous jobs constantly running in the staging area. Data needs to be backed up, retention policies has to be applied, indexes need to built and rebuilt. Cube in the OLAP system should automatically be triggered for reprocessing.

An advantage of using one vendor for the Staging and Presentation area is integrated monitoring and work-flow solution. Typically more mature vendors have an end to end solution with integrated monitoring and work-flow system.

One last important aspect to keep in mind while reviewing different vendors for operability is related to Quality Attributes and how well they handle rolling upgrades – specially in a distributed environment. Rolling upgrades are the ability to push new software versions to the production environment without having to bring down the system. Distributed systems have the added complication of ensuring that all components of the system are running compatible and consistent versions.

3.5. Time to Market

Time to market is very important factor for determining the best business intelligence (BI) strategy to follow. Most of the time, in addition to cost, time to market is the barrier for implementing a BI solution in house.

Since most of the development and testing efforts are in the staging area of the business intelligence solution, time to market features have to be carefully evaluated for ETL vendors. Many ETL vendors offer features that allow for rapid development of ETL processing. Such features include an extensive library of transformation and processing nodes. They give the ability to compose complex data pipeline by chaining together pre-built processing components, and allowing for description of metadata based ETL using logical mappings of attributes and transformations.

Many vendors also allow for flexible and automatic schema evolution and metadata driven ETL where new col-

Category	Sub Category	Scenario
Availability	Business Continuity	*Source:* Nature *Stimulus:* Earthquake destroys data center *Environment:* Typical load *Artifact:* entire system is destroyed *Response:* • No data loss • Recover within 2 hours *Response Measure:* simulate failure, switch to new geographical location, run report on old system and new system
Availability	Fail Over	*Source:* Computer machine *Stimulus:* An ETL processing machine loses power during processing *Environment:* ETL job in progress *Artifact:* Particular job fails and intermediate data is in inconsistent state *Response:* System should recover automatically *Response Measure:* manually shut down down one of the ETL processing machines. ETL job should recover with no manual intervention
Scalability	Data Size	*Source:* Users of website *Stimulus:* a new feature on website increases page view to 100 million in on hour *Environment:* ETL system is running at 80% of its capacity *Artifact:* Double the number of rows in the input web server logs *Response:* add new hardware results in no changes to response ETL finish time *Response Measure:* Duration time of ETL processing jobs
Scalability	Concurrent Queries	*Source:* Product managers *Stimulus:* Due to a new product launch, all product managers are running the 20 product managers are running the same report at the same time *Environment:* ETL load has finished for the day *Artifact:* Load is increased on the OLAP tool as well as the DBMS *Response:* Only 20% degradation in response time *Response Measure:* Run 20 simultaneous reports and measure run time
Performance	Query Response Time	*Source:* User of reporting system *Stimulus:* A user query asks for one day of data to be reported on *Environment:* Normal conditions *Artifact:* Query sent to DBMS for processing *Response:* Only required horizontal application is processed *Response Measure:* See execution plan and compare to running time of query that accesses all partitions
Performance	Query Response Time	*Source:* User of reporting system *Stimulus:* A user query asks an aggregation that only uses subset of columns *Environment:* Normal conditions *Artifact:* Query sent to DBMS for processing *Response:* Only required columns are scanned and aggregated *Response Measure:* See execution plan and compare to running time of query that accesses all columns
Operability	Terminal failure	*Source:* Un-handled error condition in ETL job *Stimulus:* A job in ETL pipeline has failed *Environment:* ETL cleansing and transformation stage *Artifact:* ETL completely stopped *Response:* Error is reported on monitoring console, operator is alerted via a pager, problem is manually rectified, operators resumes work-flow from point of failure *Response Measure:* Simulate failure in processing job by removing input data in the middle of processing and measure the end to end time it takes to resume the pipeline
Operability	Upgrades	*Source:* Service engineering team *Stimulus:* A new ETL software version needs to be rolled out to production *Environment:* ETL jobs are running

Table 1 (Part 1 of 2): Example of Utility Tree including Scenarios.

Category	Sub Category	Scenario
Time to Market	Flexibility	*Source:* Upstream changes *Stimulus:* A new pass through column is added to the web logs *Environment:* Normal conditions *Artifact:* Input schema changed *Response:* Output Schema changed with the additional column *Response Measure:* Amount of time spent developing, testing and deploying new software
Time to Market	Modifiability	*Source:* Product manager *Stimulus:* A change to the the transformation applied to one of the columns *Environment:* Normal conditions *Artifact:* ETL code needs to be modified *Response:* New code is deployed *Response Measure:* Amount of time spent developing, testing and deploying new software
Compliance	SOX	*Source:* Government auditors *Stimulus:* SOX Audit *Environment:* Normal conditions *Artifact:* Login and Log out reports requested *Response:* Generate required reports within 2 days and no additional resources *Response Measure:* Time it takes to generate and validate the required reports and the number of people used to work on the task
Compliance	A29	*Source:* Upstream changes *Stimulus:* A new private and personally identifiable information attribute is added to one of the source systems *Environment:* Normal conditions *Artifact:* Additional column is added and additional transformation is needed *Response:* Personal Information is converted to anonymous one and stored in the presentation area *Response Measure:* No personal information stored in presentation area
Data Quality	Error Detection and Correction	*Source:* Upstream data quality issue *Stimulus:* A non-nullable attribute has a null value *Environment:* ETL load in progress *Artifact:* Error detection code is triggered *Response*: Reject malformed record and log it in a separate store *Response Measure:* Verify that the malformed record is rejected and logged
Data Quality	Metric Reporting	*Source:* Upstream data quality issue *Stimulus:* 10 input records were malformed *Environment:* Normal conditions *Artifact:* 10 records are rejected and stored *Response:* Run report on data and see that there are 10 rejected records broken down by reason of rejection *Response Measure:* Simulate input and run report

Table 1 (Part 2 of 2): Example of Utility Tree including Scenarios.

ing where number of partitions can be determined and adjusted dynamically based on the input systems. Some vendors facilitate team based development by supporting integration with source control systems allowing multiple developers to work on the project easily at the same time.

Some features provided by vendors also reduce the time it takes to maintain the ETL application by providing auto-documentation features, custom annotations of different processing jobs, automatically generated lineage report that can be used as a manual for the users of the data as well as new developers and impact assessment of changes to the system such as a data type change for one of the attributes could result in changes to only a subset of the processing jobs.

3.6 Compliance

Organizations have different standards that they need to comply with depending on the nature of the data they posses and the type of analytics they perform on it. For example, business intelligence solutions that are used for revenue recognition and reporting need to adhere to the Sarbanes–Oxley Act of 2002, also known as SOX. SOX compliance applies to publicly traded US companies and is a result of financial scandals affecting companies such and Enron and costing people billions of dollars.

> “The main factors for an organization to make a decision on the best Business Intelligence (BI) strategy to follow: availability, scalability, performance, operability, time to market, compliance, and data quality”

SOX compliance requires companies to document and show the flow of transactions. From a data warehousing perspective, this translates to the ability of extract lineage out of the ETL processing jobs. It also requires detailed reports about user activities such as login/logout events. Every access to the data needs to be documented along with the type of access such as read, write or delete. This is needed to ensure that the data has not been tempered with after it has been published by the ETL process. System events such as startup and shut down or changes to the system time or audit log need to be tracked to ensure that the ETL code has not been changes without proper authorization and approvals. Also tracking of account management and user group changes needs to be tracked to ensure that only authorized users have access to the data with the right permissions.

This requires all components of the warehousing and business intelligence solution to have detailed security and auditing features as well as comprehensive and structured logging to facilitate the generation of required SOX report.

Another form of compliance requirements are requirements for protecting user privacy. This is specially needed by companies that collect user behavior or financial data. The European Privacy Directive, specially A29, requires companies to not retain any user personally identifiable data such as browser cookies, IP address or searches that the user performed on their site. Companies usually handle this by converting the private information to anonymous values that are used to identify an unique anonymous person, instead of a login name, or aggregate the information to an appropriate level such as zip code instead of IP address. Some vendors have some pre-built components that allows such transformations or have the ability for the application developer to plug in their own transformation functions in the form of a UDF – user defined function.

3.7 Data Quality

Data quality measures the consistency and accuracy of the data. It is used to determine how fit the data is to be used for decision making. Data with poor quality is considered worse than no data at all since it leads to the wrong decision making.

It is the responsibility of the staging area to ensure that the quality of the data is of high standards before publishing into the presentation area. And, it is the responsibility of the presentation area to keep data quality metrics and expose them to the users of the data.

Most data quality issues are a result of bad data from the input systems. It is best to deal with the data quality issue at the source. In addition to that, ETL vendors have features that allow the detection of bad data and configurable actions to be taken when encountering it. Options include the ability to reject and log the bad data to be analyzed and possibly corrected offline, the ability to correct or nullify bad data or the ability to halt the ETL process until the data quality issue is investigated by an operator (not recommended). It is important to verify that the ETL vendor of choice meets your data quality issues handling requirements. The ETL system also needs to aggregate the number of data quality issues encountered and publish them with the final datasets to be consumed by different users.

In the presentation area, the BI tools need to show data quality metrics to users. This can be added to all reports as as custom aggregation. Some BI tools also allow users to collaborate in a discussion about the and its quality.

3.8 Sample Business Intelligence Utility Tree

Table 1 is an example Utility Tree, represented in tabular format, for a business intelligence system. It highlights the different important quality attributes and give an example of scenarios.

4 New Trends

Although open source software has been around for a long time, it only recently became used widely as part of BI solutions. Most notably is Hadoop, an Apache based open source java implementation of Map/Reduce framework. Although Hadoop has not yet reached version 1.0, it is being used in over one hundred companies that are listed on the Hadoop page of the Apache website, <http://hadoop.apache.org/>. More so, there are different sub-projects of Hadoop that focus on making it more productive and suitable for business intelligence projects such as PIG, a procedural query and processing language on top of Hadoop. Hive is an SQL implementation on top of Hadoop and Oozie

is a workflow manger for Hadoop based jobs.

Additionally, many companies are using in-house or external cloud computing techniques to process their data. Cloud computing with regard to Business Intelligence solution entails ease of provisioning new hardware resources (scalability and performance), geographical location independence (availability), and automatic and live deployment (Maintainability).

5 Summary

Before discussing alternatives for implementing a Business Intelligence solutions, it is important that the quality attributes are documented and reviewed by all stockholders of the project. Quality attributes serve as a medium of communication across multiple teams. It also helps document and serve as a reference for the rationale and reasons behind the decisions that were made. After building the quality tree, spending time writing detailed scenarios. Get all stakeholders to review them and participate in the prioritization process. Make sure that one person is ultimately responsible for assigning the priorities for scenarios otherwise consensus on priorities maybe impossible to achieve. Spend time researching different technologies available on the market and determine how well they meet the different scenarios. Document your findings, review them and move on to implementation.

Bibliography

- L. Bass, P. Clements, and R. Kazman Rick. Software Architecture in Practice. Second Edition. Addison Wesley, 2003.
- R. Kimball and M. Ross. The data warehouse toolkit: the complete guide to dimensional modeling. Wiley Publishing Inc. 2002.
- R. Kimball and J. Caserta. The data warehouse ETL toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley Publishing Inc., 2004.
- P. Lyman, and H R. Varian. How Much Information? The Journal of Electronic Publishing, Vol. (6), Number 2, 2000.

Strategic Business Intelligence for NGOs

Diego Arenas-Contreras

This article shows how to plan and apply a Business Intelligence (BI) strategy to a nonprofit organization starting from the understanding of Non-Governmental Organizations' (NGOs) nature and goals, their organizational processes and the identification of information needs, relevant available data, and proprietary information to meet the information requirements that an organization has. These ideas are developed through the study of an actual project carried out in a NGO in Chile. The main data entities identified were "lead", "contact", "company", "institution", "volunteer", "donation", "event", and "campaign". The interaction between these entities and the understanding gained from raw data enables us to obtain valuable information to make decisions in NGOs. The whole process will be described in order to implement a successful information strategy in these organizations.

Keywords: Business Intelligence, CRM, Data Analysis, KPI, MDM, NGO, Reports.

Author

Diego Arenas-Contreras studied Civil Engineering in Computing at the *Universidad de Talca*, Chile. He holds a Diploma in Business Management with Business Intelligence and is passionately interested in the world of information systems and information visualization. He has worked in various companies and industries on such subjects and projects as Business Intelligence, Performance Management, and Data Mining. He is currently Head of Business Intelligence Solutions at Formulisa, <http://www.formulisa.cl>, a Chilean consultancy firm dedicated to understanding the behaviour of consumers and Customer Intelligence for its clients. <darenasc@gmail.com>

1 Introduction and Definitions

This paper describes how to implement a successful Business Intelligence (BI) strategy in a Non-Governmental Organization (NGO). We will understand a strategy as an action plan to gain a competitive advantage compared to an earlier stage of the organization or to a similar organization. Business Intelligence is defined as a set of tools, processes, techniques and algorithms which support the process of business decisions and bring the right information to the right person (decision maker) at the right time.

In every organization decisions are made at all levels. Decision support systems improve efficiency and help management make informed decisions, and provide a guide to an organization's continuity. Nonprofit organizations make decisions too, but their focus, goals, data and day-to-day operations are different and so their problems differ. However they can make decisions based on their specific data and can generate knowledge from their information.

Voluntarios de la Esperanza Global, Volunteers for Global Hope, a.k.a. VE Global, (VE from now on, <http://www.ve-global.org>) is the NGO where this knowledge was obtained. VE is an organization that recruits, trains and organizes volunteers to work with children at social risk in Chile, and this paper draws on VE's knowledge and experience. The lack of people specialized in information systems and the NGO's priority focus on social issues brought about the start of this project and the results commented on in this paper.

An NGO, like any other organization, interacts with people and other organizations, public and private. These interactions are constantly working to make better and more efficient decisions, but only a few are based on data.

Data generated by the organization itself is one of the most important organizational assets because no other organization has access to it. Data has to be identified in order to propose a strategy based on it. A proper information strategy for an NGO is significant not only at an organization level but also at a social level.

The first step is to identify the information needs and to discover the organization's strategic goals through meetings with the directors and by generating agreements to guide the information strategy. It is important to bear in mind that strategy information must much the actual organizational capabilities. At the beginning of the project and for this paper, it is imperative to unify semantics to facilitate communication between stakeholders and to speak a common language during the project; this will simplify communication issues and will bring agility to the project. At VE we used a shared document to write known definitions and to add terms which require definitions. Everyone could add a new column until we reached an agreement as to the definition. This task gives an active role to stakeholders and assures the knowledge obtained in the project to new participants. Wikis are recommendable for this purpose. Important examples of definitions for this paper are:

- *Lead*: Person or institution with a potential mutual relationship with VE.

> “ In every organization decisions are made at all levels; nonprofit organizations make decisions too, but their focus, goals, data and day-to-day operations are different ”

- *Contact*: Person with some relationship with VE; there are different types of relationships.
- *Organization*: Company and organization involved with VE; it allows contacts to be grouped within organizations.
- *Institution*: Household where VE works through volunteering programs.
- *Volunteer*: A person trained and led by VE who works in an institution and/or at a VE office.
- *Donation*: A monetary or in-kind donation to VE by contacts or other organizations.
- *Program*: A specific plan designed to support children in VE's partner institutions with a specific purpose, i.e.: "Liga de deportes" is a sports programme that promotes the practice of sport in children; "Vamos a Leer" is a programme that encourages children to read.

As in VE, there are specific terms in every organization and it is necessary to define them to facilitate communication.

Some tools used during this project were collaborative and open source, thereby keeping project costs down. They were very useful for effective communication:

- Google Docs, to make documents, working collaboratively, <http://docs.google.com>.
- Dropbox, for file exchange, <http://www.dropbox.com>.
- GanttProject, an open source tool to manage the schedule of the project, <http://www.ganttproject.biz/>.
- FreeMind, to draw mind maps, to express and discuss ideas, <http://freemind.sourceforge.net/wiki/index.php/Main_Page>.
- Salesforce, a sandbox testing environment of "Salesforce", where we test changes in a system without affecting the production system.

Like many organizations, VE maintains several different repositories of information. Their aim is to put these sources of information into their CRM system and to start using it as a single and reliable source of information in order to maximize the potential of their CRM system.

The first step of a BI strategy must be aligned with the organization's strategic goals. Then we must work in collaboration with the people responsible for the information with an effective communication system and transmit project goals to the stakeholders. The next step is to analyse the available data; the main processes must be documented and the data flows identified from the perspective of strategic goals. Then, we must work on the quality of the information through data quality assurance and then apply the strategy based on the organization's capabilities and anticipate future information requirements based on current data. Finally, we need to monitor system use and evaluate improvements.

2 Data, Information and Processes

Meetings with the directors of VE were arranged in order to discover the strategic goals and information needs of each area of the organization. At the first meeting, interviews were conducted to know the organization's vision and the data that they manage, then brainstorming sessions were organized in which directors were asked to imagine the available information and the reports to produce and the decisions they could make if the project was successfully implemented. This process allowed us to identify the entities and their interactions. From these activities, you can obtain two artefacts; the high level of entities, their relationships, and the required reports document. After the first brainstorming sessions, thirty minute meetings were scheduled to report on progress and to set achievable weekly goals. Stakeholders and project staff took part in these meetings. If there was something that could not be defined during a meeting, an extra meeting was scheduled during the week with the people involved in that specific task.

During the first stage, we need to find out about the organization's information needs, organizational culture, the unique data that belongs to the organization, the main information processes and actors, and the data flows. In this way we can discover what data is relevant to the organization. A BI expert's role is to show the benefits of a BI solution, to recommend the right solution according to the needs, and to define a strategy to achieve it.

Once processes are well known and the information needed is sufficient to plan a BI strategy, we need to share that knowledge with the stakeholders. We need to define a common language which is going to be the basis of the project, then we will deliver an artefact called "organizational definitions". Each term in the document has a unique and homogenous semantic. We need everyone in the project and in the organization to understand the same thing when they hear the term *volunteer* or *institution*, for example. We are assuring understanding between stakeholders and the

> “ The first step is to identify the information needs and to discover the organization's strategic goals ”

> “Some tools used during this project were collaborative and open source, thereby keeping project costs down”

definition of the vocabulary for an effective communication and a favourable development of the strategy. If we all speak the same language we can avoid any misunderstanding.

Another artefact is the "required reports" document. It is a shared document in VE where we state the owner of the report, the report's name, a summary, the reasons or the justification for writing the report, the consultancy frequency, the required data fields, and observations or indications. We have collected over forty-five required reports since the beginning of the project and we have turned the reports into a mind map in which the first level of nodes show the entities such as "Lead, Contact, Organization, Donation, Event, Campaign and Program" and, within these entities, we group required custom fields to meet reporting needs, i.e.: we put "contact data" under "information data", "lead", "entity" and so on. We also add metadata fields to know how people come to VE and other types of information. In the mind map, created with FreeMind, we can find the entities and their custom fields to complete the reports; we need to input these custom fields into the database with their correct values. We also gather reports according to their topics; by doing so we can identify VE's three main areas of interest; *Contacts*, *Donations* and *Performance*. "*Contacts"* includes reports which are used to reach people and organizations according to a number of criteria. Meanwhile "Donations" contains correct information about monetary or in-kind donations, for example: Where are they? How are we using them? What are they used for and by whom? etc. "Performance", a recurrent topic in every organization, contains reports to optimize resources, to focalize efforts, to reduce costs, and to make decisions based on data. All the above is shown in Figure 1.

A BI strategy is not a specific project with a limited outcome such as a special report or a dashboard; it is a plan to execute and to take into account information needs and to deliver the right actions to satisfy these needs according to current organizational capabilities.

Every organization generates specific data specific to the institution's activities and day-to-day operation. No other organization has access to this data so it is important to consider it as an important asset of the organization and a unique source of information to improve performance and to acquire knowledge. This *unique data* is inherent to the organization and can be surveyed by means of interviews and by looking at existing reports and the data quality of the data model. It is also possible to identify potential data which is not being kept and which is unique to the organization. This step is necessary to ensure that this information is recorded in the BI strategy.

NGOs often lack information management specialists which results in processes and data flows being made on demand, depending on the person in charge (the rotation of volunteers makes the integration process more complex). A recurrent issue is the poor utilization of systems such as CRMs, because of the low level of customization and adaptation to the specific needs of the organizational. In NGOs communication with their network of contacts means continuity. It is important to provide information about the work being carried out and to measure these contacts. To acquire available information, the NGO has to organize events in collaboration with many of its contacts. To obtain accurate statistics, the NGO should monitor volunteers' working hours, know how many people it is working with, and have the capacity to share the work done with accurate figures for a better communication within the NGO's community. Hence, the organization has a more efficient and guided information.

From the beginning of the implementation you should focus on the relevant information of all the data and processes surveyed, rank the information in terms of its importance and impact on the organization, and focus your efforts on meeting the organization's information requirements in accordance with this ranking.

3 Design and Implementation of a Strategy

A successful key factor is to have executive support from the directors and the stakeholders of the project at every stage of the implementation. They must focus on information which is relevant to the organization and assure that important issues are covered according to the ranking of that information. The aim should be to cover the most important information needs, both current and future, revealed by the survey, and also the most important processes and their impact.

A Business Intelligence strategy is independent of the tools, software, and actors involved; it is about covering

> “NGOs often lack information management specialists which results in processes and data flows being made on demand, depending on the person in charge”

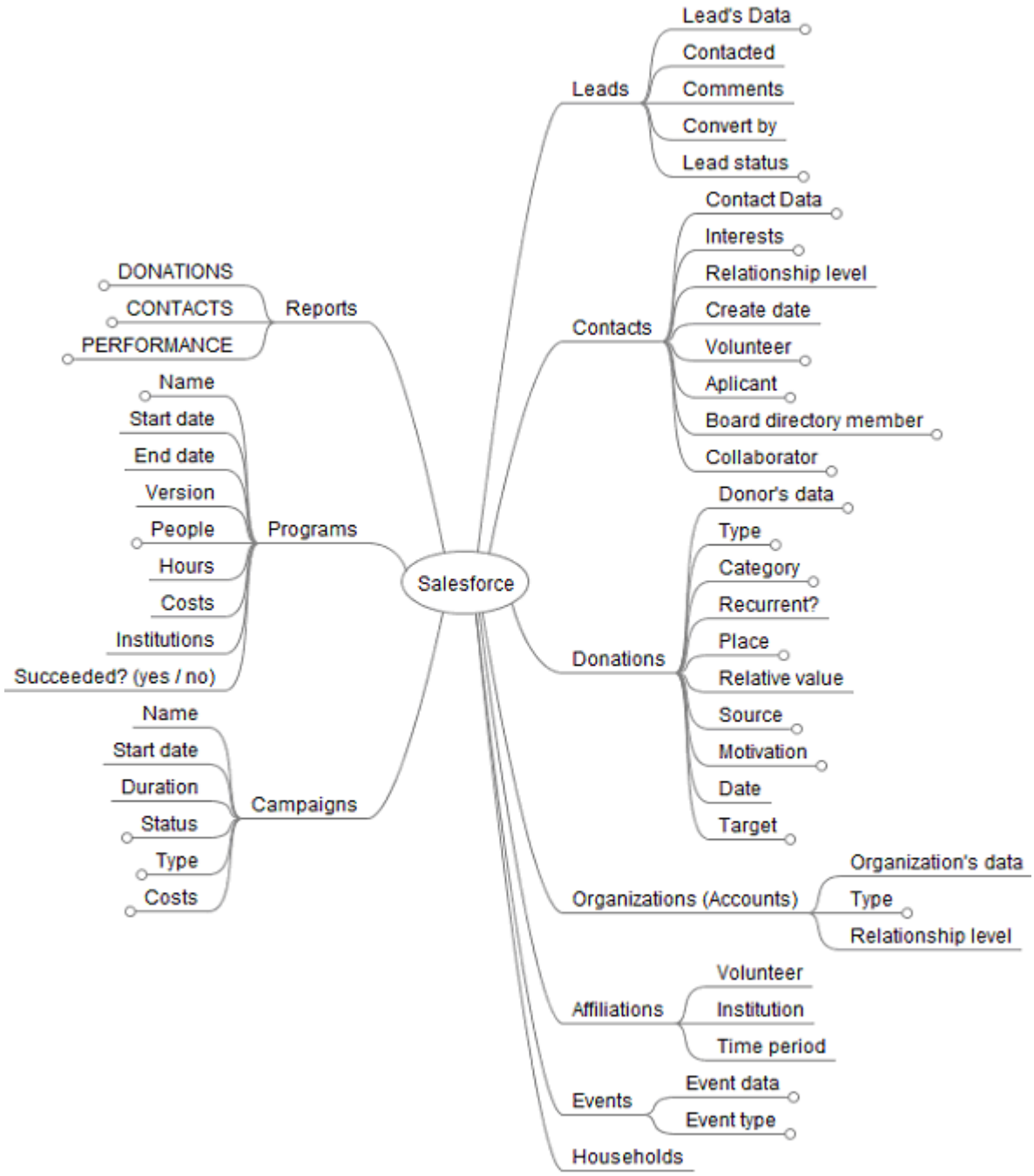

Figure 1: First Level of Mind Map showing Minimum Fields and Subjects of the Reports.

needs at every level of the organization. Information is the oxygen that allows the organization to breathe. The architecture is dependent on the information needs and is defined after the initial survey of information needs and processes of the organization. It is important for NGOs to use flexible and agile processes in order to rapidly align constant changes in information needs and to ensure the alignment of new strategic goals. The NGO has to evaluate new requirements in a short time. For example, during the first stage we evaluated and implemented new features in the system that we had not scheduled at the beginning. The flexibility of the development allowed us to do this and thereby incorporate improvements that were aligned to the goals of the project.

Data owners and information owners must be clearly defined. You must evaluate current data quality and make a plan to continue the improvement of data quality and information quality. It is also necessary to define standards for the security of data and information access because there will be sensitive data in the system. The design of the strat-

> “It is important for NGOs to use flexible and agile processes in order to rapidly align constant changes in information needs and to ensure the alignment of new strategic goals”

egy must consider a series of reports, performance indicators and measurements of the main processes.

At VE we obtained relevant data and main processes through interviews and brainstorming meetings with the directors. We ensured that this knowledge was reported in shared documents. In the mind map, we classified the reports and entities in the data model so we could ensure the minimum data to meet information needs. This data becomes a custom field in the CRM. We then evaluated the possible values and interactions with other entities before inputting these custom fields into the system. We then gave the possible values to complete the minimized open input data in the system, amending data quality such as a country list, questions with alternatives, etc. Next, we analysed the reports according to the area of interest, reducing them and maximizing their information.

The reports were given to the reports' owner so that every issue found in the reports could be dealt with directly by the owner. Opportunities for improvement and optimized processes were detected during the project such as the direct use of web forms in the CRM database. It is possible to collect data automatically, which reduces the time of the process and the duplication of the information source. The update of data allows decision makers to know the organization better and to make informed decisions, which also allows clear and segmented communication be sent to the to the network.

It is advisable to unify data sources, to select a unique repository of information, and to manage every information request from there, as was the case at VE, which chose to input data and processes into their CRM system. By identifying processes we facilitate ongoing improvement while the optimization of processes facilitates access to information and the inputting of data into the system. It is necessary to have the user's support and commitment to use the system properly. You must know the benefits of the proper use of the information which is being entered into the system and to know how much data quality and information quality depends on the user. Users will have to be trained in inputting data and reporting features. For this reason the early involvement of users is recommended.

Finally the strategy intends to give the right information to the right users. It is necessary to have traceability of data sources in reports with performance indicators. A BI strategy improves the relationship with organizations' partners and general contacts and establishes a social link thanks to their own data and information.

4 Conclusions and Future Work

The guidelines in this paper aim to show how to plan and implement a BI strategy in an NGO and how to customize it according to the organization's specific needs.

The idea is to measure program performance and campaign effectiveness so as to replicate this knowledge to similar organizations with similar objectives and to share in the success of this formula. Here on in there are many opportunities to manage by using information, such as identifying the profiles of the people and organizations in your support network, profiling donors, volunteers and contacts... It is also necessary to analyse past donation data to build a predictive data mining model for donations. Knowledge and data or information such as volunteers' working hours, missions that have been completed, the number of people in each partner institution, how the organization was led in the past and its status today, must be shared within the NGO's community.

Acknowledgments

To VE Global for giving me the opportunity to develop this project with them; it has been a great professional and personal experience. To volunteers Bushra Akram and Ben Richman who have worked hard to implement the project and bring it to fruition. To Josh Pilz, VE's executive director, for his support and enthusiasm; to directors and collaborators, Annie Rondoni, Mariah Healy, Jamie Ensey and Faith Joseph for their work and readiness to help at any time. To Clementine Bouchereau, Operations Assistant Intern, for her help in translating this paper into English.

Bibliography

- C. Hudson. Successful Business Intelligence: Secrets to Making BI a Killer App. McGraw-Hill Osborne Media, 1 edition. 2007. ISBN-10: 9780071498517.
- O. Parr Rud. Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy. Wiley, 2009. ISBN-10: 9780470392409.
- H. Rouillé D' Orfeuil. La Diplomacia No Gubernamental (Non-Governmental Diplomacy). Lom – Chile, 2008. ISBN: 9562829723.
- D. Wubbard. How to Measure Anything: Finding the Value of Intangibles in Business. Wiley; 2 edition, 2010. ISBN-10: 9780470539392.

Data Governance, what? how? why?

Óscar Alonso-Llombart

Decision-making is based on the information we obtain from business data. All decision-making involves accepting a certain degree of risk, but the truth is that it is not always possible to have complete and hard data available ... In this situation, how can we achieve real value from the data we do have and provide a consistent view of the business performance?, how can we properly analyse the available data taking into account the constant changes that take place in every organisation?

Keywords: Data Governance, Data Management, Data Ownership, Data Quality, Data Stewardship.

1 Introduction

Being able to obtain the real value of data is not an easy task. Data can be collected from multiple channels and then stored in different information systems and databases hosted on heterogeneous technology environments and in differing formats. Even when we have direct access to data, it is difficult to make use of them where, when and how we need to. Also data are often "dirty", full of errors, omissions and/or inconsistencies.

This issue is important enough to make any ICT strategic business project, initiative or even an entire company fail miserably. The data layer of an organization is a critical component, with which overly optimistic assumptions are often made and the real quality of the data is misunderstood or even ignored.

Data is used only in a technological environment is usually restricted to a process or an application with limited impact. There is also some data which is critical because it defines the most important identities (customers, products, employees, suppliers ...) and this has to be shared by multiple processes, departments and business lines. This data (called "master data") should be treated as a strategic asset.

Ensuring quality, integrity and accuracy of data is one of our greatest challenges. Ensuring a a clear and consistent view of data across departments, lines of business or other groupings in a modern company, can be critical to the achievement of business objectives.

Achieving quality data is a philosophy that aligns strategy, business culture, and technology to manage data for the overall benefit of the company. In short, this is a competitive strategy that every company can use to differentiate themselves from their competitors through their data quality and their use of data.

> “Being able to obtain the real value of data is not an easy task ... data are often "dirty", full of errors, omissions and/or inconsistencies”

Author

Óscar Alonso-Llombart is Engineer in Management Software by the *Universitat Autònoma de Barcelona*, Spain, has a Master's degree in Software Engineering by the *Universitat Politècnica de Catalunya*, Spain, and is Graduate in Data Mining by the *Universitat Oberta de Catalunya*. He works as Analysis Manager at the Spainsh company Penteo. He has over 15 years of experience in technology consulting in areas such as Business Intelligence, Datawarehousing, Corporate Performance Management, custom development, implementation of development methodologies, etc. He is author of numerous articles and studies on the application of information systems to the business strategies. Twitter: <@oalonsollombart>; Linkedin: <http://www.linkedin.com/in/oscaralonsollombart>. <o.alonso@penteo.com>

But to what extent does poor data quality affect today's business? Due to the dynamic nature of data, which is typically generated by numerous business processes and combined information sources, stored and used in various systems, it is a major challenge to establish methods to evaluate the impact of poor data quality.

Despite this challenge, it is clear that poor data quality has a real economic cost, primarily in the efficiency of processes.

From research conducted by Penteo it is clear that there is still a considerable *gap* between today's insight and true business intelligence for almost all companies. Although there are many companies that have implemented business intelligence systems, a significant percentage have only done so in isolated projects, responding to very specific needs. In the vast majority of the companies the challenge is to find and properly exploit the data and information on status and progress of the business itself (see Figure 1). This situation invariably impacts the business in terms of economics, confidence about the data, regulatory compliance, satisfaction and productivity.

2 Data Management

Business processes rely heavily on information systems, systems that interact with each other, sharing information and being able to communicate in order to provide adequate and efficient service to the organization. It is possible to make strategic decisions based on information extracted from the systems, and we must have reliable information

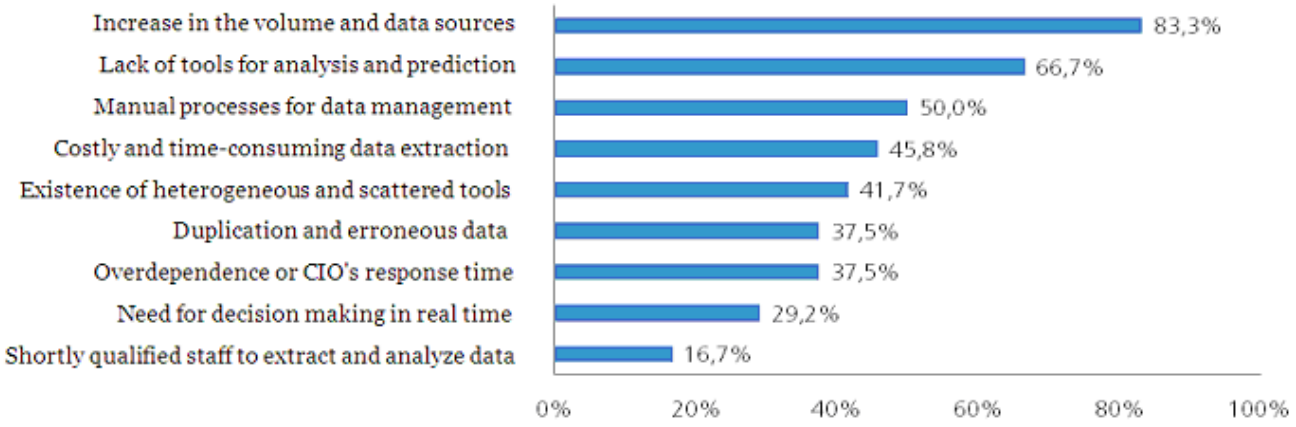

Figure 1: What are the Main Problems in making Decisions? [Source: Penteo.]

for a good corporate management.

In this situation we must realize that we are dependent on the quality of the data we have in our organization. Data as an entity in itself does not add value to business and business intelligence solutions are nothing if we do not have reliable data. Those companies that manage data quality effectively tend to avoid the problems arising from decisions based on poor or misleading information.

Data Management is the first piece on which to base an appropriate use of information (see Figure 2), after considering the data and information inferred from them as valuable business assets. The data and information must be carefully managed, like any other asset, ensuring quality, safety, integrity, availability and effective use.

The objectives of Data Management are:

- To understand the information needs of the organization.
- To capture, store, protect and ensure the integrity of data assets.
- To continuously improving the quality of data and information including accuracy, integrity, integration, relevance and usefulness of data.
- To ensure privacy and confidentiality, and prevent unauthorized and inappropriate use of data and information.
- To maximize effective use and value of data assets and information.
- To be aware of and control the cost of Data Management.
- To promote usage of and deeper and broader knowledge of the value of data assets.
- To manage information in a consistent manner throughout the organization.
- Align Data Management and the technology needed to business needs.

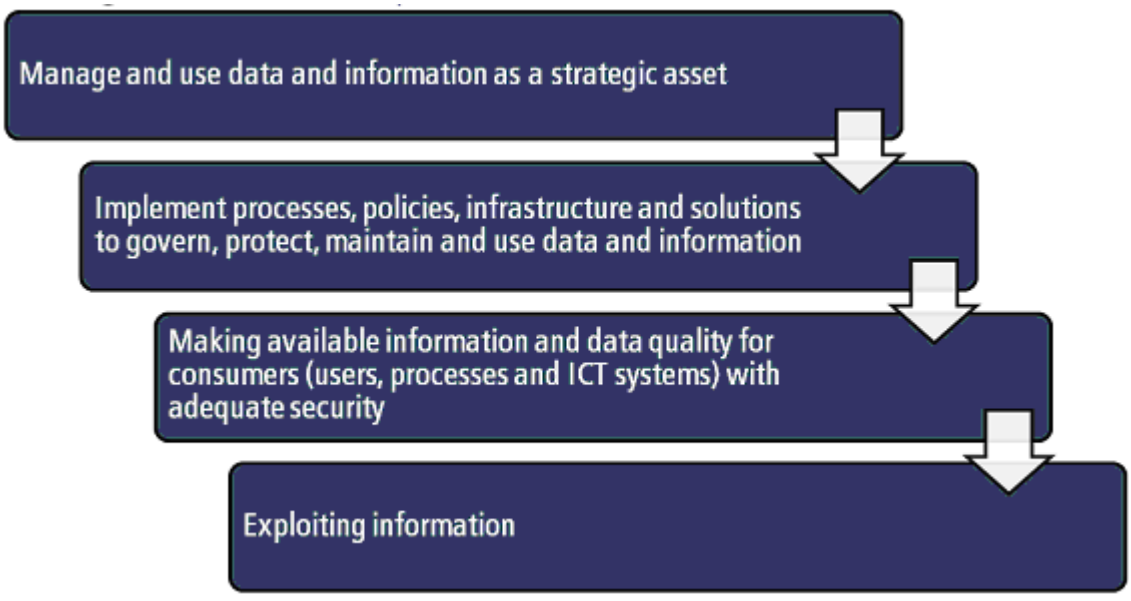

Figure 2: Data Management and Information. [Source: Penteo.]

> “Ensuring quality, integrity and accuracy of data is one of our greatest challenges”

Data Management has to be seen as a business function. Real competitive advantage is obtained through the appropriate use of the information.

3 Data Governance (Technology cannot solve the Problem by itself)

Data Governance... what is it and why is it important? What is the relationship between governance and ownership of data? Is the concept of Data Management included in data governance? Do we know the costs to the organization of having duplicate data or having no standard definitions of common data? If we are unable to answer these questions, perhaps we should consider a strategy for addressing the need to understand and use data more effectively and efficiently.

To achieve this goal companies must implement Data Governance projects, a set of policies and procedures that, combined, establish the processes that will allow you to transform data into a strategic asset and take the company to a higher level of "maturity" in the use of information, improve data quality and resolve any inconsistencies, managing change in relation to the use of data, and meet regulations and internal and external standards.

Data Governance is the cornerstone on which all practices related to Data Management underpin, that interact and influence each and every one of these, such as data quality, data integration or *warehousing* projects. Data governance is the exercise of authority and control (planning, monitoring and enforcement) on the management of data assets; data do not rule directly but governs how users access data through technology.

A Data Governance program provides guidance on how the other functions of Data Management should operate, appointing data owners on both executive and operational levels. It also has to properly balance objectives with compliance, which limit access to data, and integration of the business processes that increase access to them. The tasks that a Data Governance Program must carry out are:

- Guide information managers in making decisions.
- Ensure that information is consistently defined and understood by all stakeholders.
- Increase the use and reliability of the data as a valuable asset.
- Improve the consistency of projects across the organization.
- Ensure compliance with internal and external regulations.
- Eliminate possible risks associated with use of the data.

Data Governance implementation projects programmes are as unique as the companies that implement them. However, the structural frameworks that are used are actually quite similar to each other. There are common foundational components on which to build the initiative:

- *Organization*: structure responsible for deploying capacity of resources and administration of activities.
- *Policies*: principles and standards, guidelines for information management, and principles to ensure data standards and procedures of government.
- *Processes and practices*: establishing the principles that guide how the policies and processes are created, modified and implemented.
- *Metrics*: a measure to monitor the performance of the government initiative and the actions to significantly improve continues the quality of the data.
- *Data architecture*: including corporate standards data, metadata dictionary, and also security and privacy measures.
- *Tools and technology*: the tasks should be automated using software whenever possible, using data quality tools, data profiling, metaData Management tools, dashboards, etc.

4 Organizing a Team of Data Governance

This is an initiative that should not be considered as an ICT project, but as a continuous process of change in corporate culture. The business must lead the initiative, the implementation of Data Governance is an important change in mindset that must permeate all areas of the company.

Shared responsibility is the hallmark of Data Governance. It requires working across organizational boundaries and systems. Some decisions are primarily a business with input and guidance for ICT, while others are technical decisions and guidelines with input from users at different levels.

The different business units are represented in the "owners" of the data, while Department ICT provides the structure and processes. These data owners, experts in certain subject areas, are put forward as representatives of business interests with the data and take responsibility for the quality and use of these.

If, prior to the implementation of the data governance initiative, there have been Business Intelligence projects it is very possible that there is some sort of Data Governance team. This, although in an informal basis, should help mitigate the costs and organizational changes often required by this type of initiative, and will facilitate us having people who can occupy the profiles that are needed.

> “Poor data quality has a real economic cost, primarily in the efficiency of processes”

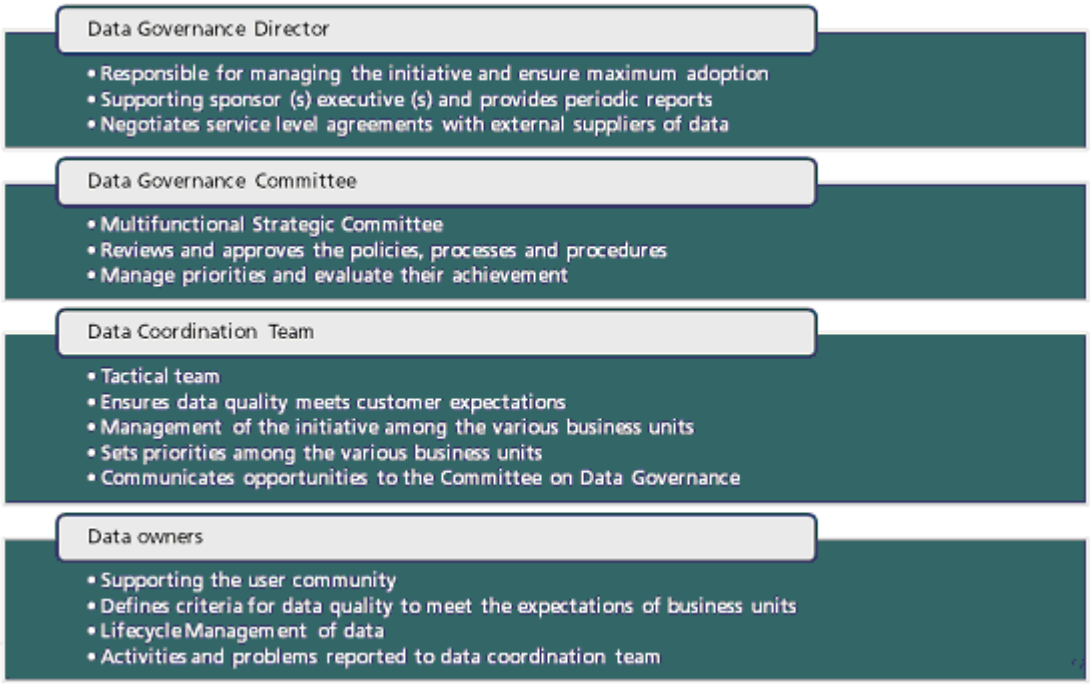

Figure 3: Organizational Chart of the Data Governance Team. [Source: Penteo.]

Staff forming part of the Data Governance team must know how to use and analyse information to facilitate decision-making, and require a mix of technical, analytical and business skills:

- Know the business, its processes, the analytical capabilities of the systems and the company's strategy to establish a master plan for data governance.
- Understand the organization and culture and channel access to information.
- Keep abreast of new capabilities that the technology can bring to the organization.

One of the historical problems in the implementation projects of Data Governance is the lack of adequate monitoring. While some organizations have successfully defined policies and government processes, often they have not put in place the necessary organisational structures to make it work properly.

The organisational framework government data programme must support the needs of all participants throughout the company. With the proper executive support, the Data Governance programme will benefit from the company's participation in the various functions required. This includes both strategic, such as data owners, and tactical, such as coordinators of data teams.

The specific roles include (see Figure 3):

- Director of Data Governance, responsible for managing the initiative and ensuring maximum adoption in the organization. This profile supports the executive sponsors and provides periodic reports project performance, as well as negotiating with external suppliers of data the associated service level agreements.
- Data Governance Committee, typically multifunctional strategic committee composed of the executive sponsor, the director of the Office of Data Governance, and the CIO of the company. Ideally, executive sponsorship should come from the business area rather than the ICT department. This committee reviews and approves the policies, processes and procedures, managing priorities and evaluates their proper discharge.
- Data coordination team, tactical team that ensures data quality meets the expectations of clients and manages the initiative among the various business units. It is the responsibility of this team to detect and communicate opportunities to the Committee on Data Governance.
- The owners of the data, which manage the lifecycle of data and provide support to the user community. These owners define the criteria for data quality to meet the expectations of the business units, and report the activities and problems with coordination team data.

5 The Need to establish Data Ownership

One of the key factors of successful implementations of Data Governance initiatives is the role of data stewardship or "Owner of the data." The ownership of data is the formalisation of responsibilities to ensure control and effective use of data assets.

> "There is still a considerable *gap* between today's insight and true business intelligence for almost all companies"

> “Data Management is the first piece on which to base an appropriate use of information”

The Data owners are business users, experts in specific subject areas designated as responsible for managing data on behalf of other users. They represent the interests of all stakeholders, including, but not limited to, the interests of their own functional areas and departments, protecting, managing and reusing data resources.

These profiles should be a business perspective to ensure quality and effective use of organisational data. The governance process will involve data owners as participants, but they will still be directly responsible for the successful management of data in their domain.

In practice there is no "silver bullet" model that fits all organizations. Basically there are five models of data ownership that organizations can apply, each of these models is unique, with its own pros and cons:

- Model 1: Property subject areas. In this model each data owner runs a particular subject area, as well as the responsibility of the customer data is different from those responsible for product data, etc. In large or complex environments, there may be more than one owner for each subject area. This model works well for companies with multiple departments to share the same data.
- Model 2: property business functions. In this case the owner of the data focuses on data that a department or line of business uses, such as data related to marketing, finance, sales, etc. Depending on the size of the organisation and management complexity, it may be that there are other owners of data by subject area, resulting in a hybrid model with the previous model.
- Model 3: property for business processes. Each business process is assigned a data controller, in this case the data owners are responsible across multiple domains of data or applications involved in a particular business process. This is a very effective model for companies with a clear orientation and a very clear definition of business processes. In organisations where there is no culture of immature processes then this approach is not the best choice.
- Model 4: Property for ICT systems. Those responsible for the data are assigned applications that generate the data they use. This model is a way to evangelize the concept of ownership of the data from the ICT department to the various business units. The data owners can report the progress of the initiative and show how the data will not only improve over time, but also will affect business results.
- Model 5: property projects. Associating the concept of data ownership with projects is a quick and practical way of introducing the culture of Data Management into the organisation. Unlike the models discussed above, this is a temporary measure which is often used as a starting point for the formal establishment of another long-term model.

To decide the ownership model of ideal data for the organisation is not a trivial task and is one in which we must consider a number of factors such as:

- Profiles and skills available in the organization for Data Management.
- The culture of the company.
- The reputation of the quality of the data.
- The current situation regarding ownership of data.
- Current use of metrics associated with data quality.
- The needs for data reuse.

6 How to tackle the Project of Data Governance?

A proper implementation of Data Governance can have a very positive direct impact on business performance. However, it is a challenge to achieve the right mix of people, processes and technologies to design a successful initiative.

To meet this challenge we must build a data governance strategy effectively, led by business objectives, providing stakeholders with improved capabilities for decision making and helping the company achieve its desired objectives. An effective strategy must ensure that company objectives, business strategy, investment and Data Governance systems are aligned.

A Data Governance initiative is nothing if not driven by the objectives of the company. Business requirements and business objectives should drive the iterations of the project. We need to establish a strategy before introducing the technology into the process.

Before beginning any work with data governance strategy, it is essential to understand and document the overall objectives to help formulate the vision and mission of government data for business growth. After documenting the initial list of objectives the major stakeholders must work to confirm the validity of the list of goals and proper prioritisation. This will ensure that we begin to build our strategy of Data Governance with a suitable base aligned with the business and users.

From Penteo’s market analysis and best practices we can draw the following:

- *Engaging business to lead the initiative*. Data Governance is not just a technology but also an important change in mentality that must transcend all areas of the company. Effective change management and communications from the start of the project are essential to ensure success. The

> “Data Governance is the cornerstone on which all practices related to Data Management underpin”

> “Shared responsibility is the hallmark of Data Governance. It requires working across organizational boundaries and systems”

project must be addressed from the component organisation and processes, but closely monitored by the ICT department. The historical existence of the role of organisation is emerging as a clear enabler of the adoption of the initiative.

- *Selling the process internally*. Deployments of Data Governance pose a significant impact on the organisation in many ways, so company CIOs should only start their data governance projects when they have reached a consensus on the decision with other officers involved in the process and when they have managed to successfully sell the project internally. Thus, the involvement in the project from different business areas is fully secured in advance and therefore the risk to address the process is much more controlled.
- *Adopting Data Governance must not be approached as a finite project*. The change of mentality, culture and the reorientation of the company together with the quality of information are indicators that identify the success of an initiative, hence it cannot be treated as a typical ICT project.
- *Manage a portfolio of strategic suppliers.* The current market situation forces us to evaluate, monitor and manage the ecosystem of our applications and road map a portfolio of solution providers to standardise and reduce risk, redundancy and cost. The selection of tools has less to do with the features and more with their ability to meet specific business requirements.
- *Planning and design prior to implementation.* This is a major initiative of high complexity so time must be takento define exactly the foundation of the future service-oriented system.

Finally, it is important that a Data Governance strategy should be designed to be agile and adaptive. It must be treated as a living entity that is constantly evolving to meet business objectives. The strategy should focus on communicating what is being planned to implement, how it will be implemented and when users will see their needs reflected in the system. Begin with general policies and guidelines and high-level diagrams as the ecosystem will mature in parallel with formal documentation and the level of detail identified in the strategy. It must be ensured that the data governance strategy evolves as part of the vision of the company as the iterative process produces more and more detail. Continuous evaluation and reinvention must be undertaken as business needs change, taking into account the current and future technology trends to support in building and delivering successful data governance strategy.

7 Conclusions

The tangible assets of organisations have a clear value and are managed through information systems and business processes. The associated data, precisely because of its intangible nature, is not collected on many occasions as strategic assets. However, accurate and available data is a pre-requisite for operations of any organisation to be effective.

Companies that are able to recognise the real value of the data, i.e.: that have established processes, policies and procedures for data quality, are aware of what data is really important or relevant to their business and ultimately rely on the quality of their data, they have become "data-driven organizations." These organizations have an obvious advantage over their competitors by managing the data as a more strategic asset, but to achieve this goal there must be an appropriate strategic vision to improve the quality of information.

The implementation of a Data Governance project requires the support of all business areas involved. Taking control of the data leads to better customer retention, increasing the success of marketing strategies, better control risks and, ultimately, allowing the company to be managed more effectively and efficiently.

Proper implementation of Data Governance eliminates discrepancies between data silos. However, those companies that have implemented these projects have realised at once that the timing of implementation varies greatly depending on the scope and simple exercises that are not technological in nature.

When taken correctly, Data Governance is a discipline helping to achieve the true value of analytic applications and should become the foundation for all initiatives in information management. To achieve proper management of these entities there must exist an appropriate strategic vision to improve the quality of information.

Are those projects that focus iteratively, starting with the set of needs and data that provide the greatest value to the business in the shortest possible time the most successful? Are you looking for a better decision making through Business Systems Intelligence? If the answer is yes to these questions then our starting point must be the analytical data. If we instead seeking to achieve greater operational efficiency or to gain consistency in processes across different transactional systems then we should start with the operational data.

Bibliography

- J. Dyche. Five Models for Data Stewardship. Baseline CONSULTING, 2009.
- D. Loshin. Data Governance for Master Data Management and Beyond. DataFlux, 2008.
- O. Alonso. Trends in the use of BI in Spain 2009. Penteo, 2009.
- O. Alonso. The problem of Data Management. Penteo, 2010.

Designing Data Integration: The ETL Pattern Approach

Veit Köppen, Björn Brüggemann, and Bettina Berendt

The process of ETL (Extract-Transform-Load) is important for data warehousing. Besides data gathering from heterogeneous sources, quality aspects play an important role. However, tool and methodology support are often insufficient. Due to the similarities between ETL processes and software design, a pattern approach is suitable to reduce effort and increase understanding of these processes. We propose a general design-pattern structure for ETL, and describe three example patterns.

Keywords: Business Intelligence, Data Integration, Data Warehousing, Design Patterns, ETL, Process.

1 Introduction

Business Intelligence (BI) methods are built on high-dimensional data, and management decisions are often based upon data warehouses. Such a system represents internal and external data from heterogeneous sources in a global schema. Sources can be operational data bases, files, or information from the Web. An essential success factor for Business Data Warehousing is therefore the integration of heterogeneous data into the Data Warehouse. The process of transferring the data into the Data Warehouse is called Extract-Transform-Load (ETL).

Although the ETL process can be performed in any individually programmed application, commercial ETL tools are often used [1]. Such tools are popular because interfaces are available for most popular databases, and because visualizations, integrated tools, and documentation of ETL process steps are provided. However, a tool does not guarantee successful data integration. In fact, the ETL expert has to cope with several issues. Many of the challenges are recurrent. Therefore, we believe that a support for ETL processes is possible and can reduce design effort. We propose the use of the pattern approach from software engineering because similarities exist between the ETL process and the software design process.

Software patterns are used in object-oriented design as best practices for recurring challenges in software engineering. They are general, re-usable solutions: not finished designs that can be transformed directly into code, but descriptions of how to solve a problem. These patterns are described in templates and often included in a catalogue. Consequently, a software developer can access these templates and implement best practices easily. The idea of design patterns has been adapted to different domains including ontology design [2], usage-interface design [3], and information visualization [4].

In the domain of enterprise system integration, the pattern approach is adapted by [5]. [6] develops patterns for the design of service-oriented architectures. In this paper, we present patterns for the design and implementation of ETL processes.

The paper is organized as follows: in Section 2, the ETL process is described, and in Section 3 we present the ETL pattern approach with three example patterns. A brief evaluation of these patterns is presented in Section 4, and in Section 5 we summarize our work.

Authors

Veit Köppen received his MSc degree in Economics from *Humboldt-Universität zu Berlin*, Germany, in 2003. From 2003 until 2008, he worked as a Research Assistant in the Institute of Production, Information Systems and Operation Research, *Freie Universität Berlin*, Germany. He received a PhD (Dr. rer. pol.) in 2008 from *Freie Universität Berlin*. He is now a member of the Database Group at the Otto-von-Guericke University Magdeburg, Germany. Currently, he is the project coordinator in the project funded by the German Ministry of Education and Research. His research interests include Business Intelligence, data quality, interoperability aspects of embedded devices, and process management. More information at <http://wwwiti.cs.uni-magdeburg.de/~vkoeppen>. <veit.koeppen@iti.cs.uni-magdeburg.de>

Björn Brüggemann studied Computer Science at Otto-von-Guericke-University Magdeburg, Germany, and received his Masters Degree in 2010. In his Masters Thesis, he focused on Data Warehousing and the ETL process in the context of Data Quality. Since 2010, he has been working at Capgemini, Berlin, Germany, in Business Intelligence and Data Warehouse projects. More information at <http://www.xing.com/profile/Bjoern_Brueggemann3>. <Brueggemann.Bjoern@gmx.de>

Bettina Berendt is a Professor in the Artificial Intelligence and Declarative Languages Group at the Department of Computer Science of K.U. Leuven, Belgium. She obtained her PhD in Computer Science/Cognitive Science from the University of Hamburg, Germany, and her Habilitation postdoctoral degree in Information Systems from Humboldt University Berlin, Germany. Her research interests include Web and text mining, semantic technologies and information visualization and their applications, especially for information literacy and privacy. More information at <http://people.cs.kuleuven.be/~bettina.berendt>. <Bettina.Berendt@cs.kuleuven.be>

2 The ETL Process

Data Warehouses (DW) are often described as an architecture where heterogeneous data sources, providing data for business analysis, are integrated into a global data schema. Besides the basis database, where data is stored at

> “Business Intelligence methods are built on high-dimensional data, and management decisions are often based upon data warehouses”

a fine-grained level, data marts for domain-specific analyses are stored, containing more coarse-grained information. Furthermore, management tools such as data-warehouse managers and metadata managers are included in the architecture. A DW reference architecture is given in [7].

The process of data integration is performed in the *staging area* in the architecture. Here, heterogeneous data are extracted from their origins. Adapters and interfaces can be used to extract data from different sources such as operational (OLTP) databases, XML files, plain files, or the Web. This extraction is followed by transformation into the DW schema. This schema depends on the DW architecture and the domain or application scenarios. In practice, relational data warehouse are used and star or snowflake schema are applied as relational On-Line Analytical Processing (ROLAP) technologies, see for instance [8]. In addition, transformations according to data formats and aggregations as well as tasks related to data quality such as the identification of duplicates are performed during this step. Finally, the data is loaded from the staging area into the basis database within the DW. Based on this, a cube or different data marts can be built, data mining algorithms applied, reports generated, and analyses performed. In Figure 1, we present the ETL process in its generic steps.

A *monitor* observes a data source for changes. This is necessary to load updated data into the DW. The monitoring strategy is defined depending on the data source. Two main strategies exist: either all changes are processed to the monitor and the delta of all changes can be computed, or the monitor can only identify that changes occurred. We distinguish the following mechanisms:

- Reactions are selected according to the event-condition-action rules for defined situations.
- Relevant data or changes are stored in an additional data store, therefore the data is replicated.
- Logs can be parsed and used, which are otherwise used for recovery.
- Applications that update data can be monitored via time stamp methods or snapshots.

The *extraction* operation is responsible for loading data from the source into the staging area. This operation depends upon monitoring the method and the data source. For example, it is possible that the monitor identifies a change, but the extraction process happens later, at a time predefined by the extraction operation. There exist different strategies for the extraction operation:

- ***Periodically***, where data is extracted continuously and recurrently at a given time interval.Tthis interval depends on requirements on timeliness as well as dynamics in the source.
- ***Query-based***, where the extraction is started when an explicit query is performed instantly. Where all changes are directly propagated into the dw.
- ***Event-based***, where a time-, external- or system-relevant event starts the extraction operation.

The *transformation* within the staging area fulfils the tasks of data integration and data fusion. All data are integrated and transformed into the DW schema, and at the same time, data quality aspects are addressed, such as duplicate identification and data cleaning. Different transformations exist and can be categorized as follows:

- ***Key handling****:* since not all database keys can be included into the dw schema, surrogates are used.
- ***Data-type harmonization***, where data are loaded from heterogeneous data sources.
- ***Conversion of encodings*** of the same domain at-

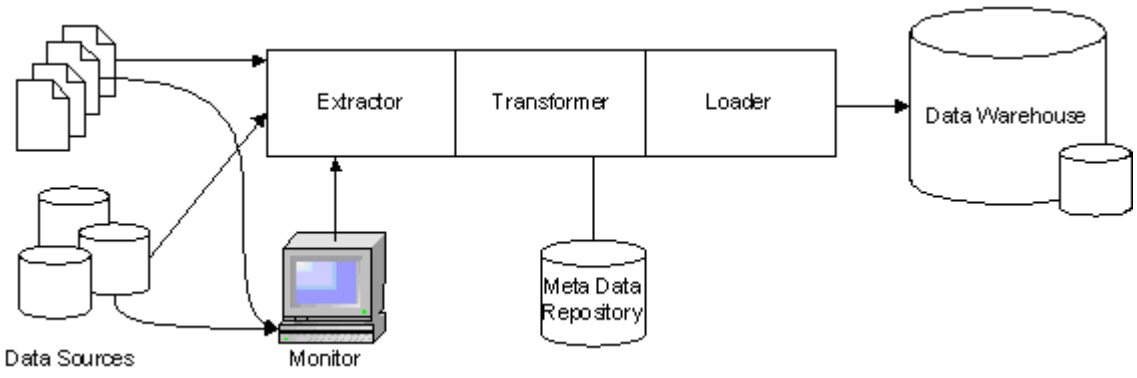

Figure 1: The ETL Process.

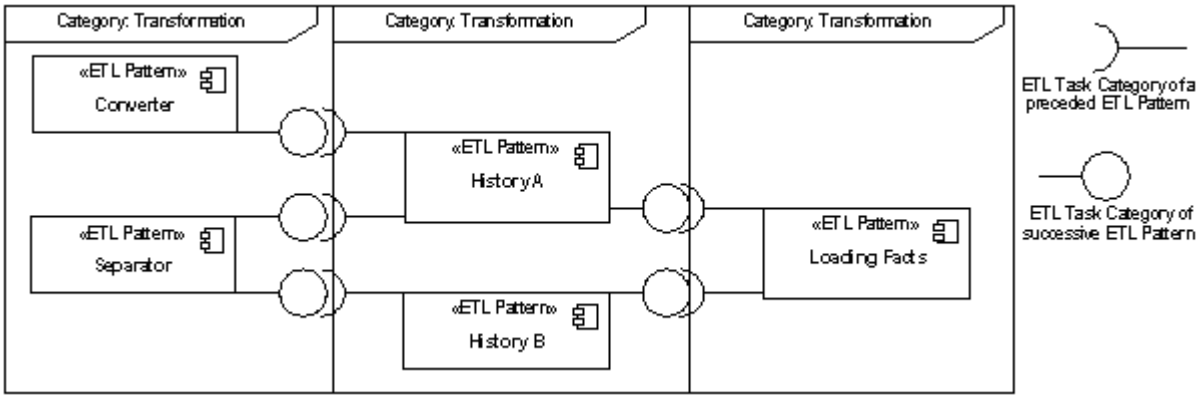

Figure 2: ETL Process with Patterns from Different Categories.

Element	META-DESCRIPTION	Mandatory?
Name	This name identifies the pattern in the catalogue.	Yes
Intention	A concise description at which use the pattern aims.	Yes
Classification	A reference to elementary or composite task with an optional refinement on the ETL steps.	Yes
Context	This describes the situation where the problem occurs.	Yes
Problem	A detailed description of the problem.	Yes
Solution	A concise description of the solution.	Yes
Resulting Context	This describes the outcome and the advantages and disadvantages of using this pattern.	No
Data Quality	Which data quality issues are addressed and which data quality dimension/s is/are improved.	No
Variants	A reference to similar and adapted patterns.	No
Alternative Naming	Other commonly used names of the pattern.	No
Composite Property	Only composite patterns use this description and state the composition property of the pattern.	No
Used in	This element describes briefly where the pattern is applied. this helps in the understanding and decision whether a pattern should be used.	No
Implementation	For various ETL tools, the solution is put into practice differently, therefore different implementations are referenced here.	No
Demonstration	A reference to an exemplary implementation of this pattern.	No

Table 1: ETL Pattern Structure.

> “The process of transferring the data into the Data Warehouse is called Extract-Transform-Load (ETL)”

tribute value to a common encoding (e.g., 0/1 and m/f for gender are mapped to m/f).

- ***Unification of strings***, because the same objects can be represented differently (e.g., conversion to lower case).
- ***Unification of date format:*** although databases handle different data formats, some other sources such as files can only provide a fixed data format.
- ***Conversion of scales and scale units***, such as currency conversions.
- ***Combination or separation of attributes***, depending on the attribute level of the heterogeneous sources and the DW.
- ***Computation and imputation***, in the case that values can be derived but are not given in the source systems.

The *loading* of the extracted and transformed data into the DW (either into the basis database or into data marts) can occur in online or offline mode. If the DW is or should be accessed while the loading takes place, an online strategy is necessary. This should be used for incremental updates, where the amount of loading is small. In the first (initial) loading of a DW, the loading is high and the DW is run in an offline mode for the users. At this time, the loading operation has exclusive access to all DW tables. Another task for the load operation is the historicization of data; old data is not deleted in a DW but should be marked as deprecated.

The ETL process can be refined into several *ETL steps*, where each step consists of an initialization, a task execution, and a completion. These steps enable ETL designers to structure their work. The following steps can be necessary in an ETL process: extraction, harmonization and plausibility checks, transformations, loading into DW dimensions, loading into DW fact tables, and updating. We use this categorization for our template approach in the next section.

3 ETL Patterns

The term "*pattern*" was first described in the meaning used here in the domain of architecture [9]. A pattern is described as a three-part rule consisting of the relations between context, problem, and solution. A pattern is used for recurrent problems and describes the core solution of this problem. For pattern users, it is necessary to identify problem, context, and solution in an easy way. Therefore, templates should be used to structure all patterns uniformly.

We derive our pattern structure from software engineering patterns because of the similarities between Software Design and ETL processes. A *template* consists of different elements such as name and description. For examples of templates in object-oriented software design see [10], for software architecture design patterns see [11], and for the domain of data movement see [12]. They all have in common that some elements are mandatory and others are optional. Mandatory elements are the name of the pattern, context, problem description, and core solution.

We see two levels of tasks in an ETL process: *elementary* and *composite tasks*. An elementary task inside an ETL process is often represented by an operator in the tools. A decomposition is not useful, although there might exist an application that allows a decomposition. We present the Aggregator Pattern as an example pattern for solving an elementary ETL task in Section 3.1.

Elementary tasks can be used in a composite task. A composite task is the sequence of several tasks or operators and therefore more complex. We can classify the composite tasks according to the ETL steps described in Section 2. Apart from the loading into the DW dimensions, all categories and consequently all ETL patterns are independent of the DW schema. We support the design of composite tasks in the ETL process by including composition properties. These *composition properties* describe categories of tasks that are executed before or after the composite task. Figure 3 depicts this composition property for the History Pattern described in Section 3.2. Before the History Task is performed, loading into the DW dimensions and transformations may be performed. After the completion of the History Task, a loading into DW fact tables or into DW dimensions is possible. Note that all elements are optional in this example.

Providing this information, a sequence structure can be defined and visualized as we present in Figure 2. In this way, the complete design of the ETL process can be given at an abstract level and customization of the ETL process can easily implemented.

We structure our ETL patterns according to the template shown in Table 1.

In the following, we present three ETL patterns as examples. In our first example, an elementary ETL task is presented, the aggregator pattern. In the other two examples, we present composite ETL tasks: the history pattern, where data is stored in the DW according to changes in DW dimensions, and the duplicates pattern for the detection of duplicates.

> “Although the ETL process can be performed in any individually programmed application, commercial ETL tools are often used”

> “A pattern is described as a three-part rule consisting of the relations between context, problem, and solution”

3.1 The Aggregator Pattern

Name: Aggregator Pattern

Intention: Data sets should be aggregated via this pattern within ETL processes.

Classification: Elementary task

Context: From a database or file data on a fine-grained level are loaded into the DW.

Problem: The DW data model does not require data at a fine-grained level. If data from the operational system is not needed at a fine-grained level, two problems may occur: more storage is required in the DW, and performance decreases due to more data having to be processed.

Solution: An ETL operator is used that collects data from the sources and transforms them into the desired granularity.

Resulting Context: A performance increase can be obtained, in the DW system as well as in the ETL process, through the reduction of data. Furthermore, the required storage space is reduced. However, one disadvantage is that there exists no inverse operation, so the inference to original data is not possible. If data granularity changes, information loss may result.

3.2 The History Pattern

Name: History Pattern

Intention: Data sets in the dimension tables should be marked and cataloged.

Classification: Composite task in the category of dimension loading for star schema.

Context: Product, Location, and Time are dimension in the DW that can change over time. Analyses in the context of master data can be done according to the dimensions.

Problem: Master data changes only occasionally, but they do sometimes change (such as the last name of a person). These changes should be taken into account in the dimension tables. However, challenges occur due to the use of domain keys that change over time, thus they cannot be use as primary keys. This is in contrast to the modeling of dimension tables in the star schema. Another problem is the use of domain keys if redundancy is required.

Solution: An important challenge is to store old and new data in the DW system. Furthermore, a relation of fact table and dimension data is necessary. For this purpose, the dimensional table has to be extended by additional attributes. In a first step, a virtual primary key is added, together with one or more attribute/s storing current or up-to-date information. The attributes *valid_from* and *valid to* are used to store the information about when the data was valid. This is described differently in the literature, for example as changes of type II dimensions [10] or as snapshot history [13]. For every data set, a decision has to be made: either it is a new dataset, an updated one, or a data set that already existed in the dimension tables of the DW. For this comparison, a key should be used that is persistent in time, such as the domain key. Every source data set is mapped with this key to dimensions. If this is not possible, a new entry is identified. If all attributes are equal for the source data set compared to a data set in the DW, an existing one is identified. Otherwise an updated data set is detected. A new data set has to be stored in the dimension tables and the attributes *valid_from* and *valid_to* as well as the virtual key have to be generated and *timeliness* set to true. For an update, the *timeliness* and *valid_to* information of the already existing dataset have to set before the source dataset can be entered into the DW.

Resulting Context: All data are historicized, however this influences performance due to the increase of the data amount in the dimension tables. The domain key has to be unique; otherwise, duplicate detection has to be performed first.

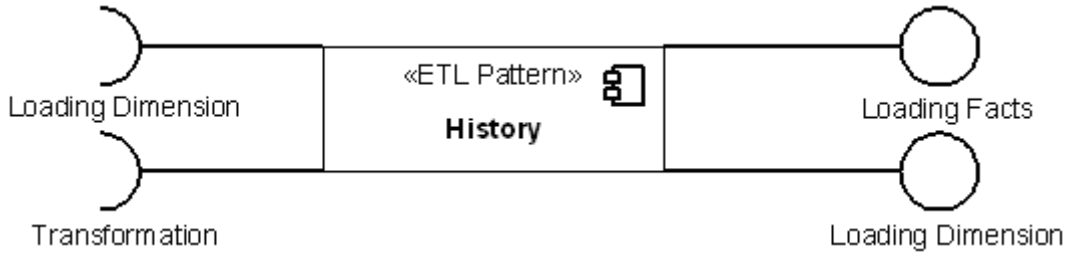

Figure 3: Composite Properties for History Pattern.

> “We derive our pattern structure from software engineering patterns because of the similarities between Software Design and ETL processes”

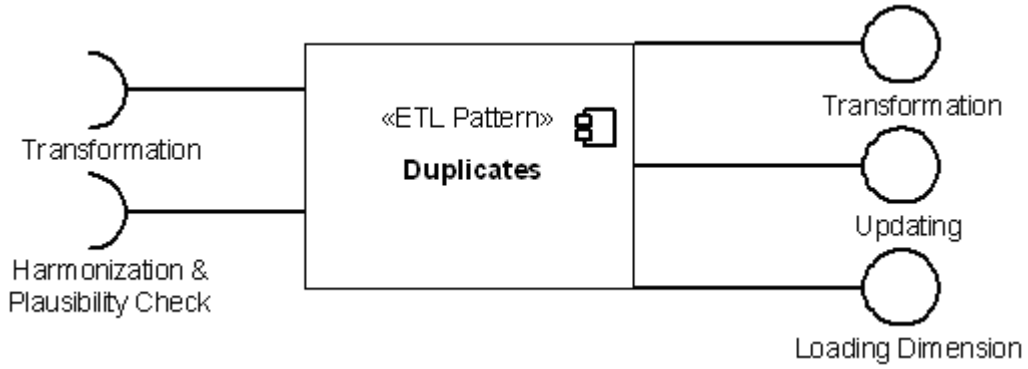

Figure 4: Composite Properties of the Duplicates Pattern.

Data Quality: All available information (*data completeness* for dimensions) is accessible for analysis with the help of the history pattern. Data *timeliness* is another advantage for data quality issues, as long as the loading is performed at short, regular time intervals.

Composite: Before an ETL task from the History Pattern is performed, patterns from the categories Loading Dimension and Transformation may be applied. The History Pattern can be followed by patterns from the Loading Facts and Loading Dimension categories.

3.3 The Duplicates Pattern

Duplicate detection is a common but complex task in ETL processes. With our pattern template, we briefly describe the solution, although in practice this should be described more comprehensively, see [14][15][16] for more details.

Name: Duplicates Pattern

Intention: This pattern reduces redundancy in the DW data; in the best case, it eliminates redundancy completely.

Classification: Composite task in the category transformation.

Context: Data from heterogeneous sources (e.g., applications, databases, files) have to be loaded into the DW.

Problem: A data hub for the integration of data is not always available, therefore master data redundancy occurs in different business applications. A duplicate are two or more data sets that describe the same real-world object. Data in the DW should give a consolidated view and must be free of duplicates.

> “The creation of complex ETL processes is often a challenging task for ETL designers. This complexity is comparable to software engineering”

Solution: Duplicates have to be identified and deleted. As a first step, data have to be homogenized. This includes conversions, encodings, and separations of all comparative attributes. Partitioning of data reduces comparison effort, but must be chosen with caution in order not to miss duplicates. The comparison is based on similarity measures that help to identify duplicates. There exist different methods and measures based on the data context.

A data fusion of identified duplicates has to be carried out. Aspects of uncertainty and inconsistencies have to be considered in this context. Inconsistency means that semantically identical attributes have different values. Uncertainty occurs if only null values are available. *Data conflict avoidance* can be carried out via the survivor strategy [17], where a predefined source entry is favored against all others, or via set-based merge [9], where the disjunction of all values is stored. In contrast, *data conflict resolution* can be carried out via a decision strategy, where an entry is determined from the sources, or a mediation strategy, where new values can be computed.

Resulting Context: Duplicates are only partially detected. Due to complexity of the duplicate detection, the ETL designer has to carefully consider data context and appropriate methods for measuring similarities or partitioning strategy.

Data Quality: The data quality issue *non-redundancy* is supported with this pattern.

Composite: The Duplicates Pattern can be preceded by patterns from the Transformation category as well as from the category Harmonization & Plausibility Check. The categories Transformation, Updating, and Loading Dimension include patterns that can be used for subsequent tasks, see Figure 4.

4 Conclusion and Future Work

The creation of complex ETL processes is often a challenging task for ETL designers. This complexity is comparable to software engineering, where patterns are used to structure the required work and support software architects and developers. We propose ETL patterns for the support of ETL designers. This provides an adequate structure for

> “We plan to create an ETL pattern catalogue with descriptions of most common ETL tasks and the corresponding challenges”

performing recurring tasks and allows developers to apply solutions more easily. In this paper we have presented a template for the general description of ETL patterns. Furthermore, we have presented three examples.

As future work, we plan to create an ETL pattern catalogue with descriptions of most common ETL tasks and the corresponding challenges. This includes an evaluation of the pattern catalogue as well as the application to different ETL tools.

References

[1] R. Schütte, Thomas Rotthowe, and Roland Holten, editors. Data Warehouse Managementhandbuch. Springer-Verlag, Berlin et al., 2001.

[2] OntologyDesignPatterns.org. <http://ontology design patterns.org>.

[3] S.A. Laakso. Collection of User Interface Design Patterns. University of Helsinki, Dept. of Computer Science. <http://www.cs.helsinki.fi/u/salaakso/patterns/index.html. 2003> [accessed July 20, 2011].

[4] J. Heer and M. Agrawala. Software Design Patterns for Information Visualization. IEEE Transactions on Visualization and Computer Graphics, 12 (5): 853, 2006.

[5] G. Hohpe and B. Woolf. Enterprise integration patterns. Designing, building, and deploying messaging solutions. Addison-Wesley, Boston, 2004.

[6] T. Erl. SOA Design Patterns. Prentice Hall PTR, Boston, 2009.

[7] A. Bauer and H. Günzel. Data-Warehouse-Systeme. Architektur, Entwicklung, Anwendung. dpunkt Verlag, Heidelberg, 2009.

[8] E.F. Codd, S.B. Codd, and C.T. Salley. Providing OLAP to user-analysts: An IT mandate. Technical report, Codd & Associates, 1993.

[9] D. Apel, W. Behme, R. Eberlein, and C. Merighi. Datenqualität erfolgreich steuern. Praxislösungen für Business-Intelligence-Projekte. Carl Hanser Verlag, 2009.

[10] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley Professional, 1995.

[11] F. Buschmann, R. Meunier, H. Rohnert, P. Sommerlad, and M. Stal. Pattern-Oriented Software Architecture. A System of Patterns. Volume 1. Wiley, 1996.

[12] P. Teale. Data Patterns. Patterns and Practices. Microsoft Corporation, 2003.

[13 H.-G. Kemper, W. Mehanna, and C. Unger. Business Intelligence - Grundlagen und praktische Anwendungen. Eine Einführung in die IT-basierte Management unterstützung. Vieweg Verlag, Wiesbaden, 2006.

[14] I. P. Fellegi and A.B. Sunter. A Theory for Record Linkage. Journal of the American Statistical Association, 64(328):1183–1210, 1969.

[15] A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios. Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering, 19:1–16, 2007.

[16] C. Batini and M. Scannapieco. Data Quality: Concepts, Methodologies and Techniques. Springer, 2006.

[17] R. Hollmann and S. Helmis. Webbasierte Datenintegration. Ansätze zur Messung und Sicherung der Informationsqualität in heterogenen Datenbeständen unter Verwendung eines vollständig webbasierten Werkzeuges. Vieweg Verlag, 2009.

Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications

Mouhib Alnoukari

Business Intelligence (BI) is a set of tools and techniques that can help organizations collect, clean and integrate all their data. Organizations can then analyse, mine and dig deeper into their data in order to make the right decisions at the right time. In this article the author sums up the knowledge and experience gained while preparing the book "Business Intelligence and Agile Methodologies for Knowledge-Based Organizations", one of the first books that focuses on the use of agile methodologies for building business intelligence applications, highlighting the integration of process modelling, agile methods, business intelligence, knowledge management, and strategic management[1] .

Keywords: Agile Methods, Business Intelligence, Knowledge Management, Process Modelling, Strategic Management.

1 Introduction

In 1996, the Organization for Economic Cooperation and Development (OECD) redefined "knowledge-based economies" as "*economies which are directly based on the production, distribution and use of knowledge and information*". According to this definition, data mining and knowledge management, and, more generally, Business Intelligence (BI), should be the foundations on which the knowledge economy is built.

However, Business Intelligence (BI) applications still face failures in determining the process model adopted. As the world becomes increasingly dynamic, traditional static modelling may not be able to deal with it. Traditional process modelling requires a great deal of documentation and reports. This prevents traditional methodology from meeting the ever changing dynamic requirements in our rapidly changing environment.

One solution is to use agile modelling, which is characterized by flexibility and adaptability. On the other hand, Business Intelligence applications require greater diversity in technology, business skills, and knowledge than typical applications, which means they may benefit from features of agile software development.

This field is addressed in the book cited in Footnote 1, which aims at providing added value for its readers for the following reasons:

- Because most organizations are using business intelligence and data mining applications to enhance strategic decision making and knowledge creation and sharing.
- Because data mining is at the core of business intelligence and knowledge discovery.
- Because most current business intelligence applica-

[1] This book, edited by Prof. Asim A. El Sheikh and Dr. Mouhib Alnoukari, will be published by IGI Global in September 2011. <http://www.igi-global.com/requests/details.asp?ID=829>.

Author

Mouhib Alnoukari received his PhD degree from the Arab Academy for Banking and Financial Sciences (Damascus, Syria). He is currently working as an ICT faculty member at the Arab International University, Damascus (Syria) and the Arab Academy for Banking and Financial Sciences, Damascus (Syria). Dr. Alnoukari has (co-)chaired many ICT symposiums and conferences in Syria including the 1st National Symposium of Business Intelligence in Syria (BISY 2010) and MENA ICT Week 2011. His research interests are in the areas of Business Intelligence, Data Mining, Data Warehousing, Agile Methodology, Software Engineering, and Databases in which he has published more than 20 journal and conferences papers. He has also (co-)edited more than 20 ICT books both in Arabic and English languages, including: “Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications” to be published by IGI Global, September 2011. He has also participated in various reference book chapters such as “Handbook of Research on Discrete Event Simulation Environments- Technologies And Applications” by Evon M. O. Abu-Taieh and Asim A. El Sheikh (eds). IGI Global. (2009), and “Business Information Systems: Concepts, Methodologies, Tools and Applications”, edited by Information Resources Management Association, USA, IGI Global, 2010. <m-noukari@aeu.ac.sy>

“ In 1996, the OECD redefined "knowledge-based economies" as '*economies which are directly based on the production, distribution and use of knowledge and information*' ”

> “Business Intelligence (BI) applications still face failures in determining the process model adopted”

tions are not able to meet the ever changing dynamic requirements of our complex environment.

- Finally because knowledge is the result of intelligence and agility.

2 Agile Modelling for Business Intelligence

Traditional process modelling are characterized by rigid mechanisms with a heavy documentation process, which make it difficult to adapt to a high-speed, high-change environment.

The manifesto and practices of agile methods were published in 2001[2] .The core ideals of the manifesto are: individuals and interactions over processes and tools; working software over comprehensive documentation; customer collaboration over contract negotiation; and responding to change over following a plan. Ultimately, by following these ideals, software development becomes less formal, more dynamic, and more customer-focused.

Agile methods share the same properties by focusing on people, results, minimal methods, and maximum collaboration. Agile approaches are best fit when requirements are uncertain or volatile; this can happen due to business dynamism and rapidly evolving markets. It is difficult to practise traditional methodologies in such unstable evolving markets [1].

Business Intelligence applications require greater diversity in technology, business skills, and knowledge than typical applications; this means it may benefit greatly from features of agile software development.

To successfully implement Business Intelligence applications in our agile and knowledge-based arena, different areas should be examined in addition to the consideration of the transition to knowledge-based economy. This book tackles the following business intelligence areas: methodologies, architecture, components, technologies, agility, adaptability, tools, strategies, applications, knowledge and history.

Applying agile methods to Business Intelligence applications is the core idea of our book. Different chapters raised the importance of using such methods by addressing the alignment between Agile principles and BI applications, analysing Agile methodologies and addressing the applicability of BI, reviewing the components and best practices of BI applications, proposing different Agile frameworks for BI applications (ASD-BI, BORM, Agile BI Delivery, etc), and applying the proposed frameworks in various areas, including higher education, e-government, regional management systems, risk management, e-marketing, IT governance, and web engineering.

3 The Knowledge Dimension in Agile Business Intelligence applications

Most experts confuse Knowledge Management (KM) with Business Intelligence. According to a survey conducted by OTR, 60 percent of consultants do not know the difference between the two [2]. We may clarify this confusion by explaining the difference between these two terms. Business Intelligence is a set of all technologies that gather and analyse data to improve decision making. Intelligence in BI notation is often defined as the discovery and explanation of hidden, inherent and decision-relevant contexts in large amounts of data. Whereas Knowledge Management is defined as a systematic process for finding, selecting, organizing, presenting and sharing knowledge in a way that improves organizations' comprehension in a specific area of interest, KM helps organizations gain insight and understanding from their own experience. This means that Business Intelligence is just one of the tools of KM which help organizations extract and share knowledge in order to enhance their competitive position in the market.

Agile methods concentrate on human-based techniques of communicating knowledge such as on-site customers, customer focus groups, daily short meetings, and post mortem sessions. The main focus when applying agile methods is to maximize the knowledge transferred and shared among various stakeholders of business intelligence applications. Knowledge capturing happens informally through the use of principles such as on-site customers and customer focus group. Knowledge sharing among all project stakeholders happens through social activities, such as short meetings and post mortem sessions.

4 Business Intelligence Government Framework

One of the main contributions of this book is a proposed business intelligence governance framework within an e-Government system. The proposed framework is based on an empirical study which demonstrates the importance of

[2] The manifesto is available at <http://agilemanifesto.org/>.

> “Traditional process modelling requires a great deal of documentation and reports. One solution is to use agile modelling, which is characterized by flexibility and adaptability”

> “Business Intelligence applications may benefit greatly from features of agile software development”

of using business intelligence in e-Government systems. It also demonstrates that using BI helps close the gap between business and IT people. This in turn can help planners and policy makers at all levels of government increase e-Government success rates.

5 Business Intelligence in Higher Education

The need for BI to achieve a competitive advantage in higher education has gained momentum in recent years. This is due to many reasons as universities are facing huge competition and they need a better understanding of business forces in order to respond effectively to the already dynamic industry. They also require tools to predict student performance, employment paths, course selection, and need to do cost-benefit analyses, trend analyses, value chain analyses, and so forth, which could be supported by BI applications.

Our main focus was the application of agile methods to a business intelligence application in higher education. One of our book’s contributors presents an ontology-based knowledge management system developed for a Romanian university. The system proves that ontology usage could improve the competency gap analysis at an individual, project and organizational level for project-oriented organizations.

Agile business intelligence has been presented in the Syrian private universities. Different models were proposed to enhance the universities’ competencies. One of the models is built on system theory, by visualizing universities as a system with input, processing, output, and feedback. Other models prove that applying agile business intelligence in higher education would help universities to dig deeper into their various data sources, thereby enhancing their decision-making process, enhancing knowledge sharing, and finally helping them implement and achieve their strategies.

They also propose a BI framework within e-Government systems, which helps facilitate and improve the delivery of e-Government services.

6 e-Government Systems

e-Government systems can benefit from business intelligence by allowing them to deal with heterogeneous and silo systems. This can enable such systems to avoid the use of sophisticated tools in order to obtain the information needed to build stronger government strategies. BI applications can also help e-Government systems by reducing the involvement and dependence of IT staff [3]. Business Intelligence can offer many advantages to e-Government systems such as: a deep understanding of citizens’ needs, increased operational effectiveness, the availability of multiple resources to government planners and decision makers, and the provision of extensive resources to support e-Government projects [4].

7 Knowledge Discovery Process Models

Business Intelligence applications ultimately depend on data mining algorithms. The data mining component is also one of the main steps of knowledge discovery from data. The book provides a detailed discussion on the knowledge discovery process models that have innovative life cycle steps including: Knowledge Discovery in Databases (KDD) Process by Fayyad et al. (1996) [5], Information Flow in a Data Mining Life Cycle by Ganesh et al. (1996) [6], SEMMA by SAS Institute (1997) [7], Refined KDD paradigm by Collier et al. (1998) [8], Knowledge Discovery Life Cycle (KDLC) Model by Lee and Kerschberg (1998) [9], CRoss-Industry-Standard Process for Data Mining (CRISP-DM) [10], Generic Data Mining Life Cycle by (DMLC) by Hofmann (2003) [11], Ontology Driven Knowledge Discovery Process (ODKD) by Gottgtroy (2007) [12], and Adaptive Software Development-Data Mining (ASD-DM) Process Model by Alnoukari et al. (2008) [13].

We also propose a categorization of existing knowledge discovery models. The following are the proposed categories for Knowledge Discovery Process (KDP) modelling: traditional KDP approach, ontology-based KDP approach, web-based KDP approach, and agile-based KDP approach.

The book provides an in-depth analysis of the strengths and weaknesses of the leading knowledge discovery process models, with their supported commercial systems and reported applications, and their matrix characteristics. The main metrics used when evaluating previous KDP models are data, process, people, adaptive, knowledge, and strategy.

8 Risk Management in Knowledge-Based Organizations

Risk management plays a crucial role in our rapidly changing environment. Many projects, especially software projects, have faced serious failures due to not knowing how to deal with the causes of failures. During the last decade, many tools and techniques were used to manage projects risks effectively. Decisions were needed to be made faster in order to address project failures in matters of minutes and sometimes seconds.

We underline the importance of using business intelligence and agile methodologies for managing risks effectively and efficiently.

9 Agile Web Engineering and Business Intelligence

Web-based systems involve "*a mixture between print publishing and software development, between marketing and computing, between internal communications and external relations, and between art and technology*". [14].

> “ Business Intelligence is just one of the tools of Knowledge Management which help organizations extract and share knowledge in order to enhance their competitive position in the market ”

Web-based applications are different from traditional applications as they need to have special features such as usability, loyalty, accessibility, and context.

Most web development methodologies such as OOHDM and WebML focus on designing approaches rather than understanding requirements.

This issue can be addressed by the adoption of Agile methodologies such as eXtreme Programming (XP). These methodologies allow systems to be built incrementally, thereby facilitating feedback from the client as the system develops.

The book highlights the main issues related to agile web engineering practices, the need for web engineering, and the agile development methodologies used in web engineering. The book also covers important topics of Web Engineering, including requirements analysis, design, architectures, technologies, test, operation and maintenance; this is complemented by in-depth knowledge about Web project management and process issues as well as important quality aspects of Web applications such as usability, performance and security.

10 Conclusion

In this article we have briefly summarized the main ideas developed in the book "Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications" (see Footnote 1), one of the first attempts to highlight the importance of using agile methodologies in business intelligence applications. Although, the research orientation is new, the book’s chapters produce very important research outcomes in different areas.

The ideas described in the book create an added value to the field because most organizations are using business intelligence and data mining applications to enhance strategic decision making and knowledge creation and sharing, and data mining is at the core of business intelligence and knowledge discovery. Also, most current business intelligence applications are not able to meet the ever changing dynamic requirements of our complex environment and, finally, knowledge is the result of intelligence and agility.

References

[1] J. Highsmith. Retiring Lifecycle Dinosaurs: Using Adaptive Software Development to Meet the Challenges of a High-Speed, High-Change Environment. Software Testing & Quality Engineering, pp 22-28, 2000.

[2] R. Hershel and N. Jones. Business Intelligence and knowledge management: The Importance of Integration. Journal of Knowledge Management, pp 44-55, 2005.

[3] Gartner. Key Issues for Business Intelligence and Performance Management Initiatives (Research note). 2008. <http://www.gartner.com> [accessed June 18, 2010)].

[4] D.I. Sandu. Operational and real-time Business Intelligence, Revista Informatica Economicã nr.3 (47)/ 2008, pp 33-36, 2008. <http://revistaie.ase.ro/content/47/06Sandu.pdf >.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge Discovery and Data Mining: Towards a Unifying Framework. Second International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp 82-88). USA: AAAI Press, 1996.

[6] M. Ganesh, E.-H. Han, V. Kumar, and S. Shekhar. Visual Data Mining : Framework and Algorithm Development. University of Minnesota Computer Science Technical Report, TR 96-021, 1996.

[7] SEMMA. Data Mining and the Case for Sampling (White Paper). SAS Institute, 1997.

[8] K. Collier, B. Carey, E. Grusy, C. Marjaniemi, and D. Sautter. A Perspective on Data Mining. Northern Arizona University, USA: Centre for Data Insight, 1998.

[9] S.W. Lee and L. Kerschberg. A Methodology and Life Cycle Model for Data Mining and Knowledge Discovery in Precision Agriculture. IEEE International Conference on Systems, Man, and Cybernetics (SMC ’98) (pp 2882-2887). San Diego, CA: IEEE Computer Society Press, 1998.

[10] CRISP-DM. CRISP-DM 1.0 – Step by Step data mining guide. CRISP-DM Consortium, 2000.

[11] M. Hofmann and B. Tierney. Proceedings of the 1st international symposium on Information and communication technologies. ACM International Conference Proceeding Series; Vol. 49 (pp 103 - 109). Dublin, Ireland: Trinity College Dublin, 2003.

[12] P. Gottgtroy, N. Kasabov, and S. Macdonell. An ontology engineering approach for Knowledge Discovery from data in evolving domains. Proceedings of the 2003 SIAM International Conference on Data Mining. San Francisco, CA, 2003.

[13] M. Alnoukari, Z. Alzoabi, and S. Hanna. Applying Adaptive Software Development (ASD) Agile Modeling on Predictive Data Mining Applications: ASD-DM Methodology. International Symposium on Information Technology (pp 1083-1087). Kuala Lumpur, Malaysia: IEEE, 2008.

[14] T.A. Powell. Web Site Engineering: Beyond Web Page Design, Prentice Hall, 1998.

Social Networks for Business Intelligence

Marie-Aude Aufaure and Etienne Cuvelier

Online social networks have been closely studied from sociology at the middle of the 20th century to today's computer science but enterprise social networks are still in infancy. Social networks can improve enterprise organization as well as business applications. This article introduces enterprise social networks and associated use cases, graphs used to model these complex networks and how to analyze the content and structure. We also present a project we are working on to integrate internal and external social networks in a public administration.

Keywords: Business Applications, Complex Networks, Graphs, Mining and Aggregating Graphs, Social Networks.

1 Introduction

We have now entered the era of knowledge. Ubiquitous computing as well as the constant growth of data and information lead to new ways of interaction. Users manipulate unstructured data – documents, emails, social networks, contacts – as well as structured data. They also want more and more interactivity, flexibility and dynamicity. Users expect immediate feedback, and want to find rather than search for. All these evolutions induce challenging research topics for Business Intelligence, such as providing efficient mechanisms for a unified access and model to both structured and unstructured data.

Business Intelligence (BI) has historically been based on a combination of data warehousing, the process of storing historical data in a structure designed for efficient processing, and data mining, the mathematical and statistical methods necessary to transform this raw data into valuable information for making business decisions. The increasing flow of information, called Big Data, implies that BI can no longer afford to focus solely on historical records stored in tabular form. BI is moving to Business Intelligence 2.0, which combines BI with elements from both Web 2.0 (a focus on user empowerment, social networks, and community collaboration), and the Semantic Web, sometimes called "Web 3.0" (semantic integration through shared ontologies to enable easier exchange of data).

Social Networks are a part of this evolution and can be defined as a set of social entities, such as individuals or social organizations connected by links created during social interactions. They correspond to a new form of organization, called Enterprise 2.0, decentralized and more flexible, and viewed as more efficient than traditional hierarchical organizations. Historically, social networks have been first studied from a sociological point of view. Georg Simmel states that the foundation of sociology is defined by the relations and interactions between individuals, and not the individuals themselves. Networks are produced by these interactions. Jacob Moreno used surveys to build a set of social data, and searched for configurations appearing regularly in relations between individuals (analytical usage). Mark Granovetter [1] defined the theory of "power of weak links", these links being the most efficient ones in a profes-

Authors

Marie-Aude Aufaure obtained her PhD in Computer Science from the University of Paris 6, France, in 1992, and her HDR from the University Claude Bernard Lyon 1, France, in 2002. From 1993 to 2001, she was associated-professor at the University of Lyon; then, she has integrated a French Research Centre in Computer Science (INRIA) during two years. She was professor at Supelec from 2003 to 2008. Now, she is full professor at *Ecole Centrale Paris* (MAS Laboratory) and head of the SAP Business Objects Chair in Business Intelligence. She is also scientific partner at INRIA since 2003. Her research interests deals with the analysis, retrieval and querying of unstructured data, and the combination of structured and unstructured data from a Business Intelligence perspective. The scientific topics developed in her team are related to semantic technologies, graphs, conceptual classification with a user-centric point of view and are applied to semantic information retrieval, question and answering over data warehouses, social networks and recommender systems (special focus on user modeling and personalization). She is reviewer for many journals and conferences and has deeply published in the fields of semantic technologies, data mining and databases in international journals, books and conferences. <marie-aude.aufaure@ecp.fr>

Etienne Cuvelier holds his PhD in Computer Science from the Facultés *Universitaires Notre-Dame de la Paix of Namur*, Belgium, since 2009. Prior to this he also obtained a MSc in Mathematics and a MSc in Computer Science and taught mathematics and computer science during several years before integrating the research world. He is now postdoctoral researcher at *Ecole Centrale Paris*, MAS Laboratory, in the Business Intelligence team. His work is related to the data mining and the social network analysis, and more specifically in functional data analysis, symbolic data analysis, conceptual analysis and graph mining. <etienne.cuvelier@ecp.fr>

> “Enterprise social networks are still in infancy but they can improve enterprise organization as well as business applications”

> “Public Administrations also need social networks as an interface so that themselves and citizens can easily understand who does what and who says what”

sional context. Milgram [2] established in 1967 the theory of six degrees of separation related to the small-world phenomenon, which is the hypothesis that the chain of social acquaintances required to connect one arbitrary person to another arbitrary person anywhere in the world is generally short.

As with other networks, social networks can be represented as a graph: a set of nodes linked by edges representing any kind of relationship between nodes. Social networks have been deeply studied for the web (online social networks), but they are also a key element in enterprises. These enterprise social networks are more complex than online social networks like Facebook, LinkedIn or Twitter, because they do not only connect people but also all the objects manipulated in an enterprise like projects, products, etc. In the enterprise context, two main reasons lead to a strong interest for social networks: (1) a technology for internal communication, collaboration and information sharing, and (2) a technology for communicating to clients, citizens, etc. Recent studies (for example those by Singh and by Nieto et al.) have shown a growing interest in transversal collaboration leading to increase the value of collaborators' competencies, and to facilitate innovation and productivity. This form of organization is generally not easy to introduce because of the organizational structure which favors intra-team and hierarchical exchanges and also the fact that sites can be geographically distributed. Social networks are adapted to build virtual communities based on common interests for example, and can be seen as a potential solution to the above mentioned problems. According to a recent study done by Coleman Parkes Research in 2008 among 500 companies around the world, only 7% of these companies use a social network and 4% are integrating a social network, while 59% declare having no strategy for having an internal social network. This study also outlined that 60% of these companies admit that social networks will be a fundamental tool in the future for collaboration in the enterprise. The main obstacles to adopting this technology are related to the security issues (76%), the inaction of the direction (57%) and the hesitation to exploit new technologies (58%).

Why are social networks of interest for companies? This is a technology for (1) internal communication, information sharing and collaboration, (2) information communication towards clients, (3) watching the gossip about the company (e-reputation, opinion mining) and (4) creating collective intelligence.

Social networks can be internal (model of the organization, social interactions between employees, etc.) and external (like Twitter than can be used for e-reputation).

In the next sections, we will define use cases social networks in an enterprise context and also for public administrations, the variety of graphs that can be used to model these networks, and how is it possible to analyze social networks.

2 Social Networks Use Cases

This section shows how social networks can be used in public administrations and in enterprises. Public administrations and enterprises have similar problems that can be partly addressed by the use of social networks. They need to manage the organization and to deliver services to citizens. Such services cover various domains such as highways maintenance, urban planning, assistance to persons, etc. Public administrations need social networks as an enterprise to analyze their internal networks (projects, organization etc.) and to analyze their external networks (suppliers, clients, partners). They also need social networks as an interface so that citizens and the Administration can easily understand who does what and who says what. We are implementing such scenarios in a project we have with a public administration (ARSA project with the city of Antibes in France). We are working in this project on internal and external social networks. The internal social network can be seen as an extension of the intranet and the main objective is to enhance transversal collaboration. The need expressed is to visualize and to share information and data, to model and visualize collaborative project and to navigate through the hierarchical structure of the organization. The objective of the external social network is to monitor what is said by citizens about the city through Twitter. The idea is to be able to visualize in real time citizens opinions and to give immediate feedback about the actions done by the administration. We will show in the social network analysis section how we monitor Twitter for defining e-reputation in real time.

Scenarios for enterprises can be organized around the centrality [3], importance and influence of actors, the identifications of groups and the identification of key actors, from a human resource, management and individual's perspective. Being central is: be the source or destination of numerous relations (degree centrality), be close to many actors (closeness), be central for many connections (betweeness). Many *scenarios* for using social networks are useful for human resources. Finding persons having the biggest influence can help to transmit good practices and improve social aspects. Identifying the most central groups helps in finding groups with a good communication, which are important elements of cohesion. The similarity between groups can be computed in order to analyze groups with "good" properties, apply observation on a group

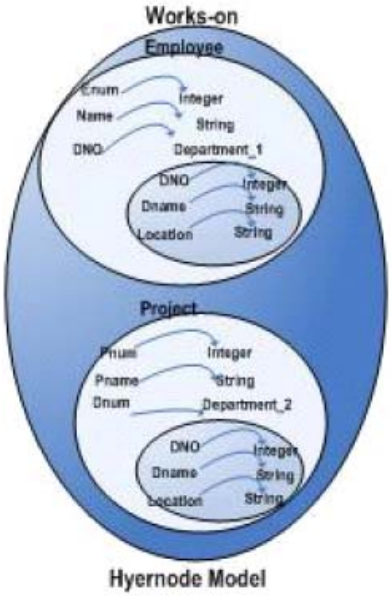

Figure 1: Example of an Hypernode Schema.

to a similar one or explain a behavior within a structure. We can also search for atypical elements, like persons who are not well integrated, and find the best group for such persons. One other application is to be able to constitute efficient teams, in order to enable innovation and to reach group cohesion. From an individual perspective, such tools can help for networking, to identify for example important relations by studying relations of other persons. Individuals can create a collaborative social network, learn from the experience of others, find an expert in the company or build a profile for finding a new job.

Applications related to this technology become more and more important, as well as the size of the resulting social networks. We need methods to be able to model (using graphs) such complex networks, to analyze and query these networks and to be able to visualize results in an efficient and intuitive way.

3 Variety of Graphs

A wide variety of graphs can be used to model social networks, from the basic mathematical definition to more complex variations. A graph $G = (V, E)$ consists of a set V of vertices (also called nodes), a set E of edges where $E \subseteq V \times V$. This definition refers to simple and undirected graphs. The term *multigraph* is generally understood to mean that multiple edges and loops are allowed. A *graph labeling* is the assignment of labels, traditionally represented by integers, to the edges or vertices, or both, of a graph. Labeling is applied to finite undirected simple graphs. A graph can also contain more information by adding attributes to nodes or edges. Such a graph is termed an *attributed graph*. Simple graphs are not sufficient to model heterogeneous graphs based on complex objects, having multiple attributes and relations. Then, the basic structure of a graph (nodes and edges) is complemented with the use of hypernodes and hypergraphs extensions that provide support for nested structures. A *hypergraph G* is a tuple (V, E, μ), where V is a finite set of nodes, E is a finite set of edges, $\mu : E \rightarrow V^*$ is a connection function where V^* means multiple nodes (an edge can connect any number of nodes). An Hypernode can encapsulate nested nodes (see Figure 1).

A variety of graph database models [4] has also been defined. All these models have their formal foundation as variations of the basic mathematical definition of a graph. The main characteristics of these models is that the schema and instance levels are distinct, like in RDF graphs, and can easily be used to model business applications. Another characteristic is that these models encapsulate the semantic in the nodes and edges.

When graphs of extracted social networks are large, effective graph aggregation [5][6] and visualization methods are helpful for the user to understand the underlying information and structure. Graph aggregations produce small and understandable summaries and can highlight communities in the network, which greatly facilitates the interpretation. Graph aggregation differs from other methods such as graph mining, based on the graph structure (edges) and graph clustering which groups similar nodes. The aggregation is a summary based on nodes content and neighborhood.

4 Social Networks Analysis

If Social Network Analysis [7] was firstly developed by sociologists, the rise of computer science and one of its fields, data analysis, enables analysis of graphs of big sizes, and permits the development of several interesting but computationally intensive tasks.

One of these is the community detection task [8]. If the notion of community seems very natural for most people, it demands a formal, and most of the time measurable, definition to be then computed. And even if there exist many definitions of what is a community in a social network, we can say, without lack of generality, that a community is a group of individuals with high concentrations of relationships within the group and low concentrations of relation-

> “As with other networks, social networks can be represented as a graph: a set of nodes linked by edges representing any kind of relationship between nodes”

with individuals outside the group. The detection of such groups in a social network can be performed using several kinds of algorithms. The choice of the algorithm depends on many criteria, but a decisive one is often the "size" of the network (the number of nodes and the number of links), because the larger the network, the more expensive in computing time, is its execution. For the sake of illustration we give here a sketch of the ideas of some of these algorithms. The "cheapest" algorithm family is the partitional ones which try to find a partition of the individuals in *a priori* given number of clusters equal to k. The "best" partition is searched using jointly, most of the time, a distance measure and a quality criterion of the found partition based on the sum of the distances of individuals to the centre of the cluster. The most popular partitional algorithm (with several variants) is the k-means clustering. Another interesting type of algorithm is the family of hierarchical clustering algorithms which are divided into two types, depending on whether the partition is refined or coarsened during each iteration: agglomerative algorithms start with a set of small initial clusters and iteratively merge these clusters into larger ones, while divisive algorithms start with all the network as one big group, and then split the dataset iteratively or recursively into smaller and smaller clusters. At each step, the clustering algorithm musts select the clusters to merge or split by optimizing a quality criterion. Several other algorithms exist.

From a business intelligence point of view, as in the case of tabular data, finding "homogeneous" groups in networks allows the study of each of the found groups in order to know their characteristics, which can be valuable knowledge in a customer relationship framework for example. From a media diffusion point of view, finding such groups can be also very interesting, not only to study the characteristics of these groups, but also to find two types of majors actors in the diffusion process: actors with the most connections with other people and actors which link two or more communities. The first type of actor, those with the highest degree of centrality, can be seen as the most popular persons in their groups - thus they became preferential and economical targets for marketing actions dedicated to adoption of new products, for propagation of information and advices, and also for monitoring opinions and mood of customers. The second type of actor, those linking communities, and which have the highest betweeness, are key actors for the spreading of information through a network because they permit the transfer of information from one community to another and they are switches, which can be used in positive or negative ways, if we wish to favor diffusion or not. This latter type of actor can be easily retrieved using, for the sake of illustration, the algorithm hierarchical divisive algorithm of Girvan and Newman [7] which attempts, successfully, to detect communities in finding the bridges between the communities. Indeed, this algorithm also permits the detection of these path between communities. Figure 2 gives a picture of three communities in the set of the Facebook friends of one of the authors of this paper. These communities, detected using the algorithm of Girvan and Newman, are completely meaningful, according to the prior knowledge of the author, because we retrieve the community of its family (nodes colored in blue), the communities of the ex-students of two universities where he taught (in red and green). And we see clearly the "bridge actors" in this figure.

> “A wide variety of graphs can be used to model social networks, from the basic mathematical definition to more complex variations”

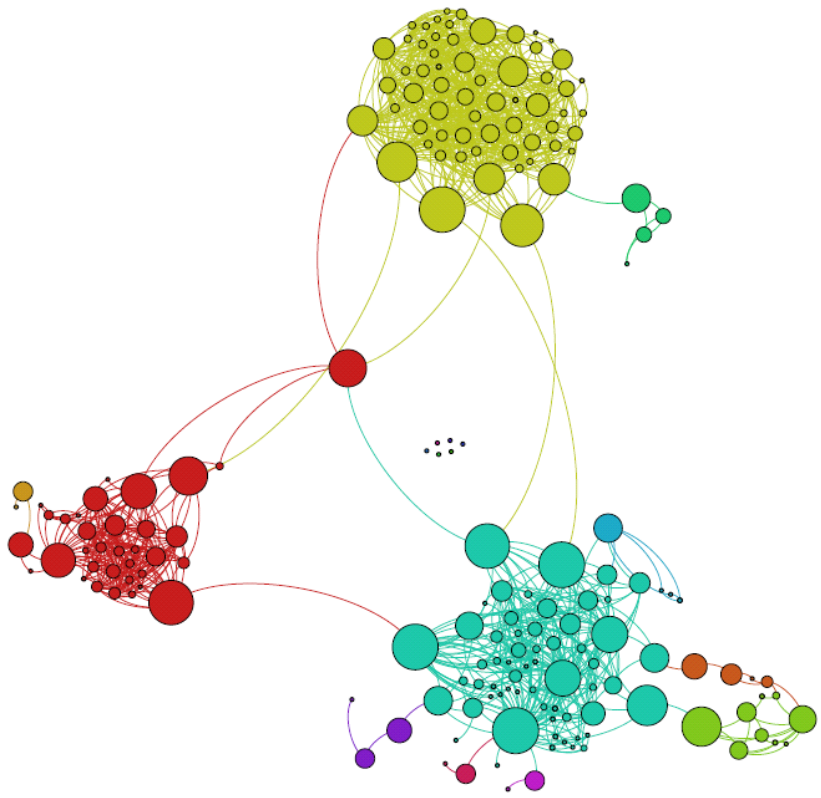

Figure 2: Communities in a Real Facebook Profile.

Detect and find communities in social networks is not the only interesting task. As already explained in the "social networks use cases" section, it could also be very valuable to study what kinds

> “The notion of e-reputation arises: ‘what is the standing of me or my society, right now on the Net?’ ”

of information cross a given network. Nowadays most online social networks are used for sharing information, mood and advice. However, very quickly following the massive adoption of such networks and practices, there arose the notion of e-reputation: "what is the standing of me or my society, right now on the Net?" Even if this e-reputation or branding is something to be built patiently, day after day, as the permanent result of an active presence on the web and social networks, any bad buzz can very quickly ruin these efforts if there is not a rapid detection and adequate reaction to such phenomenon. Efficient tools to monitor their own e-reputation become urgently needed by enterprises or official institutions in this high-speed interconnected world. A lot of such new services appears every month on the web, but mostly they use simple queries and classical statistics and then don’t give a global summary view. In [9] we have proposed a prototype of an e-reputation monitoring tool used on the Twitter network. This latter social network is one of the most used to share information, has a large audience (200 million users in April 2011,[10]), is fast growing (460,000 new accounts per day in average during February 2011, [11]) and manages a huge quantity of exchanged information, called tweets (140 millions tweets sent per day, [10]). Each piece of generated information can be forwarded, but can also be edited before the forwarding process, it then becomes a real challenge to trace the path and transformations of a "successful" tweet (i.e. a buzz).

This challenge is however crucial for an e-reputation tool. Our prototype proposes to retrieve all the groups of words that are the most forwarded on a given subject. This tool called EVARIST is developed in the framework of the ARSA project mentioned above in collaboration with the French city of Antibes. EVARIST is based on a mathematical tool called Galois lattice: briefly, if we have a set of objects and a set of attributes, with a Boolean table (called context table) such a value TRUE is given at cell (j, k) if the object j owns the property described by the attribute k, then the Galois lattice is the structure which gives all the subsets of object and attributes (called concepts) such all the objects share the same attributes. The graphical representa-

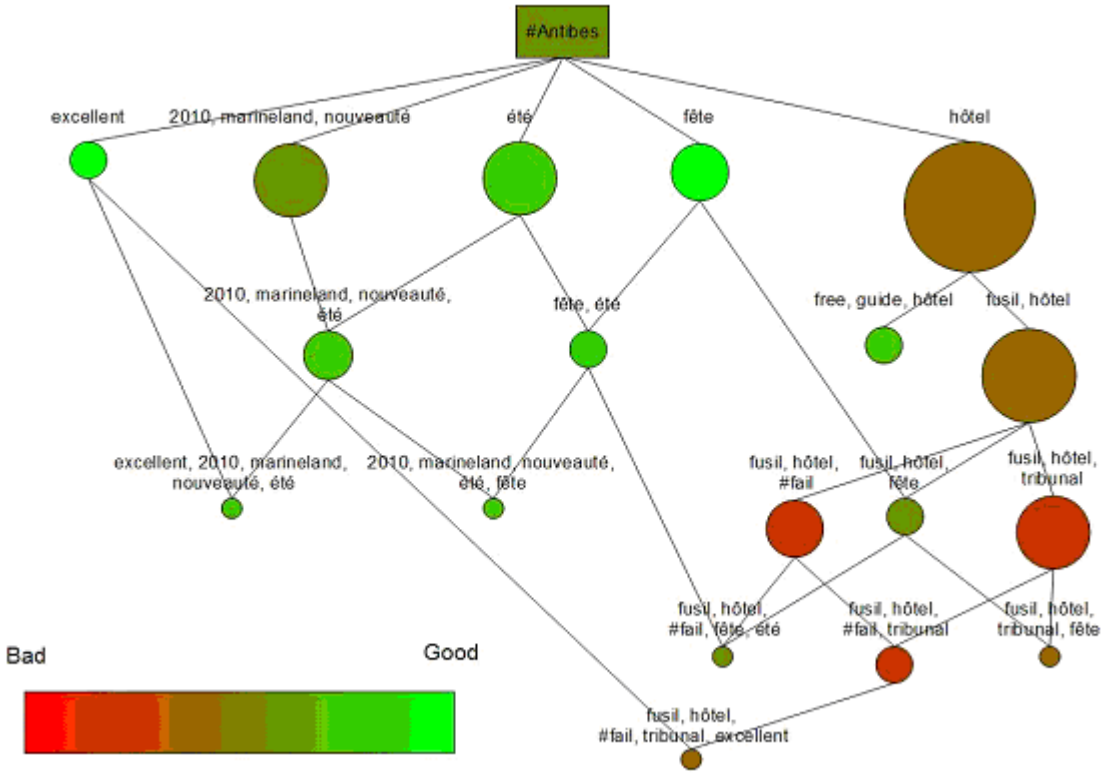

Figure 3: The EVARIST Prototype.

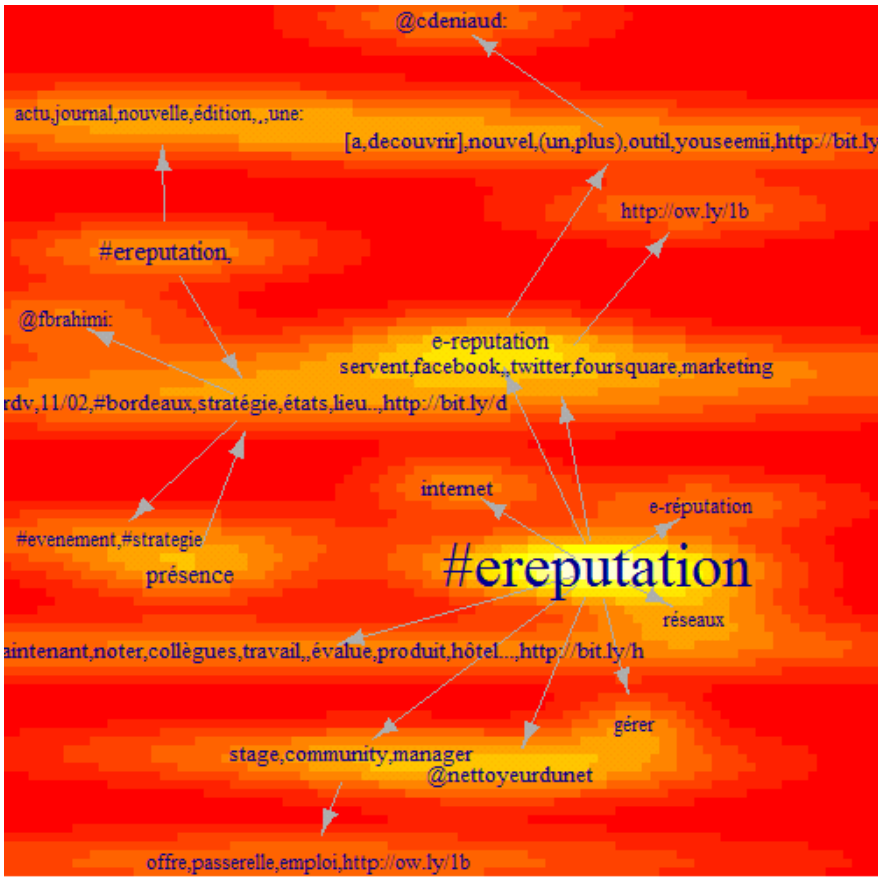

Figure 4: An EVARIST Visualization on e-Reputation.

tion of a Galois lattice is made using an Hasse diagram which show all the concepts and their inclusions.

Figure 3 shows a Galois lattice produced by EVARIST. This tool use the following steps:

1. Getting the tweets including a chosen word or group of words,
2. Cleaning the tweets (suppressing stop words, punctuations,...);
3. Stating the table of context with the tweets as objects and the words as attributes;
4. Building the corresponding Galois lattice and give to each concept a size proportional to the number of tweets owned;
5. Visualization of the results.

In Figure 3, the result of an execution of this tool for monitoring what is told on twitter about the French town is shown. In this figure, starting from the top we can see how the group of tweets can be subdivided in subgroups, and which words are shared by all the tweets of the subgroup. The size of the concept being proportional to the portion of tweets contained in the subgroup, the figure permits retrieval of the most used words in the buzzed tweets about Antibes,. A sentiment analysis process is applied to each tweet in order to determine if this latter gives a positive advice on Antibes or not. Given the color of the concept, the community manager can have also a quick look about any possible bad buzz for his town.

Figure 3 demonstrates that it can quickly become difficult to display all the concepts proportionally to their sizes, and at the same time display clearly all their attributes if the number of tweets and words increases. To reduce the number of attributes and concepts to be displayed, we can select only the concepts with a relative size greater than chosen threshold. In other words, in respect to the notion of buzz we can select the concepts with the more tweeted words. Finally, to reinforce a reading going from the most

> “Efficient tools to monitor their own e-reputation become urgently needed by enterprises or official institutions. We have developed the EVARIST tool to this end”

general to the most particular, we have proposed to add a topographic allegory called "topigraphic" network of tags. To do this, for each point of the resulting graphic, we add a level, these levels being pictured using the classical level curves. Such a "topigraphic" map is shown in Figure 4, but without the sentiment analysis result.

5 Conclusion

Integrating social networks in enterprises and public administrations is of real interest, but somehow difficult to implement. Two implementation solutions can be considered. The first is to use an existing social network, like LinkedIn for example, and create a new group. This solution is very easy to implement but has several drawbacks: you have no access to the metrics (i.e. evolution of the number of participants), you do not own data (privacy issue), you cannot link the social network to knowledge management tools and you are dependant on a business model that is not yours!. The other solution is to use a commercial or an open source tool. In that case, you have the total control of metrics, data, content published and you can also link the tool to existing social networks. The drawback is that the implementation is not always easy. In our project with the city of Antibes, we worked with SNA[1] (Social Network Analyzer), a tool developed by SAP. The idea was to allow a user to interact from an external social network to the internal one with respect to the access rights, and conversely, to allow the public administration to be aware of the information spread over the external social networks (Twitter in this project), information that can be useful for the city.

References

[1] M. Granovetter. "Introduction for the French Reader," *Sociologica* 2: 1–8, 2007; B. Wellman. "Structural Analysis: From Method and Metaphor to Theory and Substance", Pp. 19-61 in *Social Structures: A Network Approach*, edited by Barry Wellman and S.D. Berkowitz. Cambridge: Cambridge University Press, 1988.

[2] S. Milgram,. «The Small World Problem». *Psychology Today*, 1(1), May 1967. pp 60 – 67.

[3] L.C. Freeman. A set of measures of centrality based upon betweenness. Sociometry, 40 35–41, 1977.

[4] R. Angles, C. Gutierrez. Survey of graph database models. ACM Comput. Surv. 40 (2008) 1-39.

[5] Y. Tian, R.A. Hankins, J. M. Patel. Efficient aggregation for graph summarization. In SIGMOD '08 : Proceedings of the 2008 ACM SIGMOD international conference on Management of data, New York, NY, USA, (2008) 567–580.

[6] R. Soussi, E. Cuvelier, M.-A. Aufaure, A. Louati, and Y. Lechevallier. DB2SNA: an All-in-one Tool for Extraction and Aggregation of underlying Social Networks from Relational Databases, 2011. In: The influence of technology on social network analysis and Mining, Tansel Ozyer *et al.* (eds.), Springer, to appear.

[7] J. Scott. *Social Network Analysis*. Sage, 2000.

[8] M. Girvan and M.E.J.Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99:7821–7826, 2002.

[9] E. Cuvelier and M.-A. Aufaure. A Buzz and E-Reputation Monitoring Tool for Twitter based on Galois Lattices, the 19th International Conference on Conceptual Structures (ICCS 2011), 25-29 july 2011, University of Derby, United Kingdom.

[10] B. Bosker. Twitter: We Now Have Over 200 Million Accounts (UPDATE), Huffpost Tech, 04/28/11, <http://www.huffingtonpost.com/2011/04/28/twitter-number-of-users_n_855177.html>.

[11] Twitter, #numbers, <http://blog.twitter.com/2011/03/numbers.html>.

[1] A demo is available at <http://sna-demo.ondemand.com/>.