THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

DATABASE SYSTEMS JOURNAL

Vol. VI, Issue 3/2015

LISTED IN

RePEc, EBSCO, DOAJ, Open J-Gate,
Cabell's Directories of Publishing Opportunities,
Index Copernicus, Google Scholar,
Directory of Science, Cite Factor,
Electronic Journals Library

BUSINESS INTELLIGENCE

ERP

DATA MINING

DATA WAREHOUSE

DATABASE

ISSN: 2069 – 3230
dbjournal.ro

Contact

Calea Dorobanţilor, no. 15-17, room 2017, Bucharest, Romania
Web: http://dbjournal.ro
E-mail: editordbjournal@gmail.com; editor@dbjournal.ro

CONTENTS

Exploring Data in Human Resources Big Data 3
Adela BÂRA, Iuliana ȘIMONCA (BOTHA), Anda BELCIU, Bogdan NEDELCU

Electricity tariff systems for informatics system design regarding consumption optimization in smart grids 11
SimonaVasilica OPREA

Data integration approaches using ETL 19
Alexandra Maria Ioana FLOREA, Vlad DIACONITA, Ramona BOLOGA

Big Data, the perfect instrument to study today's consumer behavior 28
Cristina STOICESCU

Exploring Data in Human Resources Big Data

Adela BÂRA, Iuliana ȘIMONCA (BOTHA), Anda BELCIU, Bogdan NEDELCU
Academy of Economic Studies, Bucharest
bara.adela@ie.ase.ro, iuliana.botha@ie.ase.ro, anda.velicanu@ie.ase.ro,
bogdannedelcu@hotmail.com

Nowadays, social networks and informatics technologies and infrastructures are constantly developing and affect each other. In this context, the HR recruitment process became complex and many multinational organizations have encountered selection issues. The objective of the paper is to develop a prototype system for assisting the selection of candidates for an intelligent management of human resources. Such a system can be a starting point for the efficient organization of semi-structured and unstructured data on recruitment activities. The article extends the research presented at the 14th International Conference on Informatics in Economy (IE 2015) in the scientific paper "Big Data challenges for human resources management".

Keywords: Big Data, Business Intelligence, NoSQL Databases, Data Mining, Cloud Computing

1 Introduction

In the context of social networks development and ICT challenges, human resource recruitment and selection issues in multinational organizations is becoming more complex. At this level, flow of information, data and knowledge comes from multiple sources with various systems leading to a major effort in the process of extraction, integration, organization and analysis of data for decision-making recruitment. Also conducting the selection process cannot be performed effectively by studying profiles, resumes and recruitment sites which presents subjective heterogeneous information. The paper aims to present intelligent methods for making the best decisions in human resource selection using Big Data technologies, optimization techniques and data mining. The solutions will allow automatic acquisition of information about applicants in recruitment sites, personal web pages, social networks, websites and academic centers and will enable decision making using intelligent optimization methods. Research motivation stems from the fact that, in the current global economic crisis, making effective decisions on recruitment is a key factor for companies.

Technologies for organizing and processing large volumes of heterogeneous data, unstructured and characterized by a high velocity is in an exponential growth. The amount of data managed by different recruitment companies available over the Internet on social networks generates Big Data problem. We use intelligent methods for analyzing such data in order to obtain a competitive advantage in recruitment and thus in business development.

2. Processing HR data from heterogeneous sources

Currently, information on supply and demand in the labor market is stored electronically as CVs in the form of text databases. These semi-structured data typically come from portals and recruitment sites. But there is a huge amount of information on social networks, collaborative platforms of universities and specialized forums. This data is unstructured. In order to use both the semi-structured and unstructured data, it is necessary to use the methods and techniques of parallel processing, extraction, cleansing, transformation and

integration in a NoSQL database. The difficulty of the problem in this case is to analyze and identify solutions and technologies for Big Data that can be applied for organizing and processing.

For data analysis, data mining methods can determine patterns and profiles for optimal recruitment strategy. But traditional data mining techniques are inadequate for the volume of data. In most cases, only a small part of all available documents will be relevant for a particular candidate. In this case, the difficulty is in identifying and implementing the algorithms for data mining and text mining to compare and rank the documents in order of importance, relevance and determination of profiles of candidates for recruitment.

Due of the complexity of the technologies to be used, and the rapid changes in the labor market, the creation of an architecture that enables the introduction of new data sources, that is capable of integrating multiple and heterogeneous sources, that includes a level of complex models analysis and determination of profiles and lead to the creation of a knowledge-based management of human resources. From this point of view, the difficulty lies in choosing the elements and builds a platform enabling efficient parallel processing, extracting timely information, interactive data analysis and satisfy performance requirements imposed by the paradigm Big Data Analytics.

Set in a rapidly growing number of impressive data collected and stored on the Internet on the availability of human resources has exceeded the human ability to understand without the help of powerful tools. Thus, instead of being based on relevant information, important decisions are made intuitively concerning recruitment, subjective or based on fixed criteria, without taking into account the complexities of nature and human behavior. To obtain relevant information methods such as multivariate analysis should be used for data processing, data mining, statistical methods and mathematical methods that can be applied to large data volumes. For these applications, the data must be well organized and indexed so as to provide ease of use and easy retrieval of information. Recent studies oriented towards organization and processing data from recruitment portals [1], [2] refers to the importance of this analysis for the selection process and the impact that these techniques have on business performance.

Regarding the determination of the profiles of candidates, there are studies published in [7] and [8] concerning the application of data mining algorithms (decision trees, association rules, clustering) for selection of candidates and determine methods of training for staff recruited. However, these studies do not account for data from social networks and collaborative platforms, from sources such as universities or forums. Processing of text information and application of data mining techniques on data from these sources are taken into consideration more and more. We have developed numerous methods of text mining, but usually they are oriented selection of documents (where the query is considered as a provider of constraints) or the assessment documents (where the query is used to classify documents in order of relevance) [3]. The goal is to retrieve keywords from a query of the text documents and evaluation of each document depending on how much satisfies the query. In this way is evaluated the relevance of a document to the query performed. Another method of classifying documents is the vector-space model presented in [5] and [6]. It involves representation of a document and query vectors and the use of a measure as an appropriate similarity to determine the suitability of the query vector and document vector. Automatic classification is an important point in text mining, because when there are a large number of documents on-line, the possibility of automatic organization of these into classes

to facilitate retrieval of documents and analysis is essential.

For software development, there are now business intelligence technologies that can be used. Also, current developments in information technology have led to the emergence of concepts and new ways of organizing and processing systems in order to improve access to data and applications organizations. Cloud Computing architecture that computing power, databases, storage and software applications coexist in a complex and complete network of servers that provides users with information as a service, accessible via the Internet using mobile devices. Such a flexible architecture that allows the connection of several types of subsystems can be used to create a platform for recruitment. There are also Big Data platforms available in cloud computing architecture that can be used and adapted to prototype realization set.

3. Big Data Solutions

When the structure of data seems randomly designed (variety), when the speed of the flow of data is continuously increasing (velocity), when the amount of information is growing each second (volume) and when there is additional information hidden in the data (value), only one solution can be assigned to manage this chaos: big data. This syntagma has been so much promoted by the big software companies, that it seems no software solution is no longer viable if it has no big data capabilities. The truth is there are some domains like telecommunications, social networks, human resources, etc. that are specifically predisposed to the four V (variety, velocity, volume, value). Of course, not only the domain matters. It depends if the data is historical or not, if it's supposed to be continuously analyzed, if it's involved in decision making processes, if it's strategic or secret, if it's structured, semi-structured or unstructured etc.

Big Data represents a technology of a new generation, with a new architecture, designed to extract valuable information from a large data set composed of different sources and having a high data generation flow.

The most obvious feature of big data is its volume. More and more people are using smart devices that are connected to an Internet network and they are producing data each second. The data is growing visibly from big to huge volume. Science has now a solid ground of data for making all sorts of assumption based on the data received from patients, clients, athletes, etc. It's a paradigm that involves our whole universe in gathering, processing and distributing the data. It is important to benefit from this flow of data, by storing it properly using big data solutions.

When creating a business strategy, the main reason why you should choose NoSQL databases is for better performance, scalability and flexibility.

As source [9] states, the two most used Big Data solutions are Cassandra and HBase. Cassandra is the leader in achieving the highest throughput for the maximum number of nodes [10]. Of course, there is a reverse to it: the input/output operations take a lot of time. Cassandra is released by Facebook. HBase is part of the Apache Hadoop project and has the support of Google, being used on extremely large data sets (billions of rows and millions of columns).

Examples of Big Data processing can be found in many domains like social media [13] or insurance fraud detection [14]. Other typical fields where Big Data is issued are Telco, supermarkets, medicine, energy generators [15], sports, etc.

Riak, HBAs Apache, MongoDB and New4J are just a few examples of NoSQL databases. It is particularly important when creating a business strategy, to understand the capabilities and constraints of each type of database in order to choose the most appropriate to achieve the objectives.

Although NoSQL databases do not use schemes, they are fast and adapt easily to the needs of the companies. For example,

NoSQL can work with non-relational distributed and unstructured data, which is the type of data that it's generated by most companies.

In the past, companies have been using relational databases to store their structured data. At present, despite the enormous impact on the world of databases and unlocking data for many applications, relational databases lacks the necessary features to meet faster data transactions in the big data era. NoSQL databases are the answer that solves many of these problems because they offer a new perspective on the world of databases.

The modern technology allows efficiently storing and querying the big data sets, and the emphasis is on using the whole data set and not just samples [11]. Big Data comes hand in hand with analytics, because the final purpose of collecting the huge amount of data is to process and analyze it in order to gain information, value. Analytics don't work directly on data. Data has to be extracted from the database using a specific language and then pass it to analytical tools.

Up until Big Data, the best way to query data from databases was the SQL language, which was specific for structured relational tables. When data began to be hold in NoSQL databases, SQL became only additionally used in queries. For example, the joins are not available in NoSQL queries. One above the other, it was recently stated (September 2014), that SQL is more important that was thought for Big Data, Oracle releasing Big Data SQL, which extends SQL to Hadoop and NoSQL. This road is only at the beginning.

4. The Cloud

The cloud is not simply the latest fashionable term for the Internet. Though the Internet is a necessary foundation for the cloud, the cloud is something more than the Internet. The cloud is where you go to use technology when you need it, for as long as you need it, and not a minute more. You do not install anything on your desktop, and you do not pay for the technology when you are not using it.

NIST (National Institute of Standards and Technology) defines cloud computing as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." [12]

Cloud computing has experienced a fast growth during the last years, and it is expected to keep developing more and more.

Cloud computing can be considered primarily as a cost-saving technology that's used here and there on cost-cutting projects and for quick fixes to provide point solutions to specific operational problems. On the other hand, cloud computing can be understood in the context of an overall business strategy based on agility and responsiveness. Cloud computing certainly provides cost savings in some situations, but cost savings is not the most important benefit. The real value of cloud computing is the way in which it can be used to support an overall strategy designed to create agility for the business.

There are plenty of good reasons, for which you should consider moving to the cloud, but mainly it makes good business sense: cloud computing lets you focus on what's important, your business. This is called efficiency. This field can be used for almost all types of applications and it is clear that it saves its users money.

First of all, the hardware is fully utilized. Cloud computing brings natural economies of scale. The practicalities of cloud computing mean a high utilization and smoothing of the inevitable peaks and troughs in workloads. Sharing sever infrastructure with other organizations, allows the cloud-computing provider to optimize the hardware needs of its data centers, which means lower costs for business.

Secondly, when you run your own data center, your servers won't be fully. Idle servers waste energy, so a cloud service provider can charge you less for energy used than you're spending in your own data center. In conclusion, power costs are lower. When you run your own servers, you're looking at up-front capital costs. But in the world of cloud-computing, financing that capital investment is someone else's problem. Sure, if you run the servers yourself, the accounting wizards do their amortization magic which makes it appear that the cost gets spread over a server's life. But that money still has to come from somewhere, so it's capital that otherwise can't be invested in the business—be it actual money or a line of credit. Moving to the cloud will save you money, not just for your cloud security needs, but for many other types of data center workloads.

Overall, one of the major benefits that the companies can gain by using the cloud, comes not from cost savings for IT resources on a per-use basis, but from the revenue they earn by becoming more flexible and responsive when it comes to customers' changing needs. This would further enable businesses efficiently deliver their new products and services as well as expand successfully into new markets.

5. Proposed architecture

The proposed architecture for a HR recruitment system can be structured on three levels: data, models, interfaces. For each of these levels, the following methods and techniques can be used:

- for the data level, the system uses technologies that collect and process data from web sources, parallel processing algorithms and data organization NoSQL databases;
- the model level uses methods and algorithms for text mining and data mining to build candidates profiles;
- the interface level to achieve online platform uses tools based on business intelligence (BI).

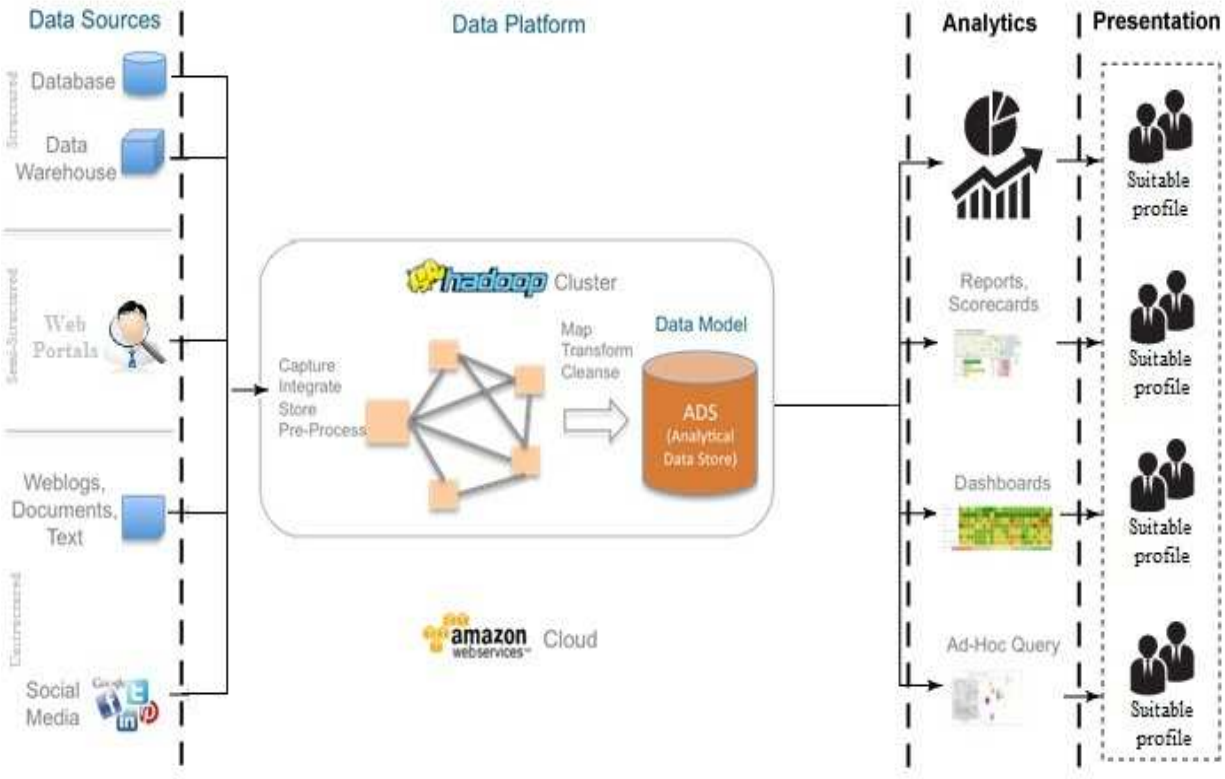


Figure 1. HR recruitment architecture

The impact of a HR recruitment system consists of: facilitating access to relevant information substantiating managers recruitment decisions; minimizing the time for the selection process through easy access to information and its synthesis; increase the information's relevance that reaches decision makers. The implementation of such a system provides a competitive advantage in terms of personnel selection which brings added value to the company and will have a major impact in the following ways:

- from an economical point of view - online platform developed on Cloud Computing architecture can lead to a more easy organization activity within human resources recruitment. By using the prototype it facilitates access to data, reduces the amount of information that reaches decision factors thus minimizing the time for recruitment decisions by easy access to information and profiles by using templates. The results of the development platform can be applied directly in the economic environment;
- in social terms - the main beneficiaries of the prototype are managers and candidates. By using an online scalable platform, company managers can directly select the candidates and increase the efficiency of the recruitment process so that future employees will add value to the company. Also, candidates will be able to publish details of experience, training, social and cultural relations directly through the online platform, providing links or documents without having to complete CV models for each type of job in the offer;
- in terms of the environment - using a scalable architecture such as Cloud Computing, companies will no longer invest in their own hardware, reducing acquisition costs, energy consumption and climate of the data center, minimizing environmental impact.

Conclusions and future work

The HR recruitment system can be developed on a flexible architecture of Cloud Computing so that it can be re-configured for other users by including training and personnel management services.

Determining candidates profiles and templates to characterize their profile can be further improved by introducing new items of interest for recruitment process.

NoSQL is increasingly seen as a viable alternative to relational databases, and should be considered especially for interactive web and mobile applications.

Cassandra or HBase seem the most proper solution for this BigData situation that requires analysis of a large volume of data regarding human resources in order to obtain profiles.

Even if not many people know about cloud computing, it became popular in the latest years. Also, famous companies like IBM adopted or created their own cloud. The major advantage is probably the effects on costs. You can save money, time, even work from home. Cloud computing services still have some disadvantages that stop many companies from adopting it, such as security and data confidentiality.

Acknowledgment

This paper presents some results of the research project: *Sistem inteligent pentru predicția, analiza și monitorizarea indicatorilor de performanță a proceselor tehnologice și de afaceri în domeniul energiilor regenerabile (SIPAMER),* research project, PNII – Collaborative Projects, PCCA 2013, code 0996, no. 49/2014 funded by NASR.

References

[1] C.Nermey - How HR analytics can transform the workplace, http://www.citeworld.com/article/2137364/big-data-analytics/how-hr-analytics-can-transform-the-workplace.html, 2014

[2] eQuest Headquarters - Big Data: HR's Golden Opportunity Arrives, http://www.equest.com/wp-content/uploads/2013/05/equest_big_data_whitepaper_hrs_golden_opportunity.pdf , 2014
[3] C.Győrödi, R.Győrödi, G.Pecherle, G. M. Cornea - Full-Text Search Engine Using MySQL, Journal of Computers, Communications & Control (IJCCC), Vol. 5, Issue 5, December 2010, pag. 731-740;
[4] D.Taniar - Data Mining and Knowledge Discovery Technologies, IGI Publishing, ISBN 9781599049618 (2008);
[5] A.Kao, S. Poteet - Natural Language Processing and Text Mining, Springer-Verlag London Limited 2007, ISBN 1-84628-175-X;
[6] A. Srivastava, M.Sahami - Text Mining: Classification, Clustering, and Applications. Boca Raton, FL: CRC Press. ISBN 978-1-4200-5940-3;
[7] H. Jantan, A. Hamdan, Z. Ali Othman - Data Mining Classification Techniques for Human Talent Forecasting, Knowledge-Oriented Applications in Data Mining, InTech Open, 2011, ISBN 978-953-307-154-1;
[8] L.Sadath - Data Mining: A Tool for Knowledge Management in Human Resource, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-2, Issue-6, April 2013;
[9] O'Reilly Media - Big Data Now, O'Reilly, September 2011, ISBN: 978-1-449-31518-4.
[10] Rabl, Sadoghi, Jacobsen, Villamor, Mulero, Mankovskii - Solving Big Data Challenges for Enterprise Application Performance Management, 2012-08-27, VLDB, Vol. 5, ISSN 2150-8097
[11] Sameera Siddiqui, Deepa Gupta - Big Data Process Analytics: A Survey, International Journal of Emerging Research in Management &Technology, Vol. 3, Nr. 7, July 2014, ISSN: 2278-9359.
[12] Bernard Golden, ''McKinsey Cloud Computing Report Conclusions Don't Add Up,'' CIO.com (April 27, 2009), www.cio.com/article/490770/ McKinsey Cloud Computing Report Conclusions Don t Add Up.
[13] V. Diaconita - Processing unstructured documents and social media using Big Data techniques, Economic Research-Ekonomska Istraživanja, Vol. 28 (1), pp. 981-993, Routledge Publisher, 2015, ISSN: 1331-677X (Print) 1848-9664 (Online).
[14] A.R. Bologa, R. Bologa, A. Florea - Big Data and Specific Analysis Methods for Insurance Fraud Detection, Database Systems Journal, Vol. 4 (4), pp. 30-39, 2013, ISSN 2069 – 3230.
[15] O. Stanescu, D. Bolborici, S. Oprea - Modeling of wind power plants generators in transient stability analysis, Journal of Sustainable Energy, Vol. 3, Issue 4, 2012, ISSN 2067-5534.

Adela BÂRA is a Professor at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. She has graduated the Faculty of Economic Cybernetics in 2002, holds a PhD diploma in Economics from 2007. She is the author of 9 books in the domain of economic informatics, over 50 published scientific papers and articles (among which over 30 articles are indexed in international databases, ISI proceedings, SCOPUS and 15 of them are ISI indexed). She participated as team member in 3 research projects and has gained as project manager two research contracts, financed from national research programs. She is a member of INFOREC professional association. From May 2009, she is the director of the Oracle Excellence Centre in the university, responsible for the implementation of the Oracle Academy Initiative program. Domains of competence: Database systems, Data

warehouses, OLAP and Business Intelligence, Executive Information Systems, Decision Support Systems, Data Mining, Big Data.

Iuliana BOTHA is a Lecturer at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. She has graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2006, the Databases for Business Support master program organized by the Academy of Economic Studies of Bucharest in 2008 and she holds a PhD diploma in Economic Informatics from 2012. She is author/co-author of 7 books, 17 published articles (4 articles ISI indexed and the other included in international databases) and 25 scientific papers published in conferences proceedings. She participated as team member in 6 research projects (among which one international research program). From 2007 she is the scientific secretary of the master program *Databases for Business Support* and she is also a member of INFOREC professional association. Her scientific fields of interest include: Database Systems, Data Warehouses, Business Intelligence, Design of Economic Information Systems.

Anda BELCIU has graduated the Faculty of Economic Cybernetics, Statistics and Informatics of the Bucharest University of Economic Studies, in 2008. She has a PhD in Economic Informatics and since October 2012 she is a Lecturer. She teaches Database, Database Management Systems and Software Packages seminars and courses at the Economic Cybernetics, Statistics and Informatics Faculty. Her scientific fields of interest and expertise include database systems, e-business, e-learning, spatial databases.

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

Electricity tariff systems for informatics system design regarding consumption optimization in smart grids

SimonaVasilica OPREA
The Bucharest University of Economic Studies
simona.oprea@csie.ase.ro

High volume of data is gathered via sensors and recorded by smart meters. These data are processed at the electricity consumer and grid operators' side by big data analytics. Electricity consumption optimization offers multiple advantages for both consumers and grid operators. At the electricity customer level, by optimizing electricity consumption savings are significant, but the main benefits will come from indirect aspects such as avoiding onerous grid investments, higher volume of renewable energy sources' integration, less polluted environment etc. In order to optimize electricity consumption, advanced tariff systems are essential due to the financial incentive they provide for electricity consumers' behaviour change. In this paper several advanced tariff systems are described in details. These systems are applied in England, Spain, Italy, France, Norway and Germany. These systems are compared from characteristics, advantages/disadvantages point of view. Then, different tariff systems applied in Romania are presented. Romanian tariff systems have been designed for various electricity consumers' types. Different tariff systems applied by grid operators or electricity supplierswill be includedin the database model that is part of an informatics system for electricity consumption optimization.
Keywords: *Time of use tariff, critical peak pricing tariff, real time pricing tariff systems, sensors, smart-metering*

1 Introduction

Apart from previous periods before electricity market liberalization, tariffs' structure for end users is continually changing almost similar to mobile phones' technology.
Modifications of electricity tariff structure are consequences of competition and dynamic electricity consumers' behaviours. They will also change in the near future due to further implementation of smart metering systems.
There are different advanced tariff systems that can be applied together with smart metering systems. They will be described in details, compared and included in the database model for electricity consumption optimization.

2. Time of use tariff system

In the time of use (ToU) tariff system, the tariffs are fixed for certain time periods. Also, tariffs for working days can be different from tariffs for weekends.
This system has several advantages such as: easy implementation, high transparency, low risk, but it also has some disadvantages such as: rewarding low potential, lack of flexibility, etc. The time of use tariff system applied in Italy, electricity consumers have to reschedule electricity consumption in order to obtain savings.
In [1] the comparison between the German project „Intelliekon" with one single fix tariff and time of use tariff system that is dependent on the moment of consumption with two distinct intervals(specific for peak and for off-peak consumption) is performed. The tariff for peak consumption period is 77% bigger than the tariff for off-peak consumption period.
In Finland, grid operators are obliged by law to implement at large scale this tariff system. By it they could flatten load curve. Similar cases were noticed in England (time of use tariff system called „Economy 7"), Spain („Interval Tariff"), Italy („OfferteBiorarie"), etc.

3. Critical peak pricing tariff system

By critical peak pricing tariff system that combines tariff system that depends on the moment of electricity consumption and different tariffs levels of ToU system for certain days notified by electricity suppliers, the French consumers could reduce annual electricity expenses by 10%. This system is characterised by certain maximum critical days a year and certain minimum time for notification.

„Tempo" tariff system is a critical peak pricing tariff system that was launched in France in 1995 for residential and small business consumers with minimum capacity of 9 kW. In 2008, tariff system „Tempo" was implemented to 350000 residential consumers and over 100000 consumers that carried out small business. This tariff system is based on time of use tariff system with different levels. It defines three types of days: blue (with low tariffs), white (with average tariffs) and red (with high tariffs). The colour of the day is announced in the previous day, around 5:30 p.m. According to **Fig. 1**, between 2009 – 2010, from 1st of November and 31st of March were identified 22 red days, 43 white days identified mainly from October to May and 300 blue days (all Sundays are blue).

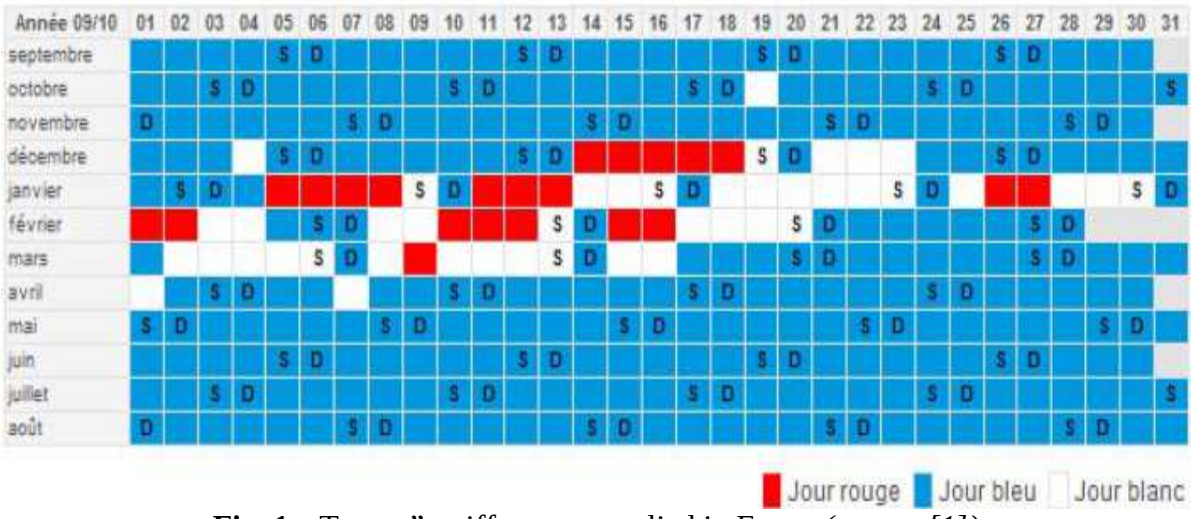


Fig. 1. „Tempo" tariff system applied in France*(source [1])*

During winter months, red and white days are numerous due to the high mix generation costs. Even five consecutive red days were recorded, that makes this tariff system more difficult to be implemented especially to those consumers with special problems.

According to **Fig. 2**, days' patternis identic with ToU tariff system, with off-peak (low tariffs) and peak (high tariffs) consumption intervals. Difference between specific tariffs for blue and red days is significant. Peak consumption tariff for red days are five times higher than peak consumption for blue days. Peak consumption tariff is five times higher than off-peak consumption tariff for red days.

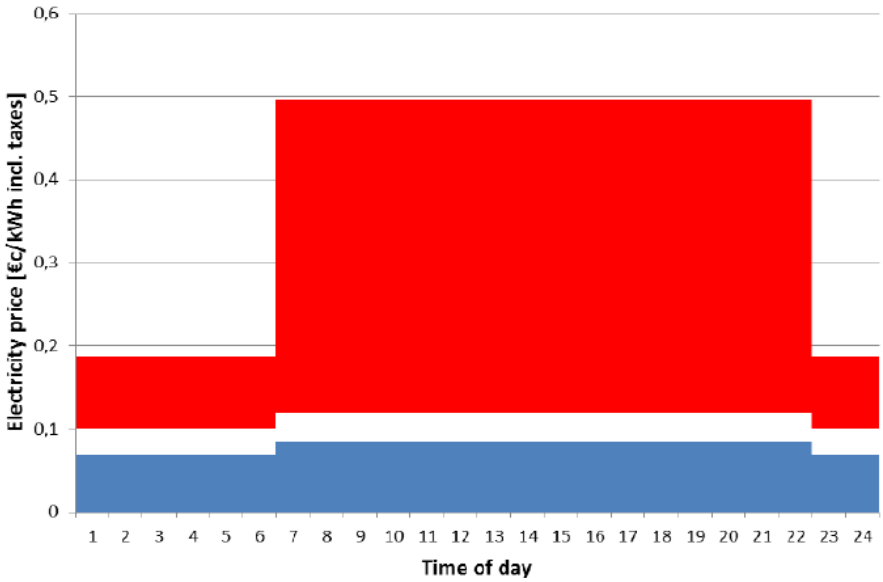


Fig. 2.„Tempo" tariff system structure applied in France*(source [1])*

The difference between peak consumption tariff and off-peak consumption tariff for white and blue days is only 20%. „Tempo" tariff system applied in France reflects energy generation costs that could vary each month, but it is more difficult to be implemented in societies with many consumers with social problems.

4. Real time tariff system

Through real time tariff system the electricity consumers know information about tariff in real time. They can respond to tariff fluctuation by raising or reducing the electricity consumption. On one hand, this tariff system allows the maximization of savings out of electricity consumption based on consumers' involvement. On the other hand, electricity consumers that are not involved in optimizing their consumption could have higher expenses.

Smart metering system allows consumers to manage in details electricity consumption. They can postpone some activities when the tariffs are higher for time intervals when tariff are low.

This tariff system is applied by some Norwegian electricity suppliers. The tariff represents the monthly average of spot price. In 2011, it varied between 3.1 €c/kWh in September and 9.3 €c/kWh in January. In 2010, the spot price was up to 57.7 €c/kWh and it recorded small variations.

Other Norwegian electricity suppliers apply tariff based on hourly average of spot price from electricity market Nordpool that could be correlated with hourly electricity consumption recorded by smart meters. According to figure 3, the price fluctuation was about ±53% compared to the average price or between 5.5 and 17.5 €c/kWh.

As shown in **Fig. 3**, fluctuations significantly depend on the season. In Norway case, the hourly fluctuations are low as a consequence of Norwegian power system particularities that are mainly based on hydro-power plants, but in other power systems, these hourly fluctuations can be high. They are intensified by the operation of power plants that are based on renewablesources.

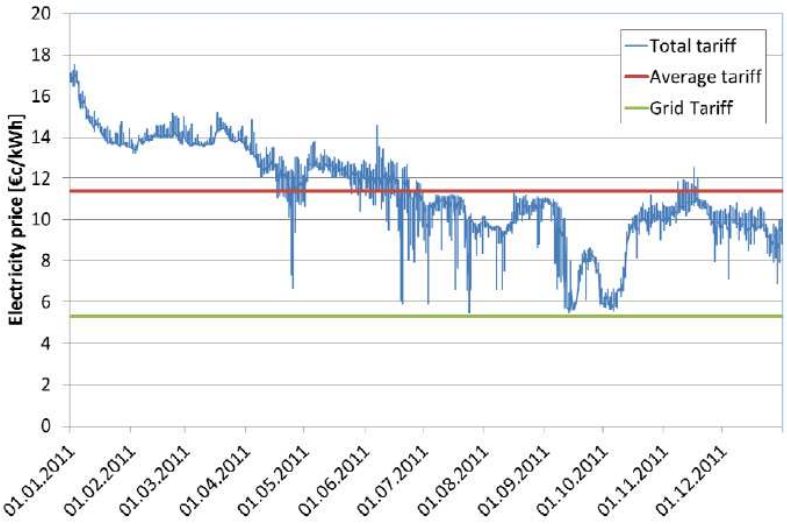


Fig. 3. Spot market price fluctuations for setting real time tariff system*(source [1])*

In Germany, some suppliers apply this tariff system. In 2010, spot market price varied between -2.945 €c/kWh and 13.179 €c/kWh. Total tariff, including grid tariff, varies between +49% and -35% compared to average of 21.1 €c/kWh.

This real time tariff system shows a series of advantages such as: increasing savings from electricity consumption, increasing the number of active electricity consumersthat are able to contribute to production-consumption balance, but also it has disadvantages, such as: increasing complexity of tariff system. In case that the settings of this system are not properly designed, price and consumption fluctuations could be uncontrollable. The real time tariff system changes the classical forecasting methods.

In **Table 1.**, I compared the three tariff systems, by shortly describing their characteristics, advantages and disadvantages.

Table 1. Comparative analysis of the advanced tariff systems

	Time of use tariff system	**Critical peak pricing tariff system**	**Real time tariff system**
Characteristics	It depends on the consumption moment; It does not need information exchange (IT&C).	It depends on the critical moments (events); It does not need information exchange (IT&C).	It depends on real time electricity request and supply; It does need information exchange (IT&C).
Advantages	Easy and simple to be implemented; Transparent, low risk.	Compared with ToU tariff system, it better reflects the market mechanisms and generation costs that depend on season.	Flexible; It involves consumers; It encourages consumption optimization; It reflects market mechanisms; Performance in dynamic control of load.

Disadvantages	It has limited potential of rewarding; Low flexibility; It involves consumers to certain extend; No performance in dynamic control of load.	Limited performance in dynamic control of load; It is difficult to estimate the critical moments; Medium complexity; It is difficult to be implemented to consumers with social problems.	Complex; Low transparency; If it is not well designed it can lead to unbalances; High volume of data.

It can be concluded that all described tariff systems are suitable for smart metering systems. It should be compatible with the most of the electricity consumers to whom it will be applied.

Taking into account that smart metering systems integration is an on-going developing process, it is reasonable to appear new improved tariff systems that will be more and more adaptable to electricity consumers' behaviour.

5. Tariff system included into the database model

In the previous paragraphs, different advanced tariff systems have been described. They are mature tariff systems, being already applied in some European countries. In this paper we design a database model that includes the tariff system. It is part of consumption records table. This table is linked to contract table that is related to electricity suppliers and consumers (figure 4).

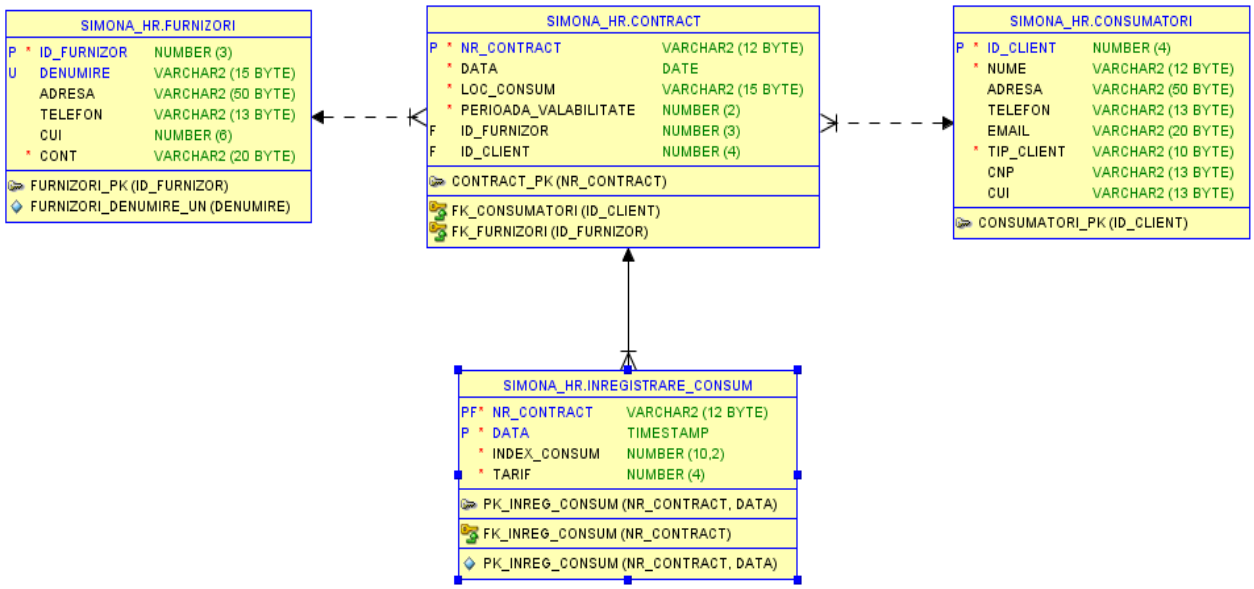


Fig.4. Database model including tariff system

For the electricity consumers the following data are stored: identification number, name, address, phone, email, type, personal numeric code, identification unique code in case of legal entities. One electricity consumers may have one or two contracts;also one electricity supplier may conclude one or more contracts. Each contract is defined by the following data: contract number, data, consumption place and validity period. For the electricity suppliers the following data are stored: id,

name, address, phone, identification unique code and bank account. For each contract, one or many consumption records can be stored. The consumption record is defined by: contract number, data (this combination will form the primary key), consumption index and tariff system, since at different contracts can be applied different tariff systems. They are included in the database model that is part of informatics system for electricity consumption optimization [3], [4].

6. Electricity tariff system applied in Romania

In Romania, electricity market has two components: regulated market and competitive market. On the regulated market, the tariff system for electricity consumers is diverse and has several components.

Social tariff (known as CS tariff) is a fix tariff, available only for electricity consumers with net monthly revenue for family member less or equal to economy minimum wage. It is recommended for monthly consumption less than 110 kWh; under this limit, it is the most economic tariff. As for the invoice calculation, the total consumption monthly is divided into three parts: the first part is defined up to 2 kWh per day (for instance 60 kWh / 30 days), that is invoiced at smallest tariff, the second part of monthly consumption is from 2 to 3 kWh per day (between 60 and 90 kWh / 30 days) that is invoiced with a medium tariff and the third part of consumption that is above 3 kWh per day (over 90 kWh / 30 days) that is invoiced at a very high tariff, design to discourage the wrong choice of tariff. If the consumer choose social tariff, but its electricity consumption exceed certain limit, his invoice could be very expensive.

Monomial tariff without reservation (known as CD tariff) is a single component tariff that includes only the price for electricity. It is suitable for monthly consumption of maximum 43 kWh. Taking into account that this tariff has no reservation and only electricity consumption is paid, when there is no consumption, the consumer has nothing to pay, but it has the disadvantage that in case of higher consumption, expenses become higher and higher compared to other tariffs. It is recommended for small monthly consumption, up to the suggested limit, for cases when social tariff couldn't be applied. It is also recommended for non-permanently inhabited houses, in case supply interruption is undesirable or for holiday houses with monthly consumption less than 43 kWh.

Monomial tariff with reservation (known as CR tariff) is the most common option for electricity consumers because it allows any daily variations of consumption and there is no limitation of monthly consumption. It is non-restrictive tariff, without risks due to inconsistency with imposed conditions of other tariff systems types.

It has two components of tariff: reservation tariff that is applied for every invoicing day and electricity consumption tariff that is applied for each kWh. Reservation component represents grid operator's maintenance expenses. It is necessary to maintain the grid in good state and permanently available to consumers.

Monomial tariff including the consumption (known as CI tariff) is recommended for daily average consumption bigger than 1 kWh. It is a tariff with subscription that includes reservation costs and daily consumption of 1 kWh.

It has two components: subscription tariff that is applied per invoice regardless the consumption and electricity tariff applied for each kWh; both components of tariff have different values depending on voltage level. Since the subscription has the same value for 1 or less than 1 kWh/day, this tariff is disadvantageous when the consumption is not performed; it will not rollover the next period.

Monomial tariff with differential reservation for two pricing zones (known as CR2 tariff) has two components:

reservation tariff that applies for each invoicing day, regardless the consumption volume and electricity tariff that applies based on the two pricing zones.
The two pricing zones are: "the day zone" that refers to Monday to Friday from 7 a.m. to 10 p.m. and "the night zone" that refers to Monday to Friday from 10 p.m. to 7 a.m. and weekend hours from Friday 10 p.m. until Monday 7 a.m.
Electricity tariff is lower at night and it is higher during the day, this difference was created in order to diminish the electricity expenses by increasing the consumption at night and weekend. This tariff can be applied only whether smart meters are implemented; regular meters are not able to keep track of different pricing zones.
Monomial tariffwith differential reservation for three pricing zones (known as CR3 tariff) is the most complex tariff for residential consumption. It requires good management of daily load curve from consumers' side.
The tariff has two components: reservation tariff that applies for each invoicing day, regardless the consumption volume and electricity tariff that applies based on voltage level, the three pricing zones and seasons. The seasons are: summer season, from 1st of April to 30th of September and winter season from 1st of October to 31st of March.
Electricity tariff has the smallest value during night hours, average value during "normal" consumption hours and highest value during peak hours.
Monomial tariff (known as CTP) has three levels of power (written in the contract): up to 3 kW, between 3 and 6 kW, over 6 kW. Based on the three levels of power, the different three tariffs are applied.It is recommended for medium-size consumers that well know each equipment from the installed power and operation simultaneity control point of view in order to determine the right absorbed power [4].
On non-regulated or competitive market, electricity consumers can negotiate with suppliers the tariff. The negotiated tariff can be:

- Variable in case of exceeding certain referential values, agreed between parties, for instance, the level of average tariff on regulated market. In these cases negotiated tariff is less than the average tariff on regulated market;
- Fix for certain time period. Whether the tariffs increase over the time, the benefit is on consumer side and whether the tariffs decrease, the benefit is on supplier side. In these cases, the supplier keeps the tariff constant;
- Otherwise agreed by parties [5].

7. Conclusions

The advanced tariff systems are essential for electricity consumption optimization due to financial incentives that could transform into savings due to consumers' behaviour changing. In this paper, advanced tariff systems (time of use, critical peak pricing, real time tariff systems) have been described.
They are applied in England, Spain, Italy, France, Norway and Germany. These tariff systems have been compared taking into account their characteristics, advantages and disadvantages. The tariff system is part of the database model that has been design by the author.
Then different tariff systems applied in Romania have been described. They have been designed for different electricity consumers types.

Acknowledgment

This paper presents results of the research project: Intelligent system for predicting, analyzing and monitoring performance indicators and business processes in the field of renewable energies (SIPAMER), research project, PNII – Collaborative Projects, PCCA 2013, code 0996, no. 49/2014 funded by NASR.

References

[1] European Parliament, Directorate-General for internal policies, Policy department, Economic and scientific policy, *Effect of smart metering on electricity prices– Briefing note*, 2012 http://www.europarl.europa.eu/document/activities/cont/201202/20120223ATT39186/20120223ATT39186EN.pdf

[2] ILungu, A Velicanu, A Bâra, I Botha, A M Mocanu, A Tudor – Spatial Databases for Wind Parks, Economic Computation and Economic Cybernetics Studies and Research Journal, ISSN: 0424-267X, nr.2/2012, pp.5-23, http://www.ecocyb.ase.ro/22012/Lungu%20Ion%20_DA_.pdf

[3] AFlorea, A Andreescu, V Diaconita, AUta, Approaches regarding business logic modeling in service oriented architectureRevistaInformaticaEconomica, 2011

[4] ElectricaMuntenia Nord, *Tariffs description*, 2015 http://www.electricafmn.ro/persoane-fizice/tarife/descriere-tarife/

[5] ANRE, *Eligible consumers' guide for choosing the supplier*

Simona-Vasilica OPREA (b. July 14, 1978) is an Assistant at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Bucharest University of Economic Studies of Bucharest. She has graduated the Polytechnic University in 2001, holds a Master Diploma in Infrastructure Management Program, Yokohama National University, Japan in 2007 and a PhD diploma from 2009. She is the author of over 25 articles. Domains of competence: wind power plant operation, investment opportunity analysis, studies of prognosis, etc.

Data integration approaches using ETL

Alexandra Maria Ioana FLOREA, Vlad DIACONITA, Ramona BOLOGA
Bucharest University of Economic Studies
alexandra.florea@ie.ase.ro, vlad.diaconita@ie.ase.ro, ramona.bologa@ie.ase.ro

Traditional data warehouses and ETL tools have been slowly pushed to expand their limits as big data has become a more and more prominent actor on the analytics stage. This paper analyzes and compares the features of Pentaho Data Integration and Oracle Data Integrator, two of the main data integration platforms. Such tools are relevant in the context of the evolution of the analytics field, which is expanding from classical business intelligence activities to the use and analysis of big data.
Keywords: *business intelligence, data warehouse, big data, ETL*

1 Introduction

In [1], published in 1958, H.P Luhn tries to define the characteristics of a Business Intelligence System, showing its flexibility in *identifying known information, in finding who needs to know it and in disseminating it efficiently either in abstract form or as a complete document*. A more modern definition of BIS is given by Howard Dresner: *concepts and methods to improve business decision-making by using fact-based support systems*.

The term business data warehouse was first used in [2], describing an information retrieval and reporting service that runs against a repository of all required business information.

The term Big Data appeared later in connection with volumes of data that are difficult to store, process and analyze using traditional database technologies [3].

There are different opinions regarding the place and purpose of Data Warehouses (DW) and Big Data within an enterprise, the similarities, and differences between those two technologies.

In [4] the authors argue that Big Data provides cheaper solutions to store raw data which usage is not predetermined and where DW is more expensive providing solutions to store cleaned, transformed, and semantically unified data for predetermined usage. Building a DW within a firm implies a certain level of business integration.

In [5], Bill Immon claims that there is no correlation between a big data solution and a data warehouse. He argues that a data warehouse is an architecture that assures corporate credibility and integrity and Big Data is just a technology for storing data. There are also opinions that state that Big Data is comprised of both technologies and architectures that can be used to extract value from large volumes of different types of data [6]. For example, Hive can provide data warehouse facilities over Hadoop, the best known Big Data ecosystem. It uses HiveQL, an SQL-type language, storing data in a distributed storage filesystem (usually HDFS).

2. Pentaho Data Integration

Pentaho is a powerful Business Intelligence open source suite that offers many features, including reporting, OLAP pivot tables and dash-boarding [7]. The Pentaho engines were developed as community projects and later integrated into the main product. Data integration can be seen as the process that combines data from a variety of sources in order to provide a coherent view. Pentaho Data Integration (PDI) draws it roots from the business intelligence tool Kettle (KDE Extraction, Transportation, Transformation and Loading Environment) originally developed for the KDE desktop environment that is mostly found in Linux distributions.

As discussed in [8], ETL is the backbone of the DW architecture, so its performance and quality are relevant for the accuracy, operability, and usability of data warehouses. The data transforming activities can be run in the target database managing system, and the process is sometimes called ELT or in a dedicated environment outside the target (the classic ETL). Many times, the umbrella term of ETL is used no matter where the transformation process takes part. PDI it's not only an ETL tool, but it can also be used in many scenarios such as loading data warehouses or data marts, integrating data in order to have a unified view, migrating data, data cleansing.

Spoon is the graphical tool that can be used to design and test every PDI process in the form of transformations and jobs. These are handled by different parts of the meta-data driven PDI engine. The jobs and transformations can be saved as XML files, in database repositories or in the PDI repository. They are created graphically, using drag and drop, but some steps can be further refined using scripting. Transformations work with streams of data, transforming the rows according to the declared steps. Jobs contain a sequence of transformations and other auxiliary tasks.

In figure 1 we built a mapper and a reducer transformation. In this case, the mapper works on a corpus of text files which it splits into rows, putting a 1 for every word it finds. The output consists of pairs of keys and values in the form of words and counts. In this case, the reducer transformation aggregates the input using the sum function.

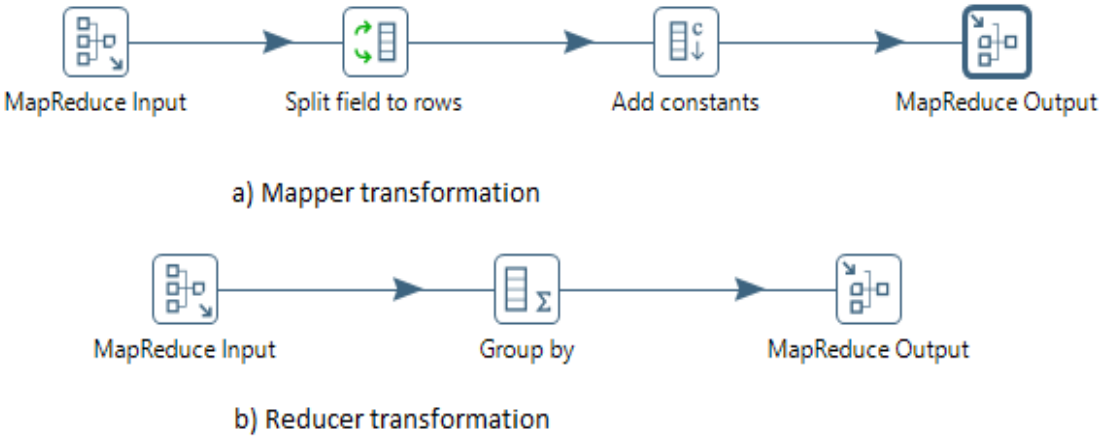


Fig. 1. Mapper and reducer transformations in PDI

In figure 2, we built a job process that connect to a Hadoop cluster, copies a file into HDFS and then runs a MapReduce process. The MapReduce process uses the two prior built transformations, the key value pairs constructed from the input files by the mapper transformation are after a combination process sent to the reducer processes that sums the appearances of every found word.

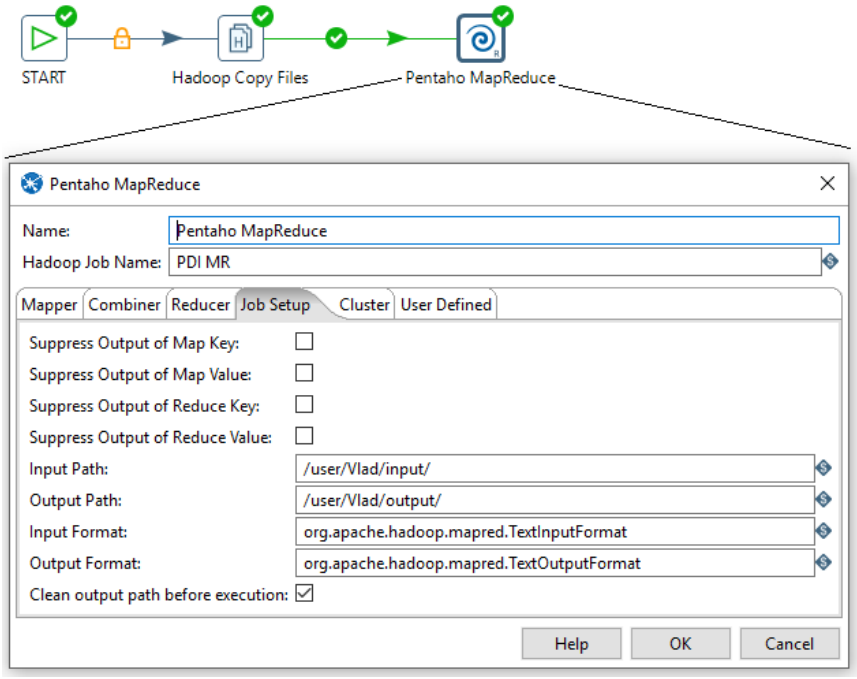


Fig. 2. A PDI job

When the job is executed, Pentaho sends the MapReduce job to the cluster. Its progress can be checked using the cluster's URL of the *Application Tracker* as shown in figure 3 and the result will be written in the declared Output Path in the cluster's HDFS.

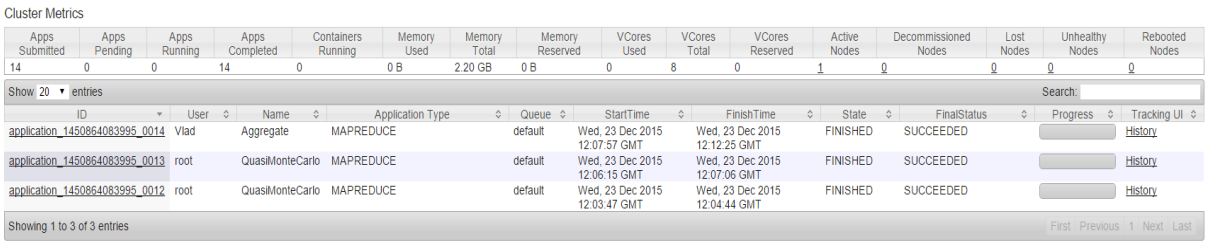


Figure 3. Hadoop Application Tracker showing the MapReduce jobs

We can also define jobs and transformations that use a MapReduce approach to run an SQL phrase in order to populate a fact table in Hive, but in our test environment that proved a very slow task. As shown in figure 4, it took 1h42 minutes to move 319 rows.

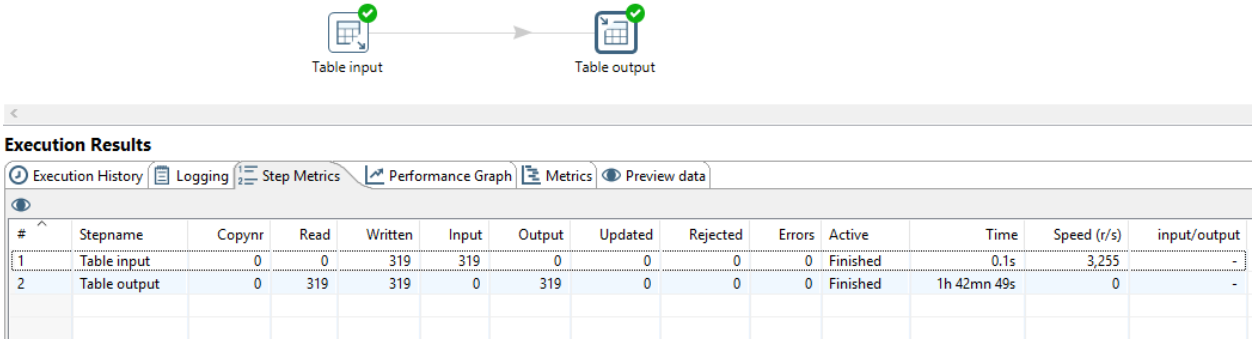


Fig. 4. Loading data from Oracle to Apache Hive using PDI

3. Oracle Data Integrator

As marketed by Oracle, ODI is a highly used, comprehensive platform that covers all of the data integration requirements, starting from high-volume, high-performance batch loads, to event-driven, trickle-feed integration processes, to SOA-enabled data services [15].

It is developed on an unusual architecture that follows the E-LT (Extract-Load-Transform) process instead of the traditional ETL one (Extract-Transform-Load). As a result there is no need for an ETL server situated between the data sources and the target server, the platform using instead the capabilities of the users' RDBMS engine. The traditional ETL process involves downtime of the data warehouse while the loads are performed, but when loading real-time data, there cannot be any downtime of the system and such new solutions must be found. [9]. ODI can perform complex transformations on the source side as well as on the target side, and a large part of these transformations occur in batch mode when there are no end-user queries to be processed by the server.

The data sources for an ODI integration can be extremely complex, varying from different types of data tables to XML files. XML files are regularly used to transmit data from different production units such as power plants [10], [11]. EXtended Markup Language or XML is the basis of all elements which represent the foundation of Web services technology. Considering platform independence, XML is the engine that enables data transfer via the Internet, also constituting the foundation of Web services.[12]

There are five main components of the ODI platform namely the Repository, ODI Studio, the agent, the console and Oracle Enterprise Manager.

The Repository is the central element of ODI's architecture and represents the main storage location where all the information that ODI handles is kept, namely, connectivity details, metadata, transformation rules and scenarios, generated code, execution logs, and statistics. It consists of two types of repositories, one Master repository which hosts sensitive data such as security and topology information as well as versioned and archived objects and several Work repositories which host project-related data.

ODI Studio represents the GUI of the platform (Graphical User Interface), and it provides access to the repositories to those who use it, regardless if they are administrators, developers or operators. It can be used to administer the infrastructure, reverse-engineer the metadata, develop projects, scheduling, operating and monitoring executions.[13]

The studio is organized in four different Navigators which usually are used by various users according to their roles and security profiles.

The Security navigator is used by system admins and DBAs to manage the roles and privileges of the regular users.

The Topology navigator is also usually utilized by the same type of users as above, in order to define the connections and credentials required for ODI to connect to the source and target systems.

Although developers will use this information most of the time, they do not have the privilege to modify it.

The design navigator is the central part of the developer's world, being used to define the needed transformations in objects called Interfaces.

Operator Navigator is the fourth component of the Studio which ensures the management and monitoring of the activities. Developers use it to check how the code has executed and to debug if necessary.

As presented in [14] the Operator Navigator has the following accordions: session List which displays all sessions organized per date, physical agent, status, keywords, and so forth; hierarchical sessions which display the execution sessions arranged in a hierarchy with their child sessions; load plan - shows the load plan runs of the load plan

instances; the scheduling accordion shows the list of physical agents and schedules; the scenarios accordion displays the list of scenarios available and the solutions accordion contains the solutions that have been created when working with version management.

In order to demonstrate the capabilities of the operator navigator, we have developed an interface based on the following scenario. We'll be moving data from a Customer System database into a data mart. The source table is called CLIENTI_MASTER (fig. 5) and the target table we will use is the CLIENTI (fig. 6) table located in the DATAMART schema.

COLUMN_NAME	DATA_TYPE	NULLABLE
CLIENTID	NUMBER	No
PREFIX	VARCHAR2(10 BYTE)	Yes
NUME	VARCHAR2(30 BYTE)	Yes
PRENUME	VARCHAR2(30 BYTE)	Yes
ADRESA	VARCHAR2(80 BYTE)	Yes
ORAS	VARCHAR2(20 BYTE)	Yes
ID_TARA	NUMBER	Yes
TELEFON	VARCHAR2(20 BYTE)	Yes
DATA_NASTERE	DATE	Yes
ID_PARTENER	NUMBER	Yes
ID_VANZATOR	NUMBER	Yes

Fig. 5. The CLIENTI_MASTER table

COLUMN_NAME	DATA_TYPE	NULLABLE
ID_CLIENT	NUMBER	No
PREFIX	NUMBER	Yes
NUME	VARCHAR2(30 BYTE)	Yes
PRENUME	VARCHAR2(30 BYTE)	Yes
ADRESA	VARCHAR2(80 BYTE)	Yes
ORAS	VARCHAR2(30 BYTE)	Yes
ID_TARA	NUMBER	Yes
TELEFON	VARCHAR2(30 BYTE)	Yes
VARSTA	NUMBER	Yes
ID_PARTENER	NUMBER	Yes
ID_VANZATOR	NUMBER	Yes
DATA_CREARE	DATE	No
ULTIMA_ACTUALIZARE	DATE	No

Fig. 6. The CLIENTI table

We can easily notice that although the two tables are quite similar there are a couple of differences that need to be addressed through mappings. Most of these will be automated but a couple must be done manually:

- The PREFIX column in our source is a character string, but in the target, it's a number. The Customer System has all the titles normalized to US prefixes, but we may wish to process the data later to have national-specific titles based on country or residence, so our data mart has all prefixes translated into numeric codes.
- Our target stores an age in the 'Varsta' field, but our source has the birth date (data_nastere).

For this, we have built an interface that deals with all the necessary mappings, as shown in figure 7.

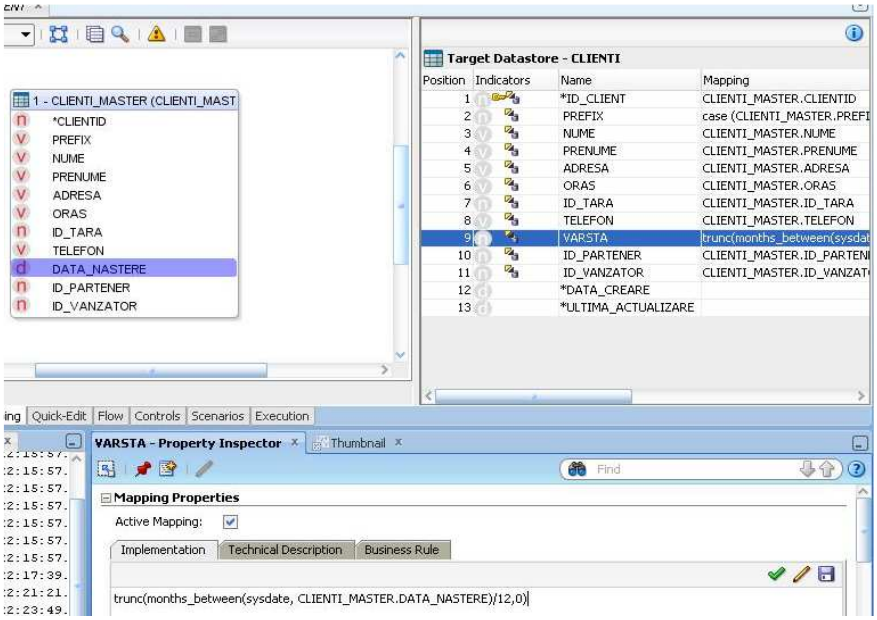


Fig. 7. ODI Interface

In order to verify the results of the interface execution, we user the operator Navigator presented previously.
We can notice that we can visualize the work sessions ordered after several criteria such as execution agent, session name, session status, the user who executed or we can see all sessions executed, ordered by their number (fig. 8).

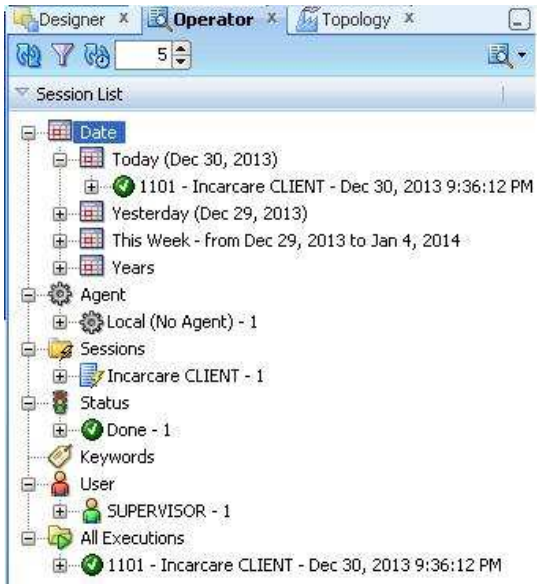


Fig. 8. ODI integration sessions

If we select the 'all executions' node and choose one of the previous sections we will be able to have some general information about that particular execution including duration, the number of inserts, updates and deletes, and the execution status of the session.
In this particular view, we can see that there have been 3 rows inserted with no updates or deletes (fig. 9).

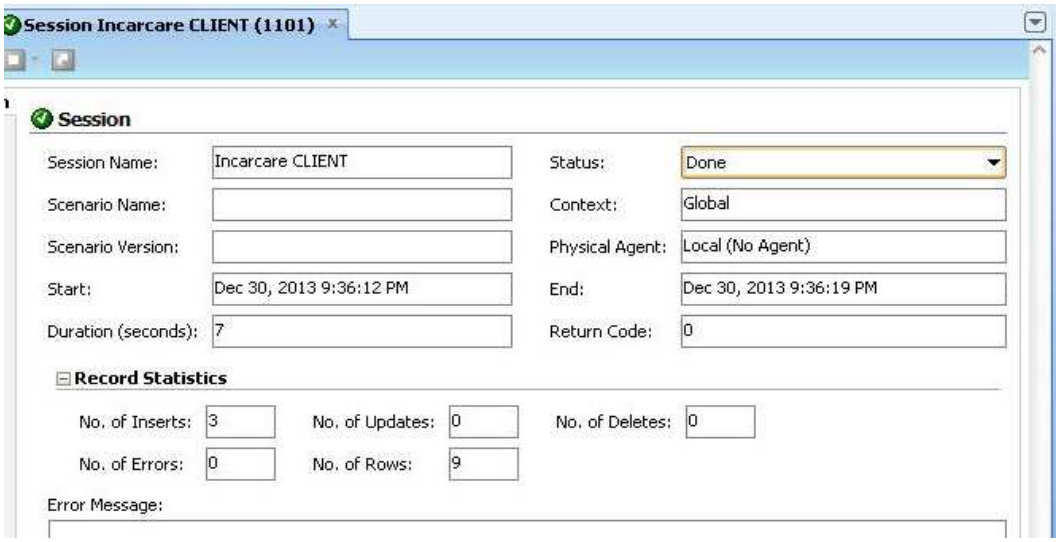


Fig. 9. Overview of an ODI integration session

Back to the list sessions, by expanding the 'Incarcare CLIENT' node, note that the session has an important step, which can then be further expanded to all the steps ODI created and executed to run the integration interface as seen in figure 10. The steps named 'Loading' have been generated by the loading knowledge

modules (LKM) while the one named 'integration' have been generated by the integration knowledge modules (IKM).

Where there is a yellow triangle, it means that a Warning had been issued, in our case when we tried deleting temporary tables that did not exist at runtime.

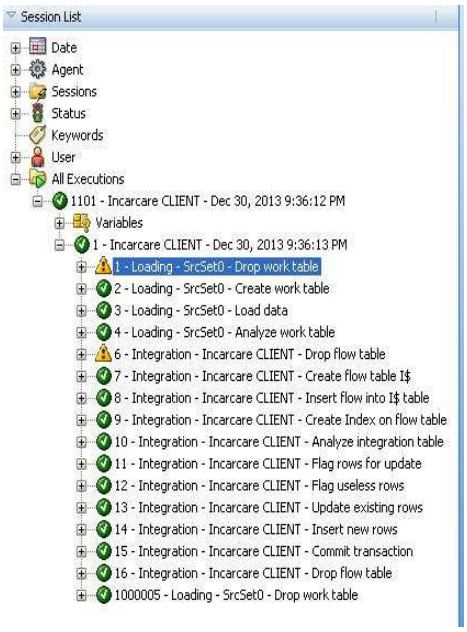


Fig. 10. Sessions' steps

Amongst the 16 steps of the execution, step 3 '*Loading – SrSet0 – Load* data' time is when the LKM module loads data into a temporary load table in the staging area. We double-click on this step to see information about execution such as before but in this case, the values in the fields for the number of inserts, updates, and deletes refer to the temporary loading work table (C$) here.

We select the Code tab and view the generated SQL source code for this step and the code correspondent to the destination for this stage (loading the working table in the staging area). We walk through the code and notice the mappings that we've previously specified for the 'age' and 'code' fields are included (fig. 11).

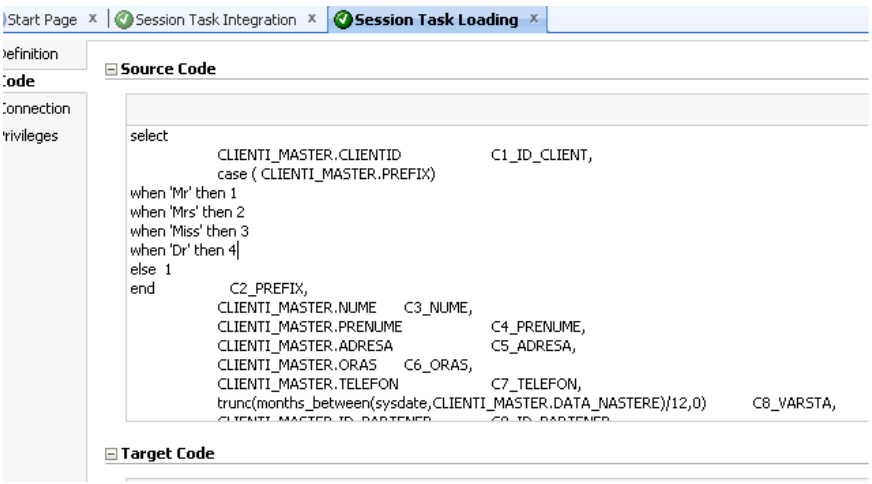


Fig. 11. Mappings information

When we return to viewing the session we choose the session step '*Integration – Incarcare PRODUS – Insert flow into I$ table*'. The window that opens shows the activity that took place on the temporary integration table. This is the step where the mappings that have been configured to take place in the staging area are executed. In this step runs mappings that have been configured to take place in the staging area.

5 Conclusions

Even though PDI was the first to handle modern data structures and offered connectors to Hadoop clusters, ODI is catching up quickly and is currently offering powerful connectors and Big Data transformations. Compared to ODI, PDI doesn't necessary need a repository, its jobs and transformations can be saved as XML files. If required, both products can use different database engines (Oracle, Microsoft SQL Server, Sybase, DB2, etc.) to install their repositories. The community, open source version of PDI can be freely downloaded. It's the code can be found on http://sourceforge.net/. ODI can be freely downloaded but only for evaluation purposes; the source code cannot be downloaded. We can conclude that both PDI and ODI are powerful, highly customizable integration tools that receive constant upgrades from its developers, so the choice for a particular integration project is more a matter of taste and budget.

References

[1] H. P. Luhn, "A Business Intelligence System," IBM Journal, 1958. [Online]. Available: http://altaplana.com/ibmrd0204H.pdf. [Accessed: 21-Dec-2015].

[2] B. A. Devlin and P. T. Murphy, "An architecture for a business and information system," IBM Syst. J., vol. 27, no. 1, pp. 60–80, 1988.

[3] I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of 'Big Data' on cloud computing: Review and open research issues," Inf. Syst., vol. 47, pp. 98–115, 2014.

[4] B. Boulekrouche, N. Jabeur, and Z. Alimazighi, "An intelligent ETL grid-based solution to enable spatial data warehouse deployment in cyber physical system context," 1st YAWL Symp. 2013, vol. 56, no. MobiSPC, pp. 111–118, 2015.

[5] B. Immon, "Big Data Implementation vs. Data Warehousing," 2013. [Online]. Available: http://www.b-eye-network.com/view/17017. [Accessed: 19-Dec-2015].

[6] R. L. Villars and L. Borovick, "Big Data and the Network," IDC White Paper, 2011. [Online]. Available: https://www.brocade.com/content/dam/common/documents/content-types/whitepaper/idc-big-data-network.pdf. [Accessed: 21-Dec-2015].

[7] R. Bouman and J. Van Dongen, Pentaho® Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL®. 2009.

[8] A. Karagiannis, P. Vassiliadis, and A. Simitsis, "Scheduling strategies for efficient ETL execution," Inf. Syst., vol. 38, no. 6, pp. 927–945, Sep. 2013

[9] J.Popeangă, I. Lungu, "Real-Time Business Intelligence for the Utilities Industry", Database Systems Journal vol. III, no. 4/2012, [Online]. Available: http://www.dbjournal.ro/archive/10/10_2.pdf

[10] I Lungu, A Velicanu, A Bâra, I Botha, A M Mocanu, A Tudor – Spatial Databases for Wind Parks, Economic Computation and Economic Cybernetics Studies and Research Journal, ISSN: 0424-267X, nr.2/2012, pp.5-23 [online]. Available http://www.ecocyb.ase.ro/22012/Lungu%20Ion%20_DA_.pdf

[11] A Bara, I. Lungu, S.V. Oprea, I. Botha, A. Chinie, "Model assumptions for efficiency of wind power

plants'operation", Economic Computation and Economic Cybernetics Studies and Research Journal, ISSN: 0424-267X, nr.2, pp.5-23 [online]. Available http://www.ecocyb.ase.ro/eng/Articles_4-2014/07%20-%20Ion%20Lungu,%20Adela%20Bara%20(T).pdf

[12] A. Florea, " Business Process Management Solutions Performance Tuning and Configuration", Database Systems Journal, Vol. II, No. 3/2011, ISSN: 2069 – 3230. Available: http://dbjournal.ro/archive/5/3_Florea.pdf

[13] Oracle, "Oracle® Fusion Middleware Getting Started with Oracle Data Integrator 12c", 20015, Available: http://www.oracle.com/technetwork/middleware/data-integrator/overview/odi-12c-getting-started-guide-2032250.pdf

[14] P.C. Boyd-Bowman, C. Dupupet, D. Gray, D. Hecksel, J. Testut, B Wheeler, " Getting Started with Oracle Data Integrator 11g: A Hands-On Tutorial", Packt publishing, Mumbai, India, 2012

[15] Oracle Data Integrator, http://www.oracle.com/technetwork/middleware/data-integrator/overview/index-0883 (accessed January 06, 2016).

Alexandra Maria Ioana FLOREA obtained her Ph.D. in the field of economic informatics in 2012 and at present, she is a lecturer at the Academy of Economic Science from Bucharest, the Economic Informatics Department. Her fields of interest include integrated information systems, information system analysis and design methodologies and database management systems. She published more than 40 papers in peer-reviewed journals and conference proceedings, many indexed by ISI or SCOPUS and is the co-author of 3 books.

Vlad DIACONITA has a Ph.D. in Statistics and Economic Cybernetics and is a member of the IEEE and INFOREC organizations and member of the technical team of the Database Systems Journal. As part of the research team, he has worked in 4 UEFISCDI funded grants. He has received an award and a scholarship as part of the EU funded Excelis project. He published more than 30 papers in peer-reviewed journals and conference proceedings, many indexed by ISI or SCOPUS. He is the co-author of four books.

Ana-Ramona BOLOGA is an associate professor at the Academy of Economic Studies from Bucharest, Economic Informatics Department. Her Ph.D. paper was entitled “Software Agents Technology in Business Environment”. Her fields of interest are: integrated information systems, information system analysis and design methodologies, and software agents.

Big Data, the perfect instrument to study today's consumer behavior

Cristina STOICESCU
University of Economic Studies, Bucharest, Romania
kris_stoicesku@yahoo.com

Consumer behavior study is a new, interdisciplinary and emerging science, developed in the 1960s. Its main sources of information come from economics, psychology, sociology, anthropology and artificial intelligence. If a century ago, most people were living in small towns, with limited possibilities to leave their community, and few ways to satisfy their needs, now, due to the accelerated evolution of technology and the radical change of life style, consumers begin to have increasingly diverse needs. At the same time the instruments used to study their behavior have evolved, and today databases are included in consumer behavior research. Throughout time many models were developed, first in order to analyze, and later in order to predict the consumer behavior. As a result, the concept of Big Data developed, and by applying it now, companies are trying to understand and predict the behavior of their consumers.

Keywords: *Big Data, consumer behavior, consumer experience, machine learning*

1 Introduction

During the last century there was an unprecedented change in the digital world that caused big shifts also in the economic and financial field. With a continuously growing middle class, more and more people actively using social networks, and new generations who grew up with technology around them, the personal and business environment has changed drastically.

According to Internet Live Stats which collects data from different international sources, around 40% of the world population has an internet connection today, while 20 years ago it was less than 1%. The number of internet users has increased tenfold from 1999 to 2013, the third billion being reached in 2014. Of it, today only one billion people are using social networks, but by 2020 this number will double, according to analysts.[16]

All these changes are transposed in day to day behavior. People want to do more activities at the same time, to improve their quality of life daily, to have a successful career and a balanced family life. In other words, people want to increase the productivity of their work by making less effort.

The consumer has changed in the same way. Decades ago, if somebody wanted to buy a book, whereof one likely learned about from his friends or relatives, first he had to go into more bookstores to see if that book exists and after that to make some price comparisons in order to decide from where to buy it from. These activities were time and money consuming.

The situation has changed radically. Now the person can learn about launching a book easily from social networks, and by simply accessing an online store such as Amazon.com, they can purchase the book by pressing a button, finally saving time and energy. So, the process of buying a product simplified in terms of time and money spent, but became more difficult in terms of decision making which has become more complex. The main reason is that people have today too many options to choose from in terms of product or service, price, quality and time. If in 20th century there were few providers for a product (most of the time it was only one provider, competition being inexistent), today competition between providers has increased significantly, each of them having more similar products to satisfy the same need for the consumer, each product having different price and different characteristics (different quality).

In conclusion, our daily life has changed. Today we are using the internet in every activity, from the most simple (like buying a book) to the most complex: when we check the weather on our phone, when we buy on-line tickets for a concert, when we check our personal or business e-mails, when we compare the offers for a holiday trip or when we organize an event. By using this services we leave behind traces of information that can be smart used by companies in taking important decisions. Today, this is the main way Big Data is created (besides data collected by enterprises that use ERP systems or paper surveys), because it is fast, real-time and easy to collect. Because today's market is mainly driven by consumers, companies all over the world understood the importance of studying the behavior of their clients in order to meet their needs increasingly more specific. They also understood that Big Data is the perfect tool, in order to achieve accurate results and increased profit. [21]

The term "Big Data" describes the accumulation and analysis of vast amounts of information. But Big Data is much more than a big amount of data. It is also the ability to extract meaning: to sort through big volumes of numbers and find the hidden patterns, unexpected correlations, and surprising connections that can be used in different industries like medical field, security and protection field or marketing field. In marketing field, companies that adopt "data-driven decision making" enjoy significantly greater productivity than those that do not. So, by using Big Data rapid deployment solutions they can lower the complexity of implementation projects, and hence project risks, while accelerating time to value.[23]

As a result, Big Data is an extraordinary knowledge revolution that is sweeping almost invisible through business, academia, government, healthcare, and everyday life. It already ensures safety and independence for older people, enables us to provide a healthier life for our children, conserves precious resources like water and energy, and peers into our own individual genetic makeup. Big Data is the perfect instrument to study today's consumer behavior. [26]

2 Evolution of consumer behavior. How do people take purchase decisions today?

The purchasing decision process began to be studied about 300 years ago by Nicholas Bernoulli (in 1783 he introduced the terms of expected utility and marginal utility in the economic theory), followed by John von Neumann and Oskar Morgenstern (they introduced the terms of risk and uncertainty, and in 1944 they published a fundamental article for microeconomics "Theory of Games and Economic Behavior"). They created a mathematical model in order to determine the utility gained after a consumer activity, people being considered pure rational beings (consumers tried only to satisfy self-interest).

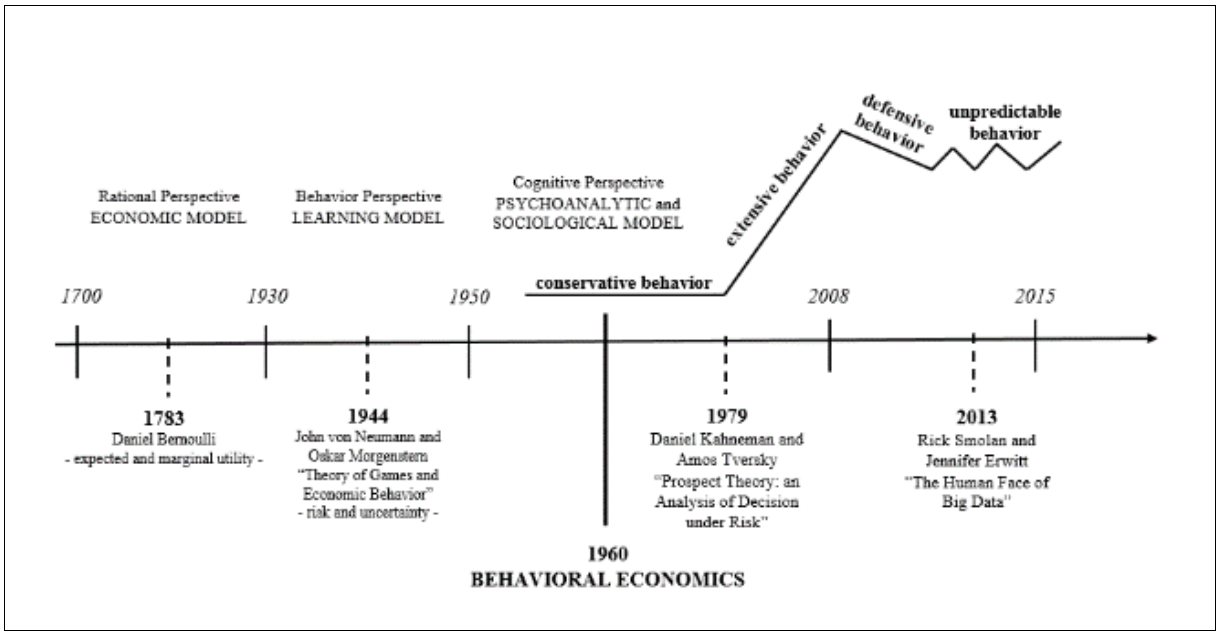


Fig. 1. Evolution of consumer behavior

Recent research shows that there are numerous factors that influence the purchase decision, besides the rational ones, like social, cognitive and emotional factors. By taking these factors into consideration when modelling the purchasing decision process, a new, interdisciplinary and emerging science appeared in 1960: the study of consumer behavior. It is a complex science that includes information from economics, psychology, sociology, anthropology and artificial intelligence.[6]

Until 1960, the economic perspective of consumer behavior and the models that described it relied on the assumption that all consumers are always rational in their purchases, so they will always buy the product that will bring the higher satisfaction. In this regard, three types of models were developed. Between 1700 and 1930, the Economic Model was used to describe consumer behavior which involved the rational perspective based on Neoclassical Economic Theory. In the next 20 years, the behavior perspective began to be applied which was based on the Learning Model, and after that the cognitive perspective which was based on the Psychoanalytic and Sociological Model.[18]

During this time, people had a conservative behavior because they were buying the same products, consumer behavior being an emergent phenomenon that has evolved along with human development. In prehistoric times this behavior was shown in a very limited way, people being organized in small family groups and having a single concern: survival. Much later, the social skills began to develop that finally led to the emergence of money, social status, wealth and ultimately shaped the consumer behavior.[17]

The main cause that determined the researchers to study the consumer behavior is the diversification of needs. In the same time, looking back a century ago, a strong connection can be observed between the moment when the population began recording a strong upward trend and the science of studying consumers behavior appeared (1960, correlation between figures 1 and 2).

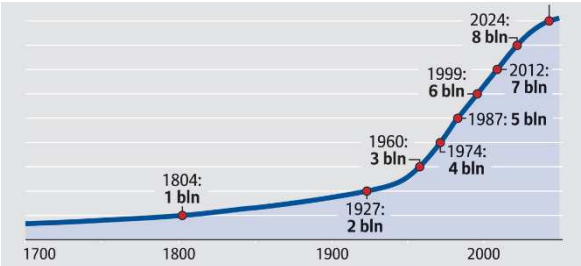


Fig. 2. Population growth on Earth

This correlation can be explained by the fact that a growing population means more needs, more products and more suppliers. Also, the life expectancy has doubled in the last century (at the beginning of 20th century the life expectancy was about 30-40 years, while in 2008 it changed to 70 years), with the same result in the change of consumer behavior: more needs to be satisfied and a consumer behavior more complex.

Also, the middle of 20th century is the moment when travel started to become accessible to all, due to the large-scale production of machinery and commercial aircraft. By travel, people had the opportunity to discover other cultures and habits and as a result their needs started to diversify. Whereas in the past most people lived in small towns, with limited possibilities to leave their community and few variations in needs, now, due to technical improvements, consumers began to have increasingly more diverse needs.

For half a century, people developed an extensive behavior, buying increasingly more products and increasingly more diverse. One of the most important paper written during this period is "Prospect Theory: an analysis of decision under risk", written by Daniel Kahneman and Amos Tversky, which proposes a new model for studying consumer behavior. In this paper, the decision making is viewed as a choice between prospects or gambles. The authors developed the new theory from the assumption that the Theory of Expected Utility (which was not challenged for over 250 years), had some flaws regarding the moment when the choice is made by the consumer. They thought that the utility is not only dependent on the actual value of a persons wealth, but also on the evolution of his or her wealth.[19]

The Prospect Theory is the most important model used at the end of the 20th century, but there were also other models created in that period of time: Nicosia model (1966), Engel, Blackwell and Miniard model (1968), Howard Sheth model (1969), Webster and Wind model (1972), Hobbes model (1984) and Veblen model (1994). [17]

The year 2008 represents another important moment in world's history that influenced the consumer behavior. The economic and financial crisis that spread all over the world led consumers to think twice before buying a product. Because consumers were buying fewer products, their behavior began to be a defensive one.

People began to use the internet on a larger scale in order to search products and to compare their price and characteristics. Online marketing began to have a decisive role in the buying process, so new techniques were developed in order to predict the consumer behavior, one of them being Big Data.[27]

Today consumers face an offer too diverse, being assaulted by marketing messages. Because of that, the opportunity cost for a product has significantly increased, making the decision process more and more complicated. According to studies, consumers may ignore the opportunity cost when they don't have to choose from more than 8 products. When the number of choices increases, the consumers become indecisive, and sometimes even give up to the buying process.

The changes in consumer behavior have had strong influences on all enterprises throughout time, a decisive moment being the mid-1970s when a significant macroeconomic change on the law of supply and demand had happened: if by that time the markets were driven by vendors, their control was taken over by buyers both in terms of influence and

bargaining power.
Companies understood that the consumer behavior is an emergent phenomenon that has evolved with human development and they became more interested in studying the behavior of their daily consumer. As a result, today's companies are empowered by the final consumer who wants instant value, mobile functionality and user-friendly services. Today, people are more informed (57% of the buying process is completed before a first interaction with sales), socially networked (53% of customers abandoned an in-store purchase due to negative online sentiment) and less loyal (59% of customers are willing to try a new brand to get better customer service). [21]
As a conclusion, the main factors that shaped consumer behavior are:

- demographic changes (the growth of population and life expectancy, had the same result in consumer behavior: more needs to be satisfied);
- evolution of technology (because people now have more ways to travel, they discovered other cultures and life styles, so their needs became more diverse);
- multiplicity (because more and more variables are integrated in every day activities – for example the movie industry has evolved from a one-dimensional to a multi-dimensional experience – also the buying act needs to become a complex experience);
- hyper efficiency (the space-time efficiency is also a daily problem, so people need faster and cheaper ways to satisfy their needs);
- risk and stress (people have too many options to choose from in order to satisfy their needs).

3 Evolution of marketing research. Big Data, the newest instrument for predicting consumer behavior

According to AMA (American Marketing Association) marketing research is "the process or set of processes that links the consumers, customers and end users to the marketer through information — information used to identify and define marketing opportunities and problems, to generate, refine, and evaluate marketing actions, to monitor marketing performance and improve understanding of marketing as a process. Marketing research specifies the information required to address these issues, designs the method for collecting information, manages and implements the data collection process, analyzes the results, and communicates the findings and their implications."
Due to the evolution of consumer behavior, the marketing research discipline has evolved in the same way, trying to adapt to the competitive economic environment. The papers regarding this discipline emphasizes its current state as being in "flux".
At the beginning of 20th century, political polling and advertising studies were the only marketing research techniques used to study the consumer behavior. Their main purpose was to find if a new solution to gain consumers was successfully or not. As people had more needs, more products were demanded, so more suppliers appeared. Due to this environmental changes, the competition became fierce, so marketing research had to adapt. As a consequence, qualitative and quantitative approaches separated into two different methods. The main qualitative methods were focus groups, in-depth discussions and observational research, while the main quantitative methods were linear models, descriptive statistics and multivariate analysis.[11]
Two of the most successfully methods of marketing research in the last years, used by international companies all over the world are: TORA (Theory of Reasoned Action) and NPS (Net Promoter Score).
The Theory of Reasoned Action was developed by Martin Fishbein and Icek Ajzen in 1975-1980, and is a model for the prediction of behavioral intention,

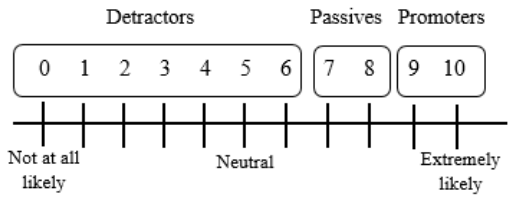
spanning predictions of attitude and predictions of behavior. In its simplest form, the TORA can be expressed with the following equation:

$$BI = (AB)W_1 + (SN)W_2$$ *Source: Hale, 2002*

where:

BI = behavioral intension
(AB) = one's attitude toward performing the behavior
W = empirically derived weights
SN = one's subjective norm related to performing the behavior

Net Promoter Score is a customer loyalty metric developed by Fred Reichheld, Bain & Company and Satmetrix, and introduced by Reichheld in 2003 in a Harvard Business Review article, "One Number you need to grow". It serves as an alternative to traditional customer satisfaction research and claims to be correlated with revenue growth. It can be calculated using the answer to a single question, using a 0-10 scale: How likely is it that you would recommend [brand X] to a friend or colleague? and the respondents are grouped using the following formula [4]:



NPS = % of Promoters - % of Detractors
Source: Bergevin , 2010

As higher this indicator, as satisfied are the consumers of a company.

In conclusion, if a century ago the change of market research techniques was linear and primarily associated with methods and data collection techniques, the current patterns indicate a shift base on a qualitative evolution. Although, the classic measurement systems were analyzing behavioral intention, not actual behavior, mainly because it has been the easiest information to collect. Until now, collecting data on actual consumer behavior has been impractical, due to the emergence of the Internet, social media and e-commerce, which have radically altered the landscape of consumer behavior data. Point-of-Sale (POS) and cash registers systems are being replaced by e-commerce sites that record every move consumers make. Casual telephone conversations with friends about recent purchases are being replaced by tweets that can be analyzed by anyone who follows those Twitter feeds.

In fact, everything that is build these days (phones, computers, cars, refrigerators) are producing terabytes and petabytes of data. Information is being extracted from everywhere, out of parking spaces, out of toll booths, out of Internet searches, out of Facebook, out of our phones, so every action that people make these days leaves a digital trace that can be recorded, stored, and after that analyzed.[23]

So while customers can tell what they think, data scientists can tell what those customers actually do, because data on actual consumer behavior and experiences is now available to be measured and analyzed. In order to do that, a new tool was developed: Big Data.

Big Data is commonly defined as the combination of volume (a large quantity of data), variety (multiple types of data) and velocity (the speed at which data is created). With traditional techniques, users can be provided with volume and variety of data, but is difficult to include velocity. Even by regularly feeding new data, this are static data and not coherent with the decisions that must be made.[21]

Table 1. The Advantages and Disadvantages of using Big Data

ADVANTAGES	DISADVANTAGES
Volume: we are recording a huge amount of data	
Big Data improves the quality of life and the customer experience by giving extra senses (today every medical aspect of a human being can be captured: the metrics can be captured by sensors, the anatomy can be captured by imaging, while the biology can be captured by using the sequence of DNA, and by having a complete view be can improve our health; by recording all consumers activities a complete view about them can be created, and in this way the buying experience can be improved)	The more data are registered, the larger the problems will be that analysts need to solve (can we find in the huge amount of data the information that we need and can it influence positive our life and/or the consumer experience)
Variety: we are recording data from different sources	
Big Data can identify hidden pattern and unexpected correlations to propose new and innovative solutions	People don't have a personal life anymore; their lives have become more transparent
Velocity: we are recording real time data	
By doing this real-time actions can be made that can solve real-time problems	At this moment it is not known who owns the data and how are they used
Veracity: we are recording inaccurate data	
Once the "dirty data" is removed, the useful and accurate data can be use to extract new information	This data can lead very easy to an avalanche of errors and incorrect results, affecting the whole business

Some papers include a fourth dimension for Big Data: veracity. Veracity is the hardest thing to achieve with big data, because due to the volume of information and the variety of its type is hard to identify the useful and accurate data from the "dirty data". The biggest problem is that the "dirty data" can lead very easily to an avalanche of errors, incorrect results and can affect the velocity dimension of Big Data. The main purpose of the Big Data can be corrupted and all the information can lead to a useless and very expensive Big Data environment, if there is not a good cleaning team.[27]

Like all technologies, using Big Data in ERP systems has advantages and disadvantages at the same time that are displayed in Table 1.

Today, there are a lot of industries that use Big Data: healthcare (treatments are becoming personalized and patient centric and predictive analysis are used to prevent diseases; for example Angelina Jolie underwent a preventive double mastectomy after learning she had 87% risk to developing breast cancer), sports (by using sensors data are collected from players during a game in order to improve their playing schemes), weather (more than 60 years of global weather analysis are used to predict the risk of future extreme events), logistics (smart trucks and smart spaces; an example is the High Bay deposit of Coca-Cola from Ploiesti, Romania), agriculture (monitoring weather and soil conditions for optimum point of harvesting), manufacturing (industry 4.0) and energy and telecommunication (smart grid and virtual plants).[23]

Hamburg Port is the largest port in Germany and the second largest in Europe. The current turnover is about 9 million containers/year but by 2025, an increase to 25 million containers is expected by using Big Data. The goal is to utilize the current infrastructure and increase container turnover by optimizing container traffic while reducing idle time for carriers.[22]

Due to the evolving consumer demands and the ever-growing digitization, the world is digitally transforming which

means that new technologies are used to driven significant business improvements. Big Data is one of the 4 channels through digital transformation is made, together with cloud, mobile and networks. The challenges for digital transforming, and therefore using Big Data as a main technology, are: digital proficiency, legacy systems, security and jobs becoming obsolete.

According to Trifu and Ivan, Big Data is a unique concept that integrates all kinds of data, not just some basic ones like in normal data warehouse. So, Big Data uses data from text to pictures, sounds, movies, music, satellite coordinates, or any other type of input or output data that came from different types of sensors. According to Sven Denecken, Global Vice President for Cloud Solutions, Big Data will stand for predictive insight driven by business strategy, new product strategies, and new consumer relationships. Using the right data in the right context will mean smarter decisions, new opportunities, and ultimately a big competitive advantage.

Using an ERP system a dynamic Big Data environment can be created, using real time data and including all four dimensions of Big Data. One of the tools that can overtake this new created environment is SAP HANA, an in-memory database platform.

SAP HANA was released in the early 2010 to allow for real-time analytics of both structured and unstructured data. In memory analytics refers to the new way computers are managing data and applications by keeping data in their main memory instead of regularly having to access the hard drive to retrieve the data. By using this new technology, users can access quasi real-time data analytics that provide meaningful information. Also, the visualization tools, such as SAP Lumira, allow this exploration and understanding of the data, and ultimately supports the decision-making process.[23]

4 The correlation between the evolution of consumer behavior and marketing research. Big data, the perfect instrument to study today's consumer behavior

In recent history, two of the most important moments that changed the world were the two World Wars. Before them, people were trying to find a balance in their life, consumer behavior being constant. The market research tools were also pretty limited (political pooling and advertising studies), their main goal being to identify if a certain marketing strategy may have an effect: IS it happening? Marketing strategy, does it have any effect on consumer behavior?

After them, consumers started to have an open-minded behavior, trying new products and buying more. Marketers adapted implementing more models in order to study this extensive behavior: surveys, focus groups, interview, trying to create a whole image of their consumers behavior: WHAT is happening? Are the consumers going to buy a certain product?

Today customers want instant value, mobile functionality and user-friendly services, so their behavior has changed. Because they are more informed (57% of the buying process is now completed before a first interaction with sales), socially networked (53% of customers are now abandoning an in-store purchase due to negative online sentiment) and less loyal (59% of customers are willing to try a new brand to get better customer service) they started to act differently. Market research is trying to adapt to this changes, implementing new tools (observation, online surveys, Big Data, neuromarketing) in order to identify WHY is it happening? (figure 3)

Big Data is not only an analysis of a lot of data. It is a complex process that can extract new information in order to understand the background of the industry in which a company operates, to assess the specific factors that apply to a company and to familiarize a company with a great

number of analytical instruments. Big Data is used to present the new information extracted from data, so that managers can understand and use it in business decisions.

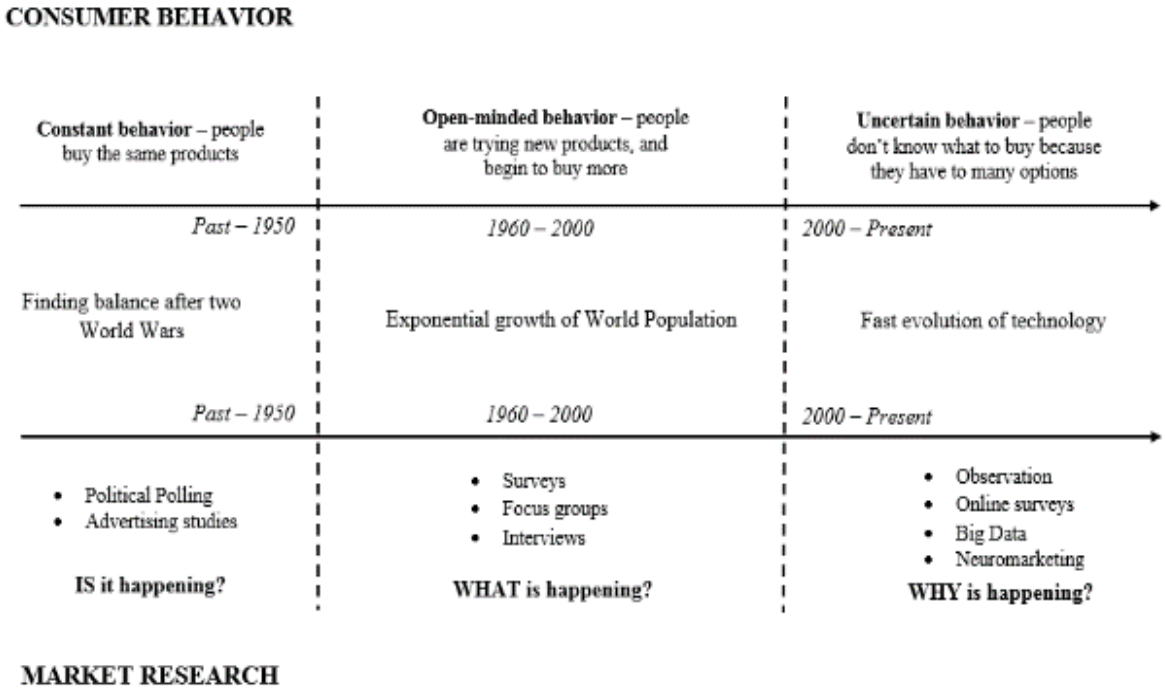


Fig. 3. The correlation between the evolution of consumer behavior and market research

Using Big Data a lot of information about consumer behavior that can improve their buying experience can be extracted, like: customer migration, customer approach, promotion analysis, acquisition analysis, priority analysis, sales according season, customer loyalty, cross sales, customer segmentation, channel of communication and media analysis, channel of distribution analysis, basket analysis, customer gain and loss analysis (churn).[11]

For example, by making a seasonality analysis companies can discover how the buying profile changes during the year and therefor create seasonal patterns in products or services that will shape the business strategy. A priority analysis will determine if there is a particular order in which customers prefer to buy products, while target marketing and niche market determination will determine if there are segments that have specific buying patterns. For example, Tesco made a study using different strategies in tandem and discovered that customers who start buying Pampers, also start buying more beers. The explanation for this behavior, which seems strange at first site, is that fathers of toddlers do not have time anymore to go to the pub, so instead they drink beer at home.[2]

But the most popular analyzes made through Big Data are basket analysis and cross-selling analysis. They determine the associations between products within the shopping basket of a consumer, increasing both the quantity of products from the same category and from complementary or even totally different categories.

In conclusion, Big Data is the perfect instrument to study today's consumer behavior, a strong bond being created between them. The succession of Big Data analytics and business decisions is an infinite single loop: users are analyzing current data and making business decisions that will generate other data, which represents the feedback of their decisions.

If the new data corresponds to the objectives of the company then the users have received a positive feedback, else they have received a negative one. In both cases, they need to analyze the new data in order to adapt again their business strategy. In both cases, using information generated by Big Data will adjust the business strategy of a company, helping it to survive in the unpredictable economic environment.

5 Using Big Data and Machine Learning to enrich consumer experience

Big Data and Machine Learning are both subfields of computer science that evolved from the study of pattern recognition. Big Data refers mainly to the large data sets, while machine learning involves the study and construction of algorithms. The list of machine learning algorithms includes the following: decision tree learning, association rule learning, inductive logic programming, support vector machines, clustering, bayesian networks, reinforcement learning, representation learning, similarity and metric learning, sparse dictionary learning, genetic algorithms and artificial neural networks.

Starting from the artificial neural networks, new research concerning consumer behavior led to a new science called neuromarketing. This science studies the reaction of consumers to different stimuli, using neuroimaging techniques such as magnetic resonance imaging, electroencephalography or magnetoencephalography. Basically, this science uses neural techniques in order to understand consumer behavior regarding brands and marketing.

Because neuromarketing brings innovation and added value, this technique has already been used by big companies like PepsiCo, Google, Coca-Cola, Disney and P&G. The complexity of this technology is given by the difficulty to transpose the results in decisions and actions without the help of specialists in the field. However, recent studies have managed to express the reactions of the human brain in only three key indicators: attention (the subject is captured or bored), emotional activity (picture conveys a positive emotion or a negative one) and capture memory (subject is able to easily memorize the received images).

The main problem of Big Data is that it can't predict with full accuracy the consumer behavior, mainly because the behavior is an emergent phenomenon of human brain. However, the prediction can be improved if neural computation is used. The goal is to identify the reactions that take place in brain and map them into mathematical equations.

Although a mapping of biological neurons has been tried, the link between artificial and biologic neural networks is only on algorithmic level. So, although researchers have tried to make a copy of the human brain, eventually creating artificial intelligence, complexity and its details have finally made this action to fail. However, artificial neural networks still use a fundamental principle, which is machine learning, also used by biological neural networks ("The Organization of Behavior," published in 1949 by Donald Hebb shows that a neural connection gets stronger as it is used, using the concept of machine learning)

Big Data and Machine Learning are both used to enrich the consumer experience. According to studies, the markets have shifted from features to experience, consumer user experience becoming the new standard. The process of buying a product or a service is no longer seen as a simple action, but as a complex experience that can determine the consumer to return to the company or to never come back.

The value of user experience can be measured through 3 variables: people, business and technology. All these variables must be taken into consideration when a product is chosen to be sold, or else 3 common mistakes can appear. First one is over engineering: the focus of the company is on the business value and

needs and the technology. In this case, the consumer desire and need is not taken into consideration, so the product is technically feasible and has a business behind it, but is too complex for the consumer (it can be handled only by professionals).

The next mistake is wishful thinking and it commonly appears when only the business and design people are selecting a product that is not technically feasible. The last mistake is vogue and it appears when design and technology people are working without taken into consideration the business part.[22]

Whatever the mistake, the consumer experience is negatively affected.

If no mistake is made, and the user experience is improved, than the company will gain both non-monetary and monetary benefits. The non-monetary benefits are increase user satisfaction, increase customer loyalty, increase solution adoption and strengthen relationship between IT and business, while the monetary benefits are productivity gain, training costs saving, users errors decrease, change requests decrease.

Today, a company that does not fulfill the standards of its industry or does not uses a certified system, may encounter a lack of integrated data. This is not the case of ERP systems which have single point access to all data. Because of this the concept of Big Data can be easily applied and new information can be extracted, that will bring benefits to the business management.

Big Data is used in a lot of industries, one of them being the Marketing industry. One of the tools that uses Big Data in order to study the consumer behavior is SAP hybris Marketing, the marketing solution of SAP (Systems, Applications & Products in Data Processing) which is integrated into the other SAP solutions like ERP (Enterprise Resource Planning) or CRM (Customer Relationship Management). The main goal of this tool is to identify what the consumer is looking for by combining the information about what the consumer is doing now and the information regarding what the consumer has done. There are a lot of similar tools that do not integrate historical data, this way excluding the most important part of consumer behavior.

SAP hybris Marketing can collect data about what the consumer have purchased in-store or on the Web shop, what he has looked at on the dot.com pages, what he has shared on social media, and even about inquiries or complains he has sent to the company. All collected information fits together into single profiles of costumer.

By creating a complete image of the customer behavior, real time individualization is possible, which can be explained by the next example. At the time that a consumer leaves a Web shop, he gets the information that his abandoned shopping cart is still available to him for further purchase. The company also does not lose this information, but instead uses it in order to send an offer to the consumer for a price reduction for the same products that were in the abandoned shopping cart. This is right-time context information delivered to the consumer.

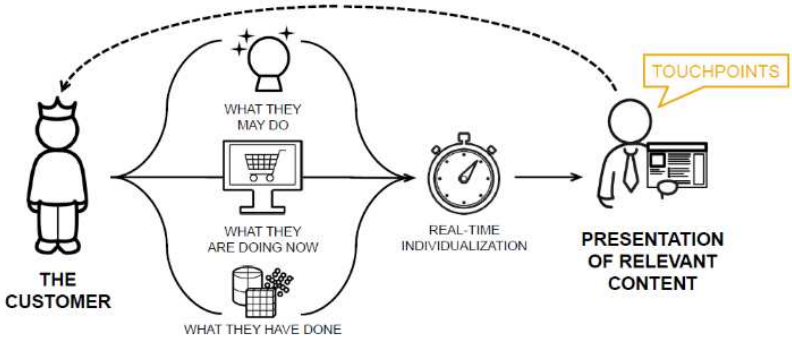


Fig. 4. Improving consumer experience (real time individualization using Big Data)

Also, by providing real time data, Big Data can be used to improve the response time of companies that will finally increase the customer satisfaction, gain customer loyalty and receive a higher degree of coming back.

The most important SAP hybris Marketing RDS (rapid deployment solutions) are marketing data management (SAP social contact intelligence), marketing segmentation (SAP audience discovery and targeting, dynamic target group integrated into SAP CRM, customer segmentation based on a predictive KPI), marketing acquisition (campaign management), marketing recommendation (modeling product recommendations) and marketing insight (customer value intelligence).

These RDS are preconfigured software and service packages that can run in the cloud, on premise or in a hybrid environment that can help companies to deploy quickly, predictable and affordably.

SAP RDS software applications provide a standardized approach through the use of such things as its Step-by-Step (SBS) guide, a tool which contains all the assets required during implementation, including accelerators and knowledge-transfer materials. The SBS guide is created specifically for each application, according to SAP, and is arranged in a specific order that mirrors the RDS implementation roadmap.

In conclusion, a company can improve its customer experience by empowering innovative new business models, value-added services and customer responsive products.

6 Conclusions

Consumer behavior is studied for 300 years and today it is the main focus for all companies. Along with its evolution, marketing research techniques have evolved in order to understand the customer behavior. The two concepts are described in detail in section 2 (evolution of consumer behavior) and 3 (evolution of marketing research) of this article, creating a complex image of their evolution. The novelty is represented by the strong correlation between their evolution which is graphically represented in the figures from this article.

In 2001, A. Hirschowitz stated that “no matter how sophisticated a company's ability to generate customer insight, it will deliver little value without the processes in place that exploit this understanding to build stronger customer relationships.” Today, the best processes that can create a complex and complete image of what consumers buy, and can also understand why they buy a certain product or service, is Big Data.

In an interview for KDnuggets, in January 2015, John Schitka, who works on the SAP Big Data Solution Marketing team, said: “Big Data is an opportunity to re-imagine our world, to track new signals that were once impossible, to change the way we experience our communities, our places of work and our personal lives.” So Big Data is the perfect instrument to study today’s consumer behavior.

Regarding Big Data, studies reflect that after 2017, this techniques of data analysis will be a competitive necessity, so companies need to start to adapt to the trends in order to survive in the dynamic and digitalized markets.

References

[1] Alioto, M.F., (2014). *The Evolution of Market Research*, article retrieved October 24, 2015, from https://rwconnect.esomar.org/change-in-the-marketing-research-discipline-evolution-paradigm-shift-and-the-emergence-of-human-intelligence/

[2] Anders, S.N., (2013). *Digilogue: How to win the digital minds and analogue hearts of tomorrow’s customer*, published by Wiley

[3] Barkworth, H., (2014). *Six trends that will shape consumer behavior this year*, article retrieved October 11, 2015, from http://www.forbes.com/sites/onmarketing/2014/02/04/six-trends-that-will-shape-consumer-behavior-this-year/

[4] Bergevin, R., Kinder, A., Siegel, W., Simpson, B., (2010). *Call Centers for Dummies*, published by John Wiley & Sons Canada, 345

[5] Bosomworth, D., (2015). *Mobile marketing statistics 2015*, article retrieved November 5, 2015, from http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/mobile-marketing-statistics/

[6] Brosekhan, A.A., Velayutham, M., Phil, M., (2003). *Consumer Buying Behaviour – A Literature Review*, Journal of Business and Management, 1(1), 8-16

[7] Choi, H., Varina, H., (2011). *Predicting the Present with Google Trends*, retrieved October 11, 2015, from http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf

[8] eMarketer, (2013). *Big Data helps reveal consumer behavior*, article retrieved October 10, 2015, from http://www.emarketer.com/Article/Big-Data-Helps-Reveal-Consumer-Behavior/1010357

[9] Fang, Z., Li, P., (2014). *The mechanism of "Big Data" impact on consumer behavior*, American Journal of Industrial and Business Management, 2014, 4, 45-50

[10] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., (1996). *From Data Mining to knowledge discovery in Databases*, American Association for Artificial Intelligence, 37-54

[11] Gamble, P.R., Tapp, A., Marsella, A., Stone, M., (2005). *Marketing Revolution*, published by Kogan Page Limited

[12] Glasgow, S., Zegler, J., (2014). *Mintel looks into its crystal ball and shares top 2015 consumer behavior trends*, article retrieved October 21, 2015, from http://www.bizjournals.com/chicago/news/2014/10/24/mintel-looks-into-its-crystal-ball-and-shares-top.html

[13] Halzack, S., (2015). *The new shopping behavior that is creating big challenges for the retail industry*, article retrieved November 2, 2015, from https://www.washintonpost.com/news/business/wp/2015/02/11/the-new-shopping-behavior-that-is-creating-big-challenges-for-the-retail-industry/

[14] Hirschowitz,A.,(2001).*Closing the CRM loop: The 21st century marketer's challenge: Transforming customer insight into customer value*, Journal of Targeting, Measurement and Analysis for Marketing 10, 168–178

[15] van Hove, M., (2015). *Big Data is not the same as Big Insight*, article retrieve November 7, 2015, from http://www.strategos.com/big-data-is-not-the-same-as-big-insight/

[16] Internet Live statistics, article retrieved October 15, 2015, from http://www.internetlivestats.com/internet-users/

[17] Jisana, T.K., (2014). *Consumer Behavior models: an overview*, Sai Om Journal of Commerce & Management, 1(5), 34-43

[18] Kahneman, D., Thaler, R.H., (2006). *Utility Maximization and Experienced Utility*, Journal of Economic Perspectives, 20(1), 221-234

[19] Kahneman, D., Tversky, A., (1979). *Prospect Theory: an Analysis of Decision Under Risk*, Econometrica, 47(2), 263-290

[20] Kurt, M., (2015). *Using Big Data and Machine Learning to enrich consumer behavior*, article retrieved October 29, 2015, from http://www.forbes.com/sites/kurtmarko/2015/04/08/big-data-machine-learning_customer-experience/

[21] Open Sap courses, (2015). *Digital Transformation and Its Impact*,

retrieved October 8, 2015, from https://open.sap.com/courses

[22] Open Sap courses, (2015). *Creating business value with user experience,* retrieved October 9, 2015, from https://open.sap.com/courses

[23] Open Sap courses, (2015). *Driving business results with Big Data,* retrieved October 10, 2015, from https://open.sap.com/courses

[24] Rajpurohit, A., (2015). *Interview: John Schitka, SAP on how to get started with Big Data*, article retrieved October 4, 2015, from http://www.kdnuggets.com/2015/01/interview-john-schitka-sap-big-data.html

[25] SAP, Rapid Deployment Solutions, retrieved October 14, 2015, from http://www.sap.com/solution/rapid-deployment.html

[26] Smolan, R., Erwitt, J., (2015). *The Human Face of Big Data,* article retrieved October 1, 2015, from http://thehumanfaceofbigdata.com/

[27] Trifu, M.R., Ivan, M.L., (2014). *Big Data: present and future,* Database Systems Journal, 5(1), 32-41

[28] Weathington, J., (2013). *Measure actual customer behavior using big data analytics*, article retrieved October 12, 2015, from http://www.techrepublic.com/blog/big-data-analytics/measure-actual-customer-behavior-using-big-data-analytics/

Cristina STOICESCU graduated the Faculty of Economic Cybernetics, Statistics and Informatics, with a bachelor degree in Economic Cybernetics in 2012. In 2014 she got her master degree from the same faculty of the Bucharest University of Economic Studies, specialization in Cybernetics and Quantitative Economics. Currently she is taking her PhD degree in Economic Cybernetics and Statistics at Bucharest University of Economic Studies, coordinated by Professor Crisan Albu. Her PhD thesis is "Experimental Economics in shaping consumer behavior".