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).

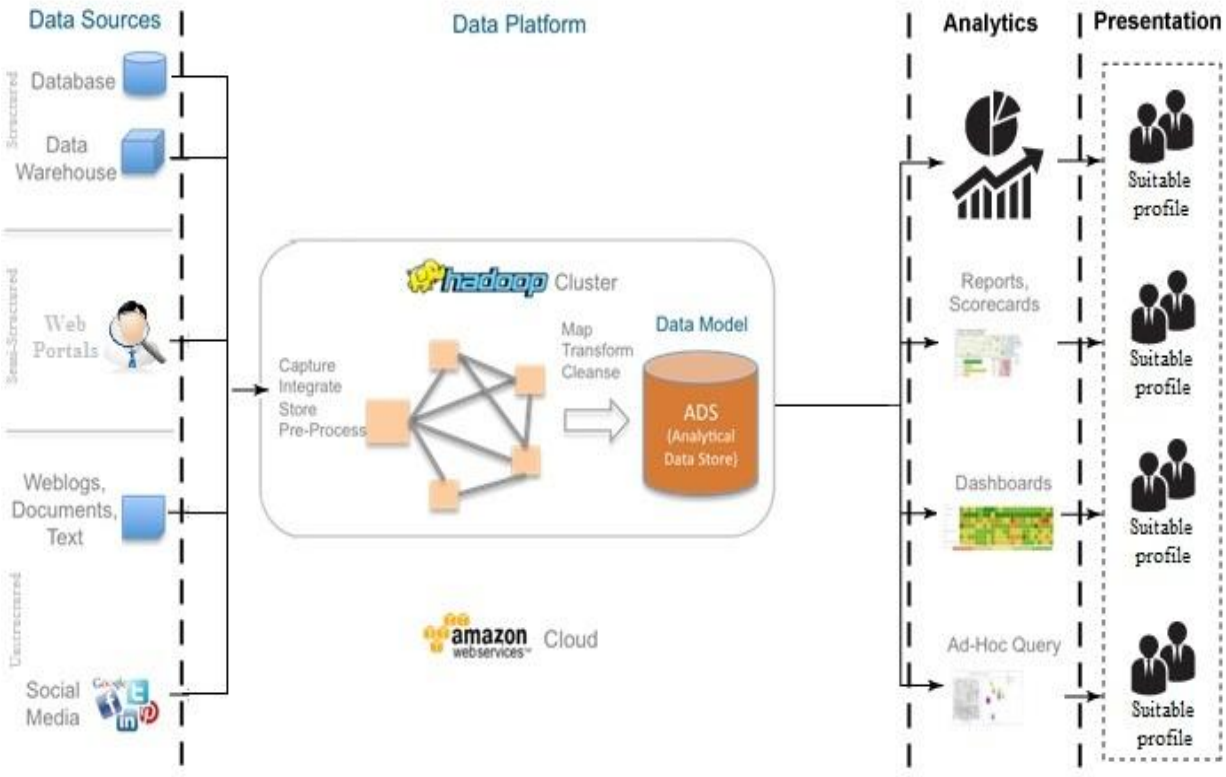

Figure 1. HR recruitment architecture

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.