THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

DATABASE SYSTEMS JOURNAL

Vol. VI, Issue 1/2015

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

Development of National Health Data Warehouse for Data Mining 3
Shahidul Islam Khan, Abu Sayed Md. Latiful Hoque

Data Mining Smart Energy Time Series 14
Janina POPEANGĂ

Big Data Analytics Platforms analyze from startups to traditional database players ... 23
Ionuţ ŢĂRANU

Using Cloud Business Intelligence in competency assessment of IT professionals 33
Elena Alexandra TOADER

Cloud Computing and its Challenges and Benefits in the Bank System 44
Bogdan NEDELCU, Madalina-Elena STEFANET, Ioan-Florentin TAMASESCU, Smaranda-Elena TINTOIU, Alin VEZEANU

In-memory databases and innovations in Business Intelligence....................................... 59
Ruxandra BĂBEANU, Marian CIOBANU

Applying BI Techniques To Improve Decision Making And Provide Knowledge Based Management 68
Alexandra Maria Ioana FLOREA

Approaches for parallel data loading and data querying 78
Vlad DIACONITA

Development of National Health Data Warehouse for Data Mining

Shahidul Islam Khan, Abu Sayed Md. Latiful Hoque

Dept. of Computer Science and Engineering (CSE),
Bangladesh University of Engineering & Technology (BUET), Dhaka, Bangladesh
nayeemkh@gmail.com , asmlatifulhoque@cse.buet.ac.bd

Health informatics is currently one of the top focuses of computer science researchers. Availability of timely and accurate data is essential for medical decision making. Health care organizations face a common problem with the large amount of data they have in numerous systems. Researchers, health care providers and patients will not be able to utilize the knowledge stored in different repositories unless amalgamate the information from disparate sources is done. This problem can be solved by Data warehousing. Data warehousing techniques share a common set of tasks, include requirements analysis, data design, architectural design, implementation and deployment. Developing health data warehouse is complex and time consuming but is also essential to deliver quality health services. This paper depicts prospects and complexities of health data warehousing and mining and illustrate a data-warehousing model suitable for integrating data from different health care sources to discover effective knowledge.

Keywords: *Data Mining, Data Warehouse, Health Informatics, Clinical Database, Data Preprocessing*

1 Introduction

Health informatics or healthcare informatics is an intersection of computer science and health care services. It deals with resources and methods needed to optimize the acquisition, storage, retrieval and use of information in medical research and applied to the areas of health care management, diagnosis, clinical care, pharmacy, nursing and public health [1, 2]. Knowledge discovery from data (KDD) is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data mining, a major part in KDD, consists of applying data analysis and learning algorithms to produce potential interesting patterns over the data [3, 4, 5].

Health data refers to any information that is contained in a patient's medical record. This information may be acquired from notes derived from a hospital admission or a doctor's visit or diagnostic report. This data comes in various forms such as text or numbers (patient identification, demographics, history, laboratory data, etc), analog or digital signals (ECG, EEG, EMG, ENG etc), images (histological, radiological, ultrasound, etc), and videos. Further complicating the storage of this data is the fact that because patient identification information cannot be publicly used. Such identifiers must be removed from other clinical parameters. Difficulty in storing this type of data is that each disease and species can only be effectively described using greatly different vocabularies and data elements. [2, 6, 7, 8].

One of the major Information Technology challenge in medical practice is how to integrate several disparate, isolated information repositories into a single logical repository to create consistent information for all users. A massive amount of health records, related documents and medical images created by clinical diagnostic equipment are generated daily. These valuable data are stored in various medical information systems such as HIS (Hospital Information System), PACS (Picture

Archiving and Communications System), RIS (Radiology Information System) in various hospitals, departments and diagnostic laboratories. Data required to make informed medical decisions are trapped within fragmented, disparate, and heterogeneous clinical and administrative systems that are not properly integrated. As a result health care suffer because medical practitioners and health care providers are unable to access this information to perform activities such as diagnostics, and treatment optimization to improve patient care [1, 6, 7].

Successful healthcare data management is an important factor in developing support systems for the clinical decision-making process. Traditional operational database system does not satisfy the requirements for critical data analysis tasks of the clinical decision-making users. It contains detailed data but do not include important historical data, and since it is highly normalized, it performs poorly for complex queries that need to join many relational tables or to aggregate large volumes of data in order to generate various clinical reports. A health data warehouse is a data store that is different from the hospital's operational databases. It can be used for the analysis of consolidated historical data [7, 8].

According to Inmon [9] A data warehouse (DW) is a subject-oriented, integrated, non-volatile, and time-variant collection of data in support of management's decisions.

Subject-oriented: as the warehouse is organized around the major subjects of the enterprise (such as customers, products, and sales).

Integrated: as DW is constructed by integrating multiple heterogeneous sources usually, such as relational databases, flat files etc.

Time-variant: as data in the warehouse is only accurate and valid at some point in time or over some time interval.

Non-volatile: as the data is not updated in real time but is refreshed from on a regular basis from different data sources.

The advantages and disadvantages of DW are given below [9, 10, 11]:

Advantages of DW:

1. Standardize data across the organization
2. Improve turnaround time for analysis and reporting
3. Easy Sharing of data
4. Remove informational processing load from operational database
5. Enhance Data Quality and Consistency
6. Provide historical intelligence and reduce cost to access historical data
7. Integrate data from multiple sources into a single repository
8. Improve data quality by providing fixing noisy data
9. Restructure the data so that it delivers excellent query performance
10. Make decision–support queries easier to write.

Disadvantage of DW:

1. Long initial development time and associated high cost
2. Data owners lose control over data, raising ownership and privacy issues

Implementing a Health DW is a complex task containing two major phases. Firstly, in the configuration phase, a conceptual view of the warehouse is specified according to user requirements (DW design). Secondly, the related data sources and the Extraction-Transform-Load (ETL) process (data acquisition) are determined. After the initial load during the operation phase, warehouse data must be regularly refreshed that is, modifications of operational data since the last DW refreshment must be propagated into the warehouse such that data stored in the data warehouse reflect the state of the underlying operational systems [5, 8, 12].

The main aim of this research is to identify the obstacles for healthcare data integration and to propose a data-warehousing model suitable for integrating fragmented data in respect to Bangladesh as well as anywhere else. The result will contribute to the advancement of knowledge in the field of medical informatics. In this paper "Health", "Clinical" "Pathological" and "Medical" these terms are used for similar meaning. The rest of this paper is organized as follows. In Section 2 we have presented selected literature reviews on DW, Health DW and KDD techniques. Section 3 describes briefly some design issues of National Health DW. In Section 4 we have shown the calculation of approximate size of our DW. Some preprocessing techniques that we have used are illustrated in Section 5. Section 6 gives readers ideas about how our DW will be used for knowledge discovery and mining. Finally Section 7 concludes the paper.

2. Literature Review

DW unifies the data scattered throughout an organization into a single centralized data structure. It is a repository of integrated information available for querying and analysis. DW may be considered a *proactive* approach to information integration, as compared to the more traditional *query driven* approaches where processing and integration starts when a query arrives [5, 6]. A health data warehouse is a repository where healthcare providers can gain access to medical data gathered in the patient care process. Extracting medical domain information to a data warehouse can facilitate efficient storage, enhances timely analysis and increases the quality of real time decision making processes. Today's healthcare organizations require not only the quality and effectiveness of their treatment, but also reduction of waste and unnecessary costs. In order to construct an operational and effective DW it is essential to combine process work, domain expertise and high quality database design [7, 8].

Electronic Health Record (EHR) describes the diseases and treatments of patients, are normally stored in hospitals or clinics, where they are created. Patients may be treated in different hospitals, clinics and, therefore, there is a need for integrating health records from different hospitals to enable any hospital to obtain a total overview of a patient's health history. Different heterogeneity problems have to be solved in order to integrate EHR systems from different hospitals and health service providers in a consistent way. The first problem is that different hospitals normally do not use a same DBMS and therefore, the traditional ACID properties of databases are missing across the different hospital locations. This may cause performance, autonomy, and consistency problems. Another heterogeneity problem is that there are several incompatible standards for EHR entries [12].

The trend of adopting data warehouses for health systems in presented in [13], where the design experience in the University of Virginia Health System is reported. Here the data warehouse is used to provide clinicians and researchers with direct, rapid access to desired patients' data. In addition they use DW also for educational and research aims, as it serves to face informatics issues−such as data capture−and to perform exploratory analyses of healthcare problems.

Medical domain has certain unique data requirements such as high volumes of unstructured data and data confidentiality. There are huge constraints and issues that limit the way the data mining is performed for medical datasets. Some of these issues are the way the data is collected; accuracy of the data, ethical, privacy and social issues that comes with patient's records [2]. Research is also done to find out impact of missing values and explore the impact of noise and how this can influence the output. Zhu et al. classified noises into class noise and attributes noise. Attribute noise include incorrect attribute values, missing or don't know attribute values and incomplete attributes or don't care values [14].

Several researches have focused on the techniques that have built in mechanism to handle noise and missing values and which are more appropriate to use for medical applications. Few techniques that have been applied and are more suited to medical data sets are studied in [15, 16]. For example decision tree, logic programs, K-nearest neighbour, and Bayesian classifiers. Lee et al recommended that Bayesian networks and decision trees are the primary techniques applied in medical information systems [17]. Obenshain claimed that that neural networks performed better then logistic regression, but the decision tree did better in identify active compounds most likely to have biological activity [18]. Wang and Wang discussed that most process models do not focus in gaining new knowledge. Medical data mining applications should follow a five stage data mining development cycle: planning tasks, developing data mining hypotheses, preparing data, selecting data mining tools, and evaluating data mining results [19].

Handling Missing Data in Pathology Databases using Multiple Imputation technique is discussed in [20]. Optimizing public health data collection for KDD using feature selection is studied in [21]. Cubillas et. al. proposed a model for improvement in appointment scheduling in health care centers [22]. Hoque et. al. discussed present structure of pathological data, requirements to formulate efficient models and the necessity to reform the present structure for predicative data mining in [23]. Kumari and Singh used Neural Network for the diagnosis of diabetes [24]. Yilmaz et. al. proposed a modified K-means Algorithm based data preparation method for diagnosis of heart and diabetes diseases [25]. Herland et. al. present recent research using Big Data tools and approaches for the analysis of Health Informatics [26].

3. Design Issues of National Health DW

The architecture of national health DW model is illustrated in Fig. 1. Health data from different govt. and private sources such as hospitals, clinics, diagnostic centers, research centers will be collected. Using ETL process data will be integrated into a temporary data repository [27].

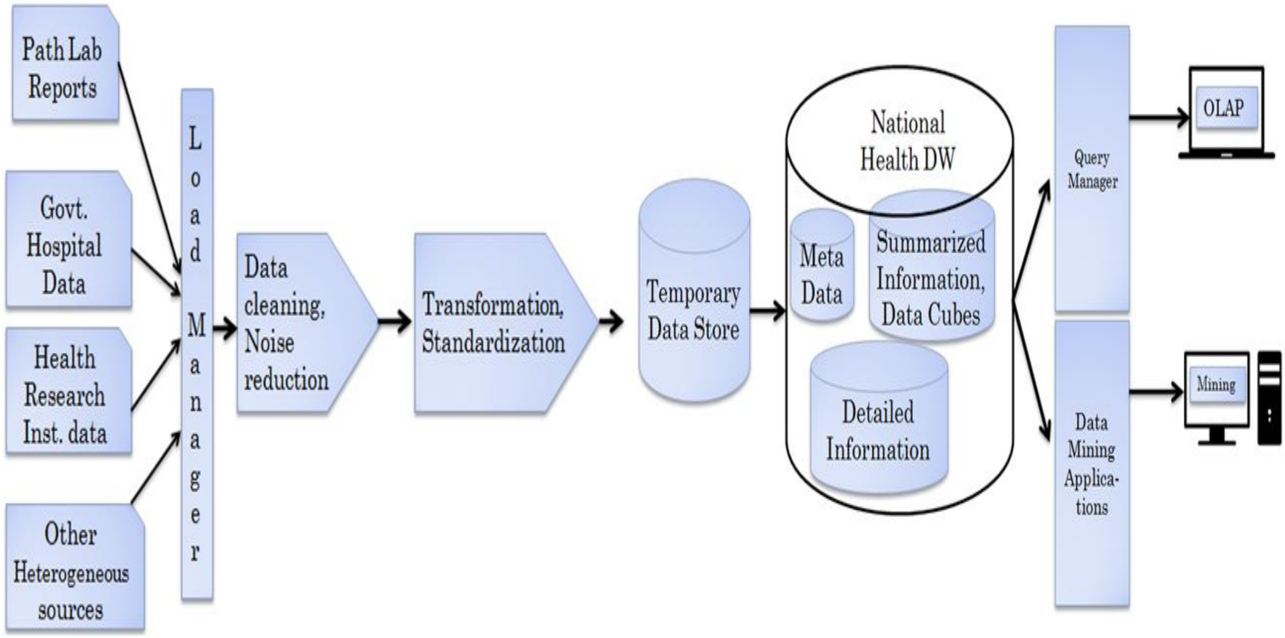

Path Lab Reports
Govt. Hospital Data
Health Research Inst. data
Other Heterogeneous sources
Load Manager
Data cleaning, Noise reduction
Transformation, Standardization
Temporary Data Store
National Health DW
Meta Data
Summarized Information, Data Cubes
Detailed Information
Query Manager
Data Mining Applica-tions
OLAP
Mining

Fig. 1 Brief Architecture of Health Data Warehouse

Cleaning, noise reduction, normalization techniques will be applied next. After that data will be loaded into DW. Online Analytical Processing (OLAP) queries and mining operations can be easily performed over the pathological DW.

4D Health data cube used for national health DW development is shown in Fig. 2. Here 0-D apex cube will provide highest level of summarization of national health data. Partial materialization is used rather than full materialization of cuboids to reduce huge space requirements [9, 10].

Logical design of DW involves the definition of structures that enable an efficient access to information. There are many logical models like Flat schema, Star schema, Star Cluster schema, Snowflake schema, Fact Constellation schema etc. Among them, star schema, snowflake schema and fact constellation schema are mostly used commercially. Efficiency is the most important factor in DW modeling because many queries access large amounts of data that may involve multiple join operations. Most suitable Logical Data Warehousing Model is the Star Schema [9, 12, 13]. We have used Star Schema in our design, illustrated in Fig. 3.

Using the building blocks of the fact table and the various dimension tables, one has thousands of ways to aggregate the data. For clinical analysis purposes, frequently needed aggregated datasets should be created in advance for the users. Having data readily and easily available is a major tenet of data warehousing. For our DW, some aggregated datasets could be:

- Patient count by Diagnosis, Gender, Age, Date
- Count of Procedures by Provider and Date
- Billing and discount information.
- Count of retesting

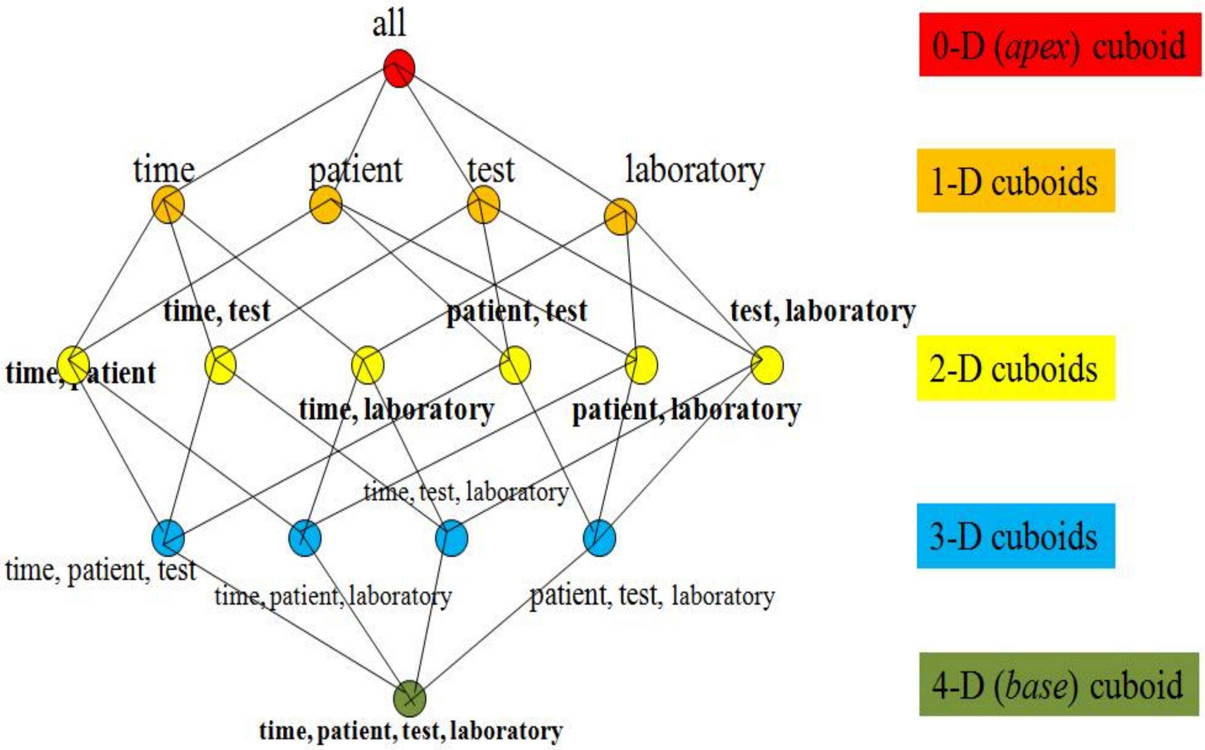


Fig. 2 Health Data Cube

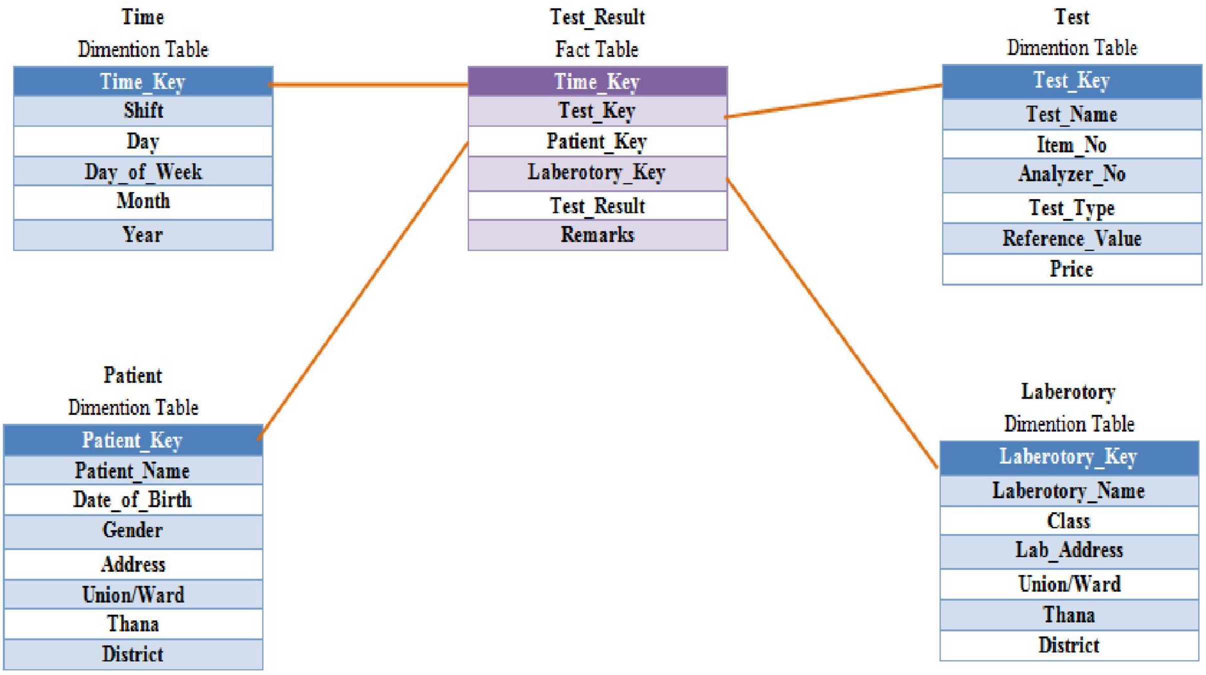


Fig. 3 Fact Table and Dimension Tables of National Health DW

4. DW Size Analysis

Let, test for a single patient = t_p ; total patients in class i = p_i

So total test for p_i patients = $\sum_{P=1}^{P=P_i} t_p$

total reports in class i = r_i ;

number of test in r_i = $\sum_{j=1}^{j=r_i} t_{rj}$

where t_{rj} is number of test in report j

Total number of tests T is given by

n=number of classes in laboratories

l_i = number of laboratories in class i

r_i =number of test reports in class i

t = number of test per report

We have derived the following equation to count the total number of test records (tuples in fact table) are generated by the different healthcare organizations such as hospitals and diagnostic centers.

$$T = \sum_{i=1}^{i=n} l_i \times r_i \times (\sum_{j=1}^{j=r_i} t_{rj} / r_i)$$

According to Directorate General of Health Services (DGHS) under the Ministry of Health and Family Welfare (MoHFW): Total number of government hospitals under DGHS is 592 and Government hospitals of secondary and tertiary levels under DGHS is 125[28], [29]. According to the Bangladesh Private Clinic and Diagnostic Owners Association (BPCDOA), 8,000 diagnostic centers of the country have DGHS approvals till December 2014 [30], [31]. So minimum number of places where pathological tests are performed is 8717. If for simplicity of calculation we consider average 500 patients reports are produced every day, each report consists 15 test attributes then putting the values in above equation we get:

T= 65377500 records(tuples /test attributes) per day.

So more than 65 million records will be added in the fact tables of National Health DW. Considering 1 record takes 0.2KB memory space than the DW will consume 12.50 GB memory/day. It is certainly falls under big data category and Bangladesh Government should go for Cloud storage and services for this [26, 32]. For fragmentation of big database there are several techniques such as CRUD matrix based fragmentation proposed by the authors of this paper [33, 34].

5. Data Preprocessing

Data preprocessing is one of the major task for developing a DW from heterogeneous sources. It includes data cleaning, missing values imputation, normalization, transformation etc. As for National Health DW, data are coming from different public and private hospitals, diagnostic centers and other sources, different preprocessing steps has been performed on data. Followings are some data preprocessing that we have performed. Table 1 and 2 present few of PCV Hct Red Cell Indices test data. Here reference values for female are 36-46 and for male are 40-50. The full dataset for this particular test contains 13296 records where the minimum and maximum results are 0.1 and 64 respectively. The results and the reference values are normalized using the Min-Max and Z-Score normalization techniques. Missing data are replaced by *class mean* method. Table 3 shows partial metadata for the same test dataset.

Table 1 Attribute subset selection and normalization of numeric data

TESTRESULT_ID	RESULT	Z Score normalized result	Min Max Normalized result
114080000000002	39.9	0.58	0.6228
114080000000204	39	0.41	0.6088
114080000000283	0.1	-6.91	0
114080000000609	0.2	-6.89	0.0016
114080000000755	0.1	-6.91	0
114080000000834	28.3	-1.6	0.4413
114080000000913	43	1.16	0.6714
114080000001138	29.7	-1.34	0.4632
114080000001279	37.8	0.18	0.59
114080000001436	37	0.03	0.5775
114080000001650	39	0.41	0.6088
114080000002071	39	0.41	0.6088
114080000002248	35	-0.34	0.5462
114080000003618	42	0.97	0.6557
114080000003766	41	0.79	0.6401
114080000003900	46	1.73	0.7183

Table 2 Reference values and their normalization

Reference values	Min-Max Normalization	Z Score Normalization
Female Lower: 36	0.561815336	-0.154685736
Female Upper:46	0.718309859	1.727135868
Male Lower: 40	0.624413146	0.598042906
Male Upper:50	0.780907668	2.479864509

Table 3 Metadata for the test result

Type	Value
Average	36.812
Maximum	64

Minimum	0.1
Standard Dev.	5.3070

Table 4 presents normalization technique of nominal data where metadata of Table 5 are used to replace result data for Urine colour diagnosis.

Table 4 Preprocessing of nominal data

TESTRESULT_ID	Result Type_ID	Min-Max Normalization
114080000002446	0	0.0000
114080000013324	1	0.1111
114080000098717	4	0.4444
114080000487792	6	0.6667
114080000743386	2	0.2222
114090000792554	3	0.3333
114090000822074	3	0.3333
114090000902763	7	0.7778
314080000143669	8	0.8889
314080000652184	5	0.5556
414080000046596	9	1.0000

Table 5 Metadata for Type_ID generation from Urine colour

Colour	Sample_Count	Type_ID
Straw	9439	0
Yellow	58	1
Reddish	32	2
L. Yellow	29	3
Milkly white	2	4
D. Yellow	2	5
Yellowish white	1	6
Reddish black	1	7
L.Reddish	1	8
Hazy	1	9

6 Data Mining from National Health DW

National Health DW can be used in many ways to improve national health standard, to provide better and prompt services to the patients and to facilitate health related research among the doctors, clinical researchers etc. In this section we are describing the use of National Health DW with two examples.

Example 1: National Reference level threshold finding

Table 6 WHO's Hemoglobin thresholds to define anemia

Age/Gender group	Hb threshold(g/dl)
Children(0.5-5yrs)	11.0
Children(5-12yrs)	11.5
Teens(12-15yrs)	12.0
Women, non-pregnant(>15yrs)	12.0
Women, pregnant	11.0
Men(>15yrs)	13.0

In Bangladesh, rule of thumbs is for *Woman > 15 years, non pregnant,* Hb> 11 Good; Hb >=10.5 ok and if Hb < 10 (g/dl), medication for the patient is needed. This is slightly different from WHO threshold [35]. Using National Health DW National Reference level for different clinical values can be found by data mining.

Example 2: Fraud Testing Awareness
If For a Costly Test T1:
Age(X,<30) ^ (Gender='M') => Negative (X, T1)
[Support =85%, Confidence=98%]
From confidence value it can be clearly identified that this test T1 has almost no impact of disease diagnosis. National awareness can be developed not to perform the test at initial level for Young Males.
In this way many other interesting patterns can be derived from National Health DW by using various data mining algorithms like association or clustering.

7. Conclusions

This paper presented the developmental stages of National Health DW platform for the management, processing and analysis of large-scale Health data modeled for e-health system. In this paper, widely accepted conceptual and logical design approaches in DW design are discussed. Considering the quality factors and the information requirements, star schema was chosen as the most suitable logical model for the purpose. Establishing a data warehouse gathering huge data from existing Health databases should give easier and better access to interesting data for researchers, health service providers and govt. authorities. In order to get maximum benefit from the model presented in this research, the conditions mentioned below should be satisfied.

1. There should be a document which clearly defines the structure of the data tables currently used by the concerned Pathological centers.
2. Stakeholders should clearly know what the data retrieval operations are going to be executed using the data warehouse.
3. There should be strong cooperative mind among different health service providers to help the governmental bodies for successful implementation.

Acknowledgment

This research is funded by ICT Division, Ministry of Posts, Telecommunications and Information Technology, Government of the People's Republic of Bangladesh.
Authors are grateful to Mr. Mamunur Rashid and Muhammed Dastagir Husain of Dept. of CSE, BUET for their valuable discussions.

References

[1] Roddick JF, Fule P, Graco WJ (2003) Exploratory medical knowledge discovery: experiences and issues. SIGKDD Explor. Newsletter, 5(1): 94-99

[2] Cios K (2002) Uniqueness of medical data mining. Artificial intelligence in medicine. 26:1-24

[3] Fayyad UM, Shapiro GP, Smyth P (1996) From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining 1–36

[4] Khosla R, Dillon T (1997) Knowledge Discovery, Data Mining and Hybrid Systems. Engineering Intelligent Hybrid Multi-Agent Systems, Kluwer Academic Publishers 143–177

[5] Inmon WH (1992) EIS and the data warehouse: a simple approach to building an effective foundation for EIS. Database Programming and Design, 5(11): 70-73

[6] Stolba N, Banek M, and Tjoa AM (2006) The Security Issue of Federated Data Warehouses in the Area of Evidence-Based Medicine. First International Conference on Availability, Reliability and Security (ARES'06, IEEE)

[7] Sahama TR, Croll PR (2007) A Data Warehouse Architecture for Clinical Data Warehousing. Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007)

[8] Lyman JA, Scully K, Harrison JH (2008) The development of health care data warehouses to support data mining. Clin Lab Med. 28(1):55-71

[9] Inmon, W (2005): Building the Data Warehouse, 4th edition, Wiley-New York.

[10] Jiawei H, Micheline K, Jian P (2012) Data Mining Concepts and Techniques 3rd Edition, Elsevier

[11] Kimball R, Ross M (2013) The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling 3rd Edition, Wiley

[12] Nugawela S (2013) Data Warehousing Model For Integrating Fragmented Electronic Health Records From Disparate And Heterogeneous Clinical Data Stores, M.Sc. Thesis, Queensland University of Technology

[13] Mullins M, Siadaty MS, Lyman J et al (2006) Data mining and clinical data repositories: Insights from a 667,000 patient data set. Comput. Biol. Med. 36: 1351–1377

[14] Zhu X, Khoshgoftaar T, Davidson I, Zhang S (2007) Special issue on mining low-quality data, Knowledge and Information Systems, 11:131-136

[15] Brown ML, Kros JF (2003) Data mining and the impact of missing data. Industrial Management & Data Systems 103: 611-621

[16] Lavrač N (1999) Selected techniques for data mining in medicine. Artificial intelligence in medicine 16(1): 3-23

[17] Lee IN, Liao SC, Embrechts M (2000) Data mining techniques applied to medical information. Medical Informatics & the Internet in Medicine 25(2): 81-102

[18] Obenshain MK, Application of Data Mining Techniques to Healthcare Data, Infection Control and Hospital Epidemiology, vol.25, no 8, pp. 690-695, 2004

[19] Wang, H, Wang S (2008) Medical knowledge acquisition through data mining,. IEEE International Symposium ITME.

[20] S. FU (2011) Missing Data in Pathology Databases. MSc Thesis, Australian National University.

[21] Partington SN, Papakroni V, Menzies T (2014) Optimizing data collection for public health decisions: a data mining approach. BMC Public Health 14: 593-598

[22] Cubillas JJ, Ramos MI, Feito FR, Ureña T (2014) An improvement in the appointment scheduling in primary health care centers using data mining. J. Med. Syst., Springer 38: 89

[23] Hoque ASML, Galib S, Tasnim M (2013) Mining Pathological Data to Support Medical Diagnostics. Workshop on Advances on Data Management: Applications and Algorithms, Department of Computer Science and Engineering, BUET, Dhaka, 71-74

[24] Kumari S, Singh A (2013) A data mining approach for the diagnosis of diabetes mellitus. IEEE 7th International Conference on Intelligent Systems and Control

[25] Yilmaz N, Inan O, Uzer MS (2014) A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases," J. Med. Syst., Springer, 38

[26] Herland M, Khoshgoftaar TM, Wald R (2014) A review of data mining using big data in health informatics. J. Big Data, Springer 1: 2

[27] Khan SI, Hoque ASML (2015) Towards Development of Health Data Warehouse: Bangladesh Perspective. Accepted in 2nd International Conference on Electrical Engineering

and Information & Communication Technology (iCEEiCT 2015).

[28] HEALTH BULLETIN 2014, 2nd Edition, DGHS, Ministry of Health and Family Welfare, Government of the People's Republic of Bangladesh

[29] http://www.dghs.gov.bd/index.php/en/health-program-progress/hpnsdp-2011-16/84-english-root/ehealth-eservice/497-hpnsdp-2011-16-brief. Accessed 20 Feb 2015

[30] http://www.bpcdoa.com/clinics_and_diagnostics.html. Accessed 22 Feb 2015

[31] http://www.thefinancialexpress-bd.com/2014/12/15/71077/print Accessed 22 Feb 2015

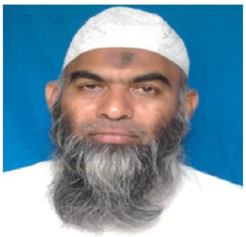
[32] Liang Z, Sherif S, Anna L, Athman B (2014) Cloud Data Management-Springer Switzerland

[33] Khan SI, Hoque ASML (2010) A New Technique for Database Fragmentation in Distributed Systems, International Journal of Computer Applications 5 (9), 20-24.

[34] Khan SI, Hoque ASML (2012) Scalability and Performance Analysis of CRUD Matrix Based Fragmentation Technique for Distributed Database, in Proceedings of 15th International Conference on Computer and Information Technology (ICCIT), 562-567.

[35] World Health Organization (2008). Worldwide prevalence of anaemia 1993–2005. Geneva: World Health Organization. ISBN 978-92-4-159665-7

Shahidul Islam Khan obtained his B.Sc. and M.Sc. Engineering Degree in Computer Science and Engineering (CSE) from Ahsanullah University of Science & Technology (AUST) and Bangladesh University of Engineering & Technology (BUET), Dhaka, Bangladesh in 2003 and 2011 respectively. He is now a Doctoral Student in the Department of CSE, BUET, which is the highest ranked technical university of Bangladesh. His current field of research is data mining and health informatics. He has more than 10 published papers in international conferences and journal. He is also an Assistant Professor (Currently in study leave) in the Dept. of CSE, International Islamic University Chittagong (IIUC), Bangladesh.



Abu Sayed Md. Latiful Hoque graduated from the Dept. of Electrical and Electronic Engineering (EEE), Bangladesh University of Engineering & Technology (BUET), Dhaka, Bangladesh in 1986. He obtained Post Graduate Diploma in 1992 from Asian Institute of Technology (AIT), Thailand and Ph.D. in CS from University of Strathclyde, Glasgow, UK in 2003. He is a professor of the Dept. of CSE, BUET and a prominent international researcher in the field of Database, Data Mining, E-Learning. He has near about 50 published papers in reputed international journals and conferences. He is also author of a book on Database Systems which is taught in Universities.

Data Mining Smart Energy Time Series

Janina POPEANGĂ
University of Economic Studies, Bucharest, Romania
janina.popeanga@yahoo.com

With the advent of smart metering technology the amount of energy data will increase significantly and utilities industry will have to face another big challenge – to find relationships within time-series data and even more - to analyze such huge numbers of time series to find useful patterns and trends with fast or even real-time response.

This study makes a small review of the literature in the field, trying to demonstrate how essential is the application of data mining techniques in the time series to make the best use of this large quantity of data, despite all the difficulties.

Also, the most important Time Series Data Mining techniques are presented, highlighting their applicability in the energy domain.

Keywords: *Time Series Data Mining, Clustering, Classification, Motif Discovery, Data Reduction*

1 Introduction

The increasing deployment of Automated Meter Reading (AMR) systems has created new challenges for utilities industry in terms of how to utilize the recorded data, not only to improve the day-to-day operations. The smart grid will provide a large volume of sensor and meter data that will require intelligent analytics that move further than data management, querying and reporting. To make the best use of this large quantity of data, it is essential to apply data mining techniques to extract relevant information valuable to the utilities industry. [1]

Data mining is defined as a “type of database analysis that attempts to discover useful patterns or relationships in a group of data. The analysis uses advanced statistical methods, such as cluster analysis, and sometimes employs artificial intelligence or neural network techniques. A major goal of data mining is to discover previously unknown relationships among the data, especially when the data come from different databases.” [2]

The paper is structured as follows: Section 2 focuses on the concept of Time-Series Data Mining. Section 3 presents the most important techniques widely used in this domain. Section 4 makes a review of relevant literature. Section 5 presents some software tools for time-series data mining, while Section 6 ends this article presenting the conclusions.

2. The Time Series Data Mining Concept

Time series data type, also called chronological series or simply time series represent results of measurements made on the characteristics of a unit of population studied, over time, at successive moments of its evolution in some time intervals.

A time series T of size n is an ordered set of n real-value variables, where

$$T = (t_1, t_2, \ldots, t_n).$$

A subsequence of length p of time series T is a sampling of length p < n of adjacent positions from T, where

$$Ti,p = (t_i, t_{i+1}, \ldots, t_{i+p-1}),$$
$$\text{for } 1 \le i \le n - p + 1.$$

The time intervals for which smart energy measurements are made may include: hours or fractions of hours, days, weeks, decades, months, quarters, semesters, years. Since the intervals are equal and represent the passage of time, the observations resulting from these measurements are successive usually equidistant in time.

There are two ways to analyze time series data:

- ***Time domain analysis*** which studies how a time series process evolves through time;
- ***Frequency domain analysis*** (spectral analysis) which studies how periodic components at different frequencies illustrate the evolution of a time series.

According to Hui et. al., Data Mining "*is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses, or other information repositories.*" The most used data mining functions in commercial and research analysis are association, classification, prediction and clustering.

These data mining techniques have also been applied to other types of data such as time-series, telecommunications, web, spatial, and multimedia data. [3]

Applying the principles and techniques of classical data mining in the time series analysis had led to the concept called ***Time Series Data Mining.***

Energy data (production and consumption) recorded over a period of time at fixed intervals is a classic time series data mining problem.

The steps taken in the entire process are:

- collect data from multiple sources: web, text, databases, data warehouses, sensors, smart devices;
- data filtering by eliminating errors. When using a data warehouse, this process is removed because a process of extraction, transformation and loading (ETL) was already applied on the data;
- establishing key data attributes that will participate in the DM process, by selecting those properties that interest the analysis;
- application of templates and detection/analysis of new knowledge;
- visualization, validation and evaluation of results.

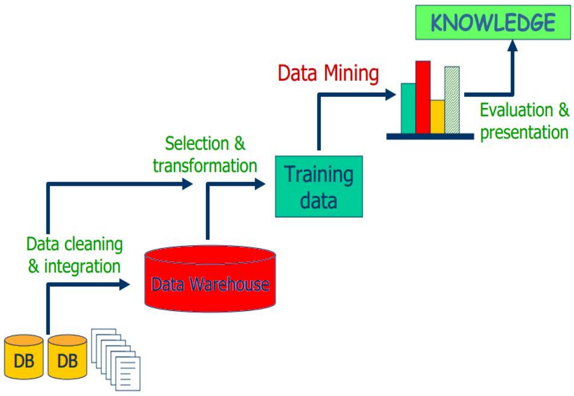

KNOWLEDGE
Data Mining
Evaluation & presentation
Selection & transformation
Training data
Data cleaning & integration
Data Warehouse
DB
DB

Fig. 1. Data Mining process

With the increasing deployment of a large number of sensors and other smart devices, the amount of time-series data is growing rapidly, often in the order of gigabytes per day or even per minute. The big challenge is to find correlation relationships within time-series data and even more - to analyze such huge numbers of time series to find similar or regular patterns and trends with fast or even real-time response.

3. Time Series Data Mining Techniques

Classification and clustering have the effect of dividing the objects in some classes, but the difference is how this division is made. The classification classes are predetermined, and at clustering they are built taking into account the characteristics of the analyzed objects.

The ***time series clustering technique*** can be divided into three main categories:

- *whole series clustering* – the purpose is to regroup complete time series into clusters so that the time series are as similar to each other as possible within each cluster;
- *subsequence clustering* – the purpose is to find similarity among different time subsequences and to group them in the same cluster. It is used

for discovering structures or patterns in time series data;

- *time point clustering* – the purpose is to find clusters of the time point. It is similar to time series segmentation, however, time point clustering is different from segmentation by the fact that not all points must be assigned to the cluster, some of points being considered noise.

The ***time series classification technique*** can be divided into two types of classification:

- *time series classification* – the purpose is to mapped complete time series into predefined classes. It is used to match a whole sequence to a class of other sequences;
- *subsequence classification* – the purpose is to check if a subsequent belongs to a certain sequence.

Association finds some rules and models in data analyzed and sequence analysis seeks some models on some data ordered in time, trying to determine a way to order these models, assuming that between events there are certain implications.

According to the number of involved time series, the ***association rules technique*** can be divided into two categories:

- *association rules mining from single series* – the purpose is to find rules and sequential patterns;
- *association rules from multiple series* – it is divided into two types:
- *intra-transactional association rules mining* – the purpose is to make known the co-relations of multiple time series at same time. Taking the renewable energy production as an example, we can find rules like "If wind speed goes up and sun intensity raises then the renewable energy production increases on the same day";
- *inter-transactional association rules* – the purpose is to make known the co-relations of different series at different time. Taking the energy consumption and renewable energy production as an example, we can find rules like "If the number of solar panels installed increases on the first day and the number of hours of sunlight increases on the second day then the energy consumption from the electricity common network will decrease on the third day".

Regression treats a general trend that, if carried out in time, can be used at forecast, while the **forecast** function takes into account other factors, including cyclical events etc. Many time series prediction application involve regression analysis, for example, predicting future energy consumption based on historical data and other information like outside temperature, house characteristics, living habits, number of person hours spent at home, electrical appliances and devices, number of electric heaters in the home, etc.

Exception analysis may be related to clustering and classification, which may find exceptions either observing that certain groups are composed of a single object very different from the other (grouping) or when an object does not fall into any class (classification).

The purpose of detecting anomalies in a time series is to find abnormal subsequences in that series, which means to find subsequences that do not follow the model of a series normal behavior.

Another data mining technique, available for time series is ***motif discovery***. The purpose of this task is to find every sequence that appears repeatedly in a time series. The sequence can be known from the beginning or not. Given a sequence as pattern, this technique performs a search to find other sequences that are similar with the pattern, but the search for unknown motifs is a more complex problem because all subsequences of all possible lengths have to be compared.

In my opinion, data mining techniques can be used in energy management to solve, for example, problems such as:

- ***Classification*** – determining consumer profiles based on different variables, determining the possibility of purchase and install special equipment for renewable energy generation, based on user profile;
- ***Regression*** – time trend analysis of the energy consumption and production, monitoring the effect of energy policies and measures;
- ***Forecast*** – predicting future energy production and consumption;
- ***Anomaly detection*** – consumer fraud detection, network intrusion and other unusual and rare events that are hard to find;
- ***Motif discovery*** – identify energy relationships that can aid in the process of forecasting, identify patterns that can be used to predict customers behavior.
- ***Association rules*** – analyzing the links between certain factors that could cause increased/decreased energy consumption or production;
- ***Clustering*** – locating fraud or high energy consumption, getting a group of them around certain areas.

Contrasting the traditional techniques for the time series analysis and limiting assumptions, the methods in the Time Series Data Mining network can be successfully applied to identify the complex characteristics, and to predict the non-periodic, non-linear, irregular and chaotic time series.

4. Research Trends and Issues

Data Mining techniques have been used by many researchers in energy system applications.

Tso and Yau have used three data mining techniques (regression analysis, decision tree and neural networks) for understanding energy consumption patterns and predicting electricity consumption. [4]

Azadeh et al. have proposed an integrated fuzzy system, data mining and time series framework to estimate and predict electricity demand for seasonal and monthly changes in electricity consumption especially in developing countries such as China and Iran with non-stationary data. [5]

Kusiak et. al. have applied eight data-mining algorithms to model the nonlinear relationship among energy consumption, control settings (supply air temperature and supply air static pressure), and a set of uncontrollable parameters, for minimization of the energy to air condition a typical office-type facility. [6]

Yu et al. have demonstrated that the use of decision tree method can classify and accurately predict building energy demand levels, using real data from Japanese residential buildings. The competitive advantage of decision tree method over other broadly used modelling techniques, such as regression method and ANN method, lies in the ability to generate accurate predictive models with interpretable flowchart-like tree structures that enable users to quickly extract useful information. [7]

Figueiredo et. al. have presented an electricity consumer characterization framework based on DM techniques. Using real data of consumers from the Portuguese distribution company, the load profiling module created a set of consumer classes using a clustering operation and the representative load profiles for each class. The classification module builded a classification model able to assign different consumers to the existing classes. [8]

Dent et. al. have described the bottom up and clustering methods for defining representative load profiles for domestic electricity users in the UK. Investigation of electricity consumption in order to determine similarities between types of consumers required that the day's usage pattern to be summarized so it can be compared with others. [9]

Time Series Data Mining has been an ever-growing and stimulating field of study that has continuously raised challenges and research issues over the past years.

As one of the most prominent issues arises from the high dimensionality of time series data, researchers have proposed methods of how to reduce the size of the data without substantial loss of information.

Data reduction is often the first step to combating massive time series data, as it will provide a summary of the data.

As Miles et. al. [10] clarify, “data reduction is not something separate from analysis. It is part of analysis. The researcher’s decisions—which data chunks to code and which to pull out, which evolving story to tell—are all analytic choices. Data reduction is a form of analysis that sharpens, sorts, focuses, discards, and organizes data in such a way that final conclusions can be drawn and verified”.

Discrete Fourier Transform, Singular Value Decomposition, Discrete Wavelet Transform and Random Projection (Sketches) are the most frequently used dimensionality reduction techniques.

Discrete Fourier Transform, first introduced by Agrawal et. al., is used to transform time series into frequency domain and selects the most important coefficients.

Giving the fact that energy storage can be a practical solution to balance energy production against its consumption, Makarov et. al. have proposed to use Discrete Fourier Transform to decompose the required balancing power into different time-varying periodic components, i.e., intra-week, intra-day, intra-hour, and real-time. This approach was used in a study conducted for the 2030 Western Electricity Coordinating Council (WECC) system model and was considered a success. [11]

Xiao et. al. have proposed to use Discrete Fourier Transform (DFT) for coordinated sizing of energy storage and diesel generators in an isolated microgrid. The DFT-based coordinated dispatch strategy allocates balance power between the two components through frequency-time domain transform. [12]

Saenthon et. al. have demonstrated the effectiveness and high accuracy of the Discrete Fourier Transform to automatically identify load type of electrical appliances by transforming time domain to frequency domain. [13]

Discrete Wavelet Transform is used to approximate the data by dividing the sequences into equal-length sections and saving the weighted value of these sections.

Chang has used the Wavelet Transform (WT) of the time-frequency domain to analyze and detect the transient physical behavior of loads during the load identification. According to his research the discrete wavelet transform (DWT) is more suitable than short-time Fourier transform for transient load analyses. [14]

Nieto-Hidalgo et. al. have demonstrated that the Wavelet Transform could be used to identify simpler electrical consumption patterns as a part of total consumption curve. They have proposed an innovative method based on WT to decompose the global power consumption in elemental loads corresponding to each appliance. [15]

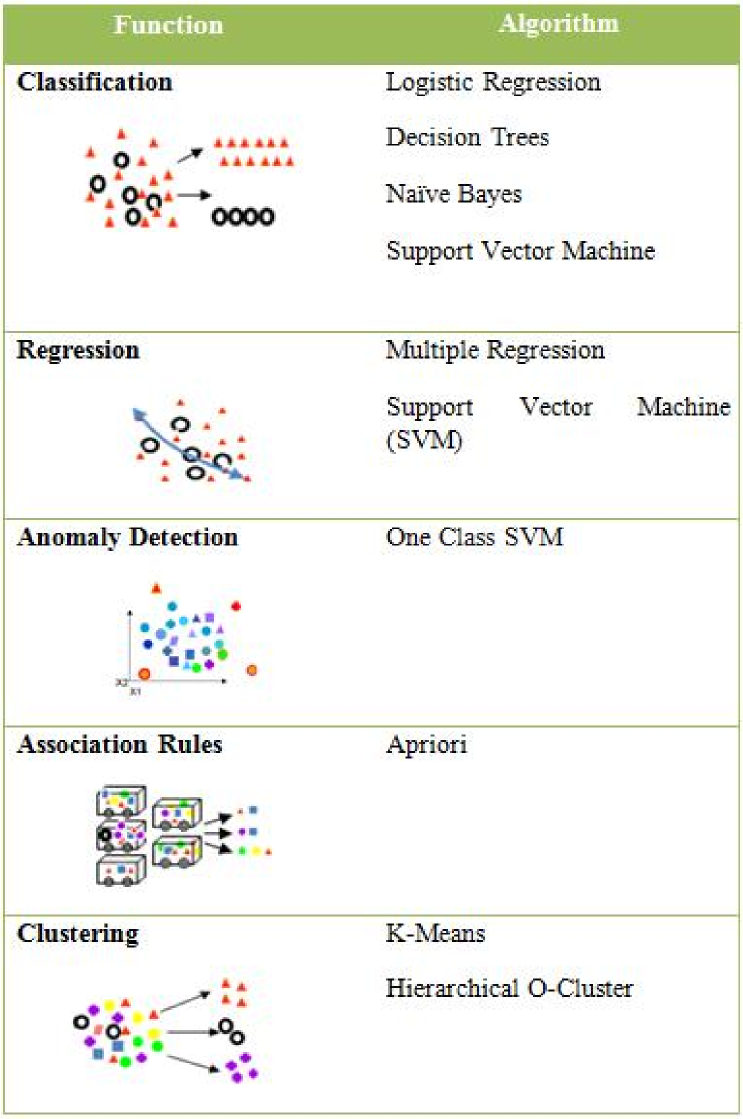
Abu-Shikhah and Elkarmi have proposed the using of the SVD (singular value decomposition) technique in a new combination which uses hourly loads of successive years to predict hourly loads and peak load for the next selected time span. After filtering out the load trend of the Jordanian power system, they have applied the SCV technique to de-noise the resulting signal. [16]

Dieb Martins and Gurjao have applied the random projection technique in order to obtain a reduced version (sketch) of smart meters' original data, thus increasing the processing throughput of the utility. Using real smart meters measurements, they have demonstrated that processing using sketches sized 50% smaller than original data can achieve a 2% average relative error while presenting greater data rates. [17]

5. Time Series Data Mining Tools

In recent years, Oracle offered more and better statistical and time series analysis capabilities through *Oracle Data Mining*. ODM provides several kinds of data-mining algorithms for functions such as clustering, classification, regression, anomaly detection, association rules, sequence similarity etc.

Working directly with ODM to analyze time series data stored in Oracle databases or data warehouses, can reduce resources consumption and the time needed for data transmission.



Function	Algorithm
Classification	Logistic Regression Decision Trees Naïve Bayes Support Vector Machine
Regression	Multiple Regression Support Vector Machine (SVM)
Anomaly Detection	One Class SVM
Association Rules	Apriori
Clustering	K-Means Hierarchical O-Cluster

Fig.2. SQL Data Mining Algorithms

Microsoft SQL Server Analysis Services (SSAS) contains online analytical processing (OLAP) and data mining functionality for business intelligence applications.

SSAS provides data mining techniques like: classification, regression, clustering, association algorithms, sequence analysis etc.

Mining historical data using SSAS can offer new visions and create a basis for forecasting, and this may be extremely interesting for analysis of time series energy data.

The *Microsoft Time Series* algorithm offers regression algorithms that are optimized to forecast continuous values, such as energy production or consumption, over time.

The Time Series Data Mining nodes in *SAS Enterprise Miner* (Time Series Data Preparation Node, Time Series Similarity Node, and Time Series Exponential Smoothing Node) significantly improve the time series analysis and data preparation capabilities of the data miner. Finding time series that unveil similar statistical characteristics permits identifying customer behaviors in large volumes of time series energy data.

Also, with the large volumes of energy consumption and production data stored in time series, the power to integrate this data into analysis workflows will help utilities to build valuable models more easily.

R is a free software with statistical and graphical capabilities which runs on a wide variety of UNIX, Windows and MAC OS platforms. R provides time-series analysis and data mining techniques such as: TS forecasting, TS clustering, classification and association rules etc. The most popular models used to predict energy consumption with R based on known past data are: autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA).

Weka's time series framework proposes a data mining approach to model time series by transforming the data into a form that can be processed by standard data mining algorithms. The solution is to remove the temporal ordering of individual input examples by encoding the time dependency via additional input fields. These fields are sometimes referred to as "lagged" variables. Various other fields are also computed automatically to allow the algorithms to model trends and seasonality. After the data has been transformed, any method capable of predicting a continuous target can be applied. This approach to time series analysis and forecasting is often more powerful and more flexible that classical statistical techniques such as ARMA and ARIMA. [18]

Handling time series analysis in a tool like *RapidMiner* requires advanced skills, but this did not prevent its success. RapidMiner has over 3 million total downloads and has over 200,000 users including eBay, Intel, PepsiCo and Kraft Foods as paying customers.

According to Bloor Research, RapidMiner provides 99% of an advanced analytical solution through template-based frameworks that speed delivery and reduce errors.

RapidMiner's approach to time series is based on two main data transformation processes [19]:

- Windowing to transform the time series data into a generic data set
- Applying any of the data mining algorithms to predict the target variable and thus predict the next time step in the series.

For the second year in a row, Gartner Research has placed RapidMiner in the Leaders Quadrant in the Magic Quadrant for Advanced Analytics Platforms, describing this tool as a "*platform that supports an extensive breadth and depth of functionality, and with that it comes quite close to the market Leaders.*" [20]

6. Conclusions

The concept behind this paper, Time Series Data Mining, can be defined as the process of discovering useful patterns and significant structures, unknown associations and relationships, anomalies and motifs, which can be used to predict future events and behaviors.

We have made a small review of the most important techniques, highlighting our proposal of their applicability in the energy domain.

As can be seen from the research trends presented, Time Series Data Mining methods have been applied successfully in a wide range of energy system applications and for issues arisen from the high dimensionality of time series data, have been proposed various methods of how to reduce the size of the data without substantial loss of information.

Acknowledgment
This work was cofinanced from the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/142115 „Performance and excellence in doctoral and postdoctoral research in Romanian economics science domain”

References

[1] J. Popeangă, *Building Consumer Profiles through Time-Series Data Mining Techniques*, Abstracts Book, 4th World Conference on Business, Economics and Management (BEM-2015), 30 April - 02 May 2015.

[2] Sunita, Prachi, *Efficient Cloud Mining Using RBAC (Role Based Access Control) Concept*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013, page 679.

[3] Hui, S. C., Jha, G., *Data mining for customer service support*, Information &Management, Volume 38, Issue 1,2000, pp. 1-13.

[4] G. K. F. Tso, K. K. W. Yau, *Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks*, Energy 32, 2007, 1761–1768.

[5] A. Azadeh, M. Saberi, S. F. Ghaderi, A. Gitiforouz , V. Ebrahimipour, *Improved estimation of electricity demand function by integration of fuzzy system and data mining approach*, Energy Conversion and Management 49 (2008) 2165–2177.

[6] A. Kusiak, M. Li, F. Tang, *Modeling and optimization of HVAC energy consumption*, Applied Energy 87, 2010, 3092–3102.

[7] Z. Yua, F. Haghighata,, B.C.M. Fungb, H. Yoshinoc, *A decision tree method for building energy demand modelling*, Energy and Buildings 42 (2010) 1637–1646.

[8] V. Figueiredo, F. Rodrigues, J. G. Gouveia, *An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques*, IEEE Transactions on Power Systems, 20(2005) 596-602.

[9] I. Dent , U. Aickelin and T. Rodden, *The application of a data mining framework to energy usage profiling in domestic residences using UK data*, Proceedings of the Research Students' Conference on “Buildings Don't Use Energy, People Do?” – Domestic Energy Use and CO2 Emissions in Existing Dwellings 28 June 2011, Bath, UK.

[10] M. B. Miles and M. A. Huberman, *Qualitative Data Analysis: An Expanded Sourcebook* (2nd ed.), Thousand Oaks, Calif.: Sage, 1994.

[11] Y. V. Makarov, M. C.W. Kintner-Meyer, P. Du, C. Jin, and H. F. Illian, *Sizing Energy Storage to Accommodate High Penetration of Variable Energy Resources*, IEEE Transactions on Sustainable Energy, (Volume:3, Issue: 1), ISSN :1949-3029, 2011.

[12] J. Xiao, L. Bai, F. Li, H. Liang, C. Wang, *Sizing of Energy Storage and Diesel Generators in an Isolated Microgrid Using Discrete Fourier Transform (DFT)*, IEEE Transactions on Sustainable Energy, ISSN :1949-3029, 2014.

[13] A. Saenthon and S. Kaitwanidvilai, *Load Identification in Household Apparatus Equipment using Discrete Fourier Transform with Proper Window Function*, Proc. Power and Energy Systems (AsiaPES 2013), track 800-144, 2013.

[14] H.Chang, *Non-Intrusive Demand Monitoring and Load Identification for Energy Management Systems Based on*

Transient Feature Analyses, Energies, 2012, vol. 5, issue 11, pages 4569-4589.

[15] M. Nieto-Hidalgo, F. J. Ferrández-Pastor, J. M. García-Chamizo, V. Romacho-Agud and F. Flórez-Revuelta, *Using Wavelet Transform to Disaggregate Electrical Power Consumption into the Major End-Uses*, Paper presented at the 8th International Conference on Ubiquitous Computing & Ambient Intelligence UCAmI 2014.

[16] N. Abu-Shikhah and F. Elkarmi, *Medium-term electric load forecasting using singular value decomposition, Energy*, Volume 36, Issue 7, July 2011, Pages 4259-4271.

[17] A. D. Martins and E.C. Gurjao, *Processing of smart meters data based on random projections*, 2013 IEEE PES Conference On Innovative Smart Grid Technologies Latin America (ISGT LA), ISBN: 978-1-4673-5272-7, Sao Paulo, 2013.

[18] Pentaho, *Time Series Analysis and Forecasting with Weka*, 2015, Link: http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka

[19] Simafore, *Using RapidMiner for time series forecasting in cost modeling: 1 of 2*, 2012, Link: http://www.simafore.com/blog/bid/106430/Using-RapidMiner-for-time-series-forecasting-in-cost-modeling-1-of-2

[20] RapidMiner, *RapidMiner: Leader in Gartner Research Magic Quadrant for Advanced Analytics Platforms*, Garnter, 2015

Janina POPEANGĂ graduated in 2010 the Faculty of Cybernetics, Statistics and Economic Informatics, Economic Informatics specialization. The title of her Bachelor's thesis is "*Distributed Databases*". In 2012, she graduated the Databases for Business Support master program with the thesis "*Monitoring and management of electric power consumption using sensorial data*". Janina's interests are broadly in the fields of databases and distributed systems. Since 2012 she is a Ph.D. Student in the Doctoral School of Bucharest Academy of Economic Studies. Her research focuses on real-time database systems, business intelligence analytics, sensor data management, smart grid and renewable energy.

Big Data Analytics Platforms analyze from startups to traditional database players

Ionuţ ŢĂRANU
Bucharest University of Economic Studies
ionut.tanaru@gmail.com

Big data analytics enables organizations to analyze a mix of structured, semi-structured and unstructured data in search of valuable business information and insights. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. With so many emerging trends around big data and analytics, IT organizations need to create conditions that will allow analysts and data scientists to experiment. "You need a way to evaluate, prototype and eventually integrate some of these technologies into the business," says Chris Curran[1]. In this paper we are going to review 10 Top Big Data Analytics Platforms and compare the key-features.

Keywords: *Big data, In-memory, Hadoop, Data analysis*

1 Introduction

The growth of data – both structure and unstructured – will present challenges as well as opportunities for organisations over the next five years.

With growing data volumes, it is essential that real-time information that is of use to the business can be extracted from its IT systems, otherwise the business risks being swamped by a data deluge. Meanwhile, competitors that use data to deliver better insights to decision-makers stand a better chance of thriving through the difficult economy and beyond. To analyze such a large volume of data, big data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, forecasting and data optimization. Collectively these processes are separate but highly integrated functions of high-performance analytics.

Today's advances in analyzing Big Data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Facebook. The business cases for leveraging Big Data are compelling. For instance, Netflix mined its subscriber data to put the essential ingredients together for its recent hit House of Cards, and subscriber data also prompted the company to bring Arrested Development back from the dead.

Another example comes from one of the biggest mobile carriers in the world. France's Orange launched its Data for Development project by releasing subscriber data for customers in the Ivory Coast. The 2.5 billion records, which were made anonymous, included details on calls and text messages exchanged between 5 million users. Researchers accessed the data and sent Orange proposals for how the data could serve as the foundation for development projects to improve public health and safety. Proposed projects included one that showed how to improve public safety by tracking cell phone data to map where people went after emergencies; another showed how to use cellular data for disease containment.[2] So it seems that data analysis is a do-or-die requirement for today's businesses. We analyze below notable vendor choices, from Hadoop upstarts to traditional database players.

2. Top 10 Big Data Analytics Platforms

2.1. Actian (**Fig.1.**– Actian Analytics Platform).

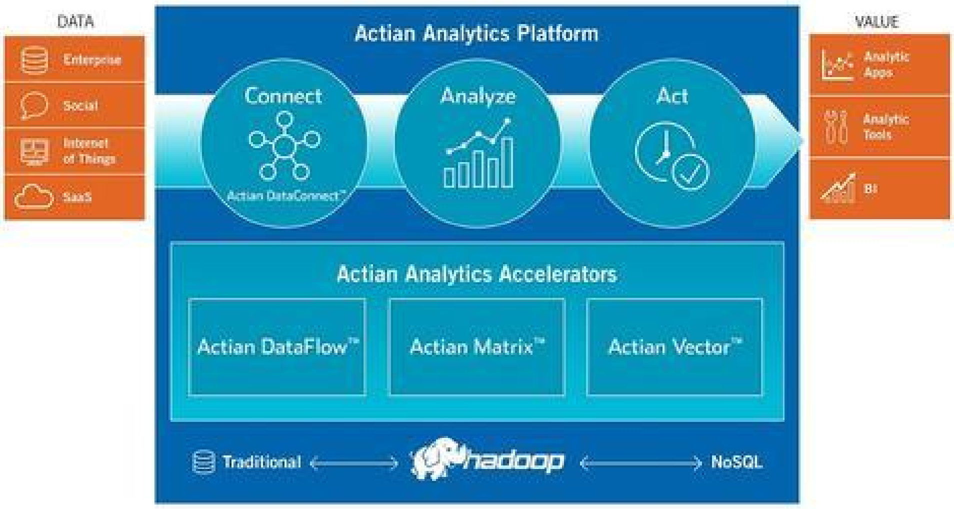


Fig. 1. Actian Analytics Platform

- **Analytical DBMS:** Actian Matrix (formerly ParAccel), Actian Vector (formerly Vectorwise).
- **In-memory DBMS**: Actian Matrix In-Memory Option (data stored to both memory and disk).
- **Hadoop distribution:** None.
- **Stream-processing technology:** None.
- **Hardware/software systems:** None (software-only vendor).

The company is counting on the combination of fast, analytical DBMS options, cloud services, and data-integration and -analytics software geared to a world in which Hadoop is a prominent fixture of the data-management architecture. Actian DataFlow includes SQL-, ETL-, and data-cleansing-on Hadoop options that work with distributions from Apache, Cloudera, Hortonworks, and others [3]

2.2. Amazon

- **Analytical DBMS:** Amazon Redshift service (based on ParAccel engine); Amazon Relational Database Service.
- **In-memory DBMS:** None. Third-party options on AWS include Altibase, SAP Hana, and ScaleOut.
- **Hadoop distributions:** Amazon Elastic MapReduce. Third-party options include Cloudera and MapR.
- **Stream-processing technology:** Amazon Kinesis.
- **Hardware/software systems:** Not applicable.

AWS is located in 11 geographical "regions": US East (Northern Virginia), where the majority of AWS servers are based, US West (northern California), US West (Oregon), Brazil (São Paulo), Europe (Ireland and Germany), Southeast Asia (Singapore), East Asia (Tokyo and Beijing) and Australia (Sydney). There is also a "GovCloud", based in the Northwestern United States, provided for U.S. government customers, complementing existing government agencies already using the US East RegionEach Region is wholly contained within a single country and all of its data and services stay within the designated Region.

Amazon Web Services 2009 (**Fig. 2.**– Amazaon Web Service) hosts a who's who list of data-management services from third-party players -- Cloudera, Microsoft, Oracle, SAP, and many others -- but the cloud giant has its own long-term ambitions where big-data analysis is concerned.[4] Building on its Elastic Compute Cloud (EC2) and Simple Storage Service (S3) storage infrastructure, Amazon launched its Hadoop-based Elastic MapReduce service way back in. In 2013, AWS added the Redshift Data Warehousing service (based on the ParAccel DBMS), which is supported by another who's who list of independent data-integration, business intelligence, and analytics vendors. Rounding out AWS's big-data capabilities are the DynamoDB NoSQL database management service and Kinesis Stream Processing service.

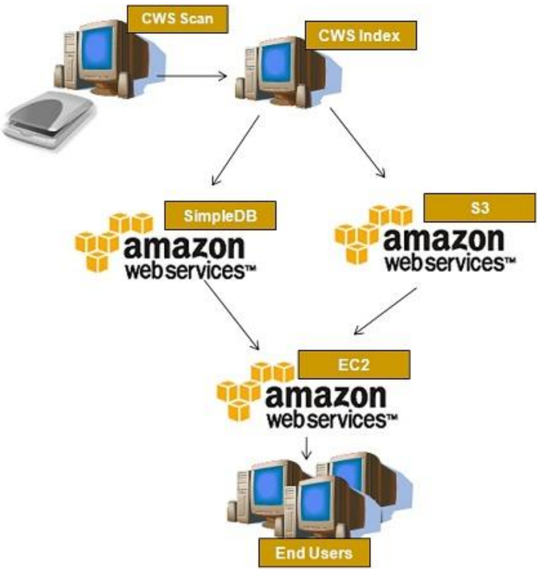


Fig. 2. Amazon Web Service

- Amazon DynamoDB provides a scalable, low-latency NoSQL online Database Service backed by SSDs.
- Amazon ElastiCache provides in-memory caching for web applications. This is Amazon's implementation of Memcached and Redis.
- Amazon Relational Database Service (RDS) provides a scalable database server with MySQL, Oracle, SQL Server, and PostgreSQL support.[22]
- Amazon Redshift provides petabyte-scale data warehousing with column-based storage and multi-node compute.
- Amazon SimpleDB allows developers to run queries on structured data. It operates in concert with EC2 and S3 to provide "the core functionality of a database".
- AWS Data Pipeline provides reliable service for data transfer between different AWS compute and storage services (e.g., Amazon S3, Amazon RDS, Amazon DynamoDB, Amazon EMR). In other words this service is simply a data-driven workload management system, which provides a simple management API for managing and monitoring of data-driven workloads in cloud applications.[23]
- Amazon Kinesis streams data in real time with the ability to process thousands of data streams on a per-second basis. The service, designed for real-time apps, allows developers to pull any amount of data, from any number of sources, scaling up or down as needed.[5]

2.3. Cloudera

- **Analytical DBMS:** HBase, and although not a DBMS, Cloudera Impala supports SQL querying on top of Hadoop.
- **In-memory DBMS:** Although not a DBMS, Apache Spark supports in-memory analysis on top of Hadoop.
- **Hadoop distributions:** CDH open-source distribution, Cloudera Standard, Cloudera Enterprise.
- **Stream-processing technology:** Open-source stream-processing options on Hadoop include Storm.
- **Hardware/software systems:** Partner appliances, preconfigured hardware, or both available from Cisco, Dell, HP, IBM, NetApp, and Oracle.

Cloudera Inc. is an American-based software company that provides Apache Hadoop-based software, support and services, and training to business customers.[6]

Cloudera's open-source Apache Hadoop

distribution, CDH (Cloudera Distribution Including Apache Hadoop), targets enterprise-class deployments of that technology. Cloudera says that more than 50% of its engineering output is donated upstream to the various Apache-licensed open source projects (Apache Hive, Apache Avro, Apache HBase, and so on) that combine to form the Hadoop platform. Cloudera is also a sponsor of the Apache Software Foundation [7]

2.4. HP HAVEn

- **Analytical DBMS:** HP Vertica Analytics Platform Version 7 (Crane release).
- **In-memory DBMS:** Vertica is not an in-memory database, but with high RAM-to-disk ratios the company says it can ensure near-real-time query performance.
- **Hadoop distribution:** None.
- **Stream-processing technology:** None.
- **Hardware/software systems:** HP ConvergedSystem 300 for Vertica, plus a choice of reference architectures for Cloudera, Hortonworks, and MapR Hadoop distributions.

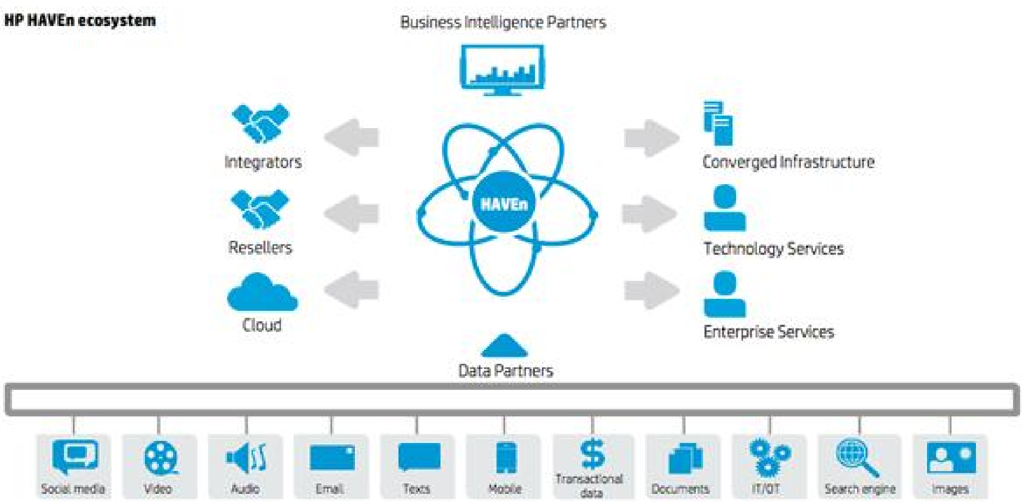


Fig. 3. HAVEn Ecosystem

HP calls its big-data-platform architecture HAVEn (**Fig. 3. -** HAVEn Ecosystem), an acronym for Hadoop, Autonomy, Vertica, Enterprise Security, and "n" applications.

The cluster-based, column-oriented Vertica Analytics Platform is designed to manage large, fast-growing volumes of data and provide very fast query performance when used for data warehouses and other query-intensive applications. The product claims to drastically improve query performance over traditional relational database systems, provide high-availability, and petabyte scalability on commodity enterprise servers.

Its design features include:

- Column-oriented storage organization, which increases performance of sequential record access at the expense of common transactional operations such as single record retrieval, updates, and deletes.[9]
- Standard SQL interface with many analytics capabilities built-in, such as time series gap filing/interpolation, event-based windowing and sessionization, pattern matching, event series joins, statistical computation (e.g., regression analysis), and geospatial analysis.
- Out-of-place updates and hybrid storage organization, which increase the performance of queries,

insertions, and loads, but at the expense of updates and deletes.

- Compression, which reduces storage costs and I/O bandwidth. High compression is possible because columns of homogeneous datatype are stored together and because updates to the main store are batched.[10]
- Shared nothing architecture, which reduces system contention for shared resources and allows gradual degradation of performance in the face of hardware failure.
- Easy to use and maintain through automated data replication, server recovery, query optimization, and storage optimization.
- Support for standard programming interfaces ODBC, JDBC, and ADO.NET.
- High performance and parallel data transfer to statistical tools such as Distributed R, and the ability to store machine learning models, and use them for in-database scoring.[11][12]

2.5. Hortonworks

- **Analytical DBMS:** HBase; although not a DBMS, Hive is Hortonworks' option for SQL querying on top of Hadoop.
- **In-memory DBMS:** Although not a DBMS, Apache Spark supports in-memory analysis on top of Hadoop.
- **Hadoop distributions:** Hortonworks Data Platform (HDP) 2.0, HDP for Windows, Hortonworks Sandbox (free, single-node desktop software offering Hadoop tutorials).
- **Stream-processing technology:** Open-source stream-processing options on Hadoop include Storm.
- **Hardware/software systems:** Partner appliances, preconfigured hardware, or both available from HP, Teradata and others.

On the matter of customer acquisition, six-year-old Cloudera probably has a slight lead over three-year-old Hortonworks (**Fig. 4. -** Hortonworks Data platform), but only just. Analysts estimate Cloudera's base of paying subscribers at around 350, while Hortonworks' CEO Rob Bearden says his company has acquired 250 customers over the past five quarters.

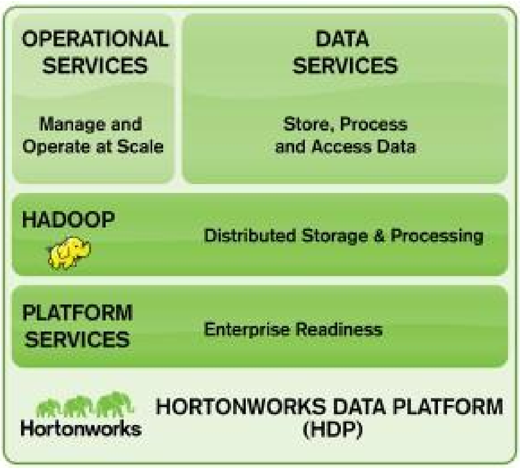

OPERATIONAL SERVICES
Manage and Operate at Scale
DATA SERVICES
Store, Process and Access Data
HADOOP
Distributed Storage & Processing
PLATFORM SERVICES
Enterprise Readiness
Hortonworks
HORTONWORKS DATA PLATFORM (HDP)

Fig. 4. Hortonworks Data platform

The most significant point of disagreement between Cloudera and Hortonworks lies in their answers to a single question – and the one that, arguably, matters most to enterprise customers: should Hadoop complement or replace traditional enterprise data warehouse (EDW) investments?

2.6. IBM

- **Analytical DBMS:** DB2, Netezza (**Fig. 5.** - IBM Netezza platform).
- **In-memory DBMS:** DB2 with BLU Acceleration, solidDB.
- **Hadoop distribution:** InfoSphere BigInsights.
- **Stream-processing technology:** InfoSphere Streams.
- **Hardware/software systems:** PureData System For Operational Analytics (DB2), IBM PureData System for Analytics (Netezza); PureData System for Hadoop (BigInsights).

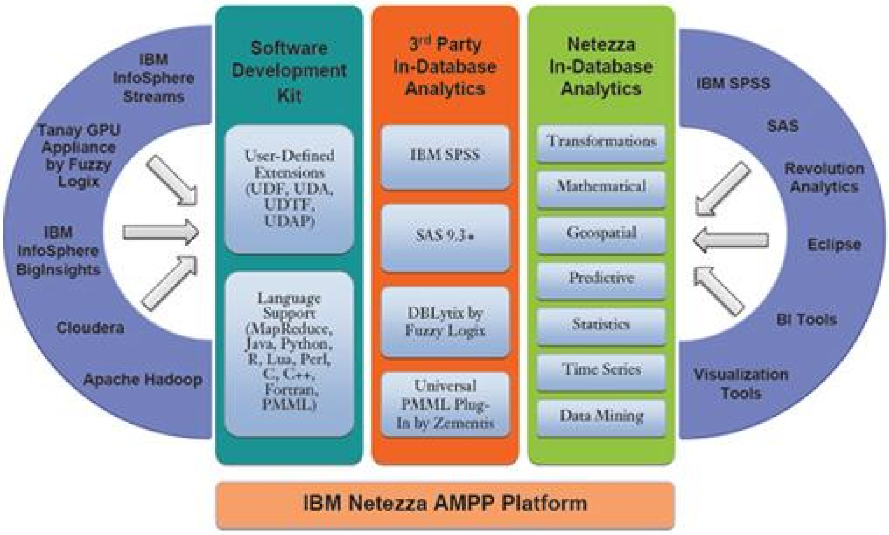

IBM InfoSphere Streams
Tanay GPU Appliance by Fuzzy Logix
IBM InfoSphere BigInsights
Cloudera
Apache Hadoop
Software Development Kit
User-Defined Extensions (UDF, UDA, UDTF, UDAP)
Language Support (MapReduce, Java, Python, R, Lua, Perl, C, C++, Fortran, PMML)
3rd Party In-Database Analytics
IBM SPSS
SAS 9.3+
DBLytix by Fuzzy Logix
Universal PMML Plug-In by Zementis
Netezza In-Database Analytics
Transformations
Mathematical
Geospatial
Predictive
Statistics
Time Series
Data Mining
IBM SPSS
SAS
Revolution Analytics
Eclipse
BI Tools
Visualization Tools
IBM Netezza AMPP Platform

Fig. 5. IBM Netezza platform

Although IBM has plenty of products and services, it's not a product-oriented provider of technology. IBM leads with its deep integration and consulting expertise in a consultative approach focused on building business-differentiating "solutions" that might incorporate multiple products.
IBM Netezza Analytics' advanced technology fuses data warehousing and in-database analytics into a scalable, high-performance, massively parallel advanced analytic platform that is designed to crunch through petascale data volumes. This allows users to ask questions of the data that could not have been contemplated on other architectures. IBM Netezza Analytics is designed to quickly and effectively provide better and faster answers to the most sophisticated business questions. [13]

2.7. Microsoft

- **Analytical DBMS:** SQL Server 2012 Parallel Data Warehouse (PDW).
- **In-memory DBMS:** SQL Server 2014 In-Memory OLTP (option available with SQL Server 2014, set for release by second quarter of 2014).
- **Stream-processing technology:** Microsoft StreamInsight.
- **Hadoop distribution:** HDInsight/Windows Azure HDInsight Service (based on Hortonworks Data Platform).
- **Hardware/software systems:** Dell Parallel Data Warehouse Appliance, HP Enterprise Parallel Data Warehouse Appliance.

The Microsoft Analytics Platform System (**Fig.6.** - Microsoft Analytics Platform System) is a turnkey big data analytics appliance, combining Microsoft's massively parallel processing (MPP) data warehouse technology–the SQL Server Parallel Data Warehouse (PDW)–together with HDInsight, Microsoft's 100% Apache Hadoop distribution, and delivering it as a turnkey appliance. To integrate data from SQL Server PDW with data from Hadoop, APS offers the PolyBase data querying technology.[14]

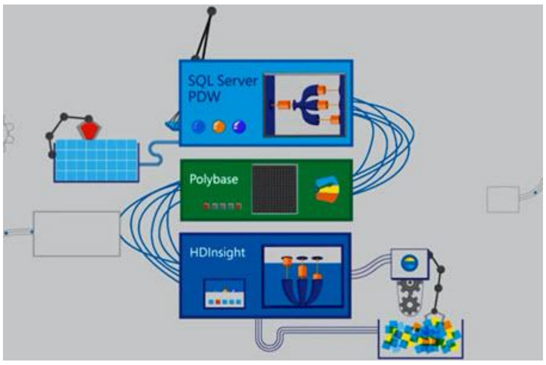
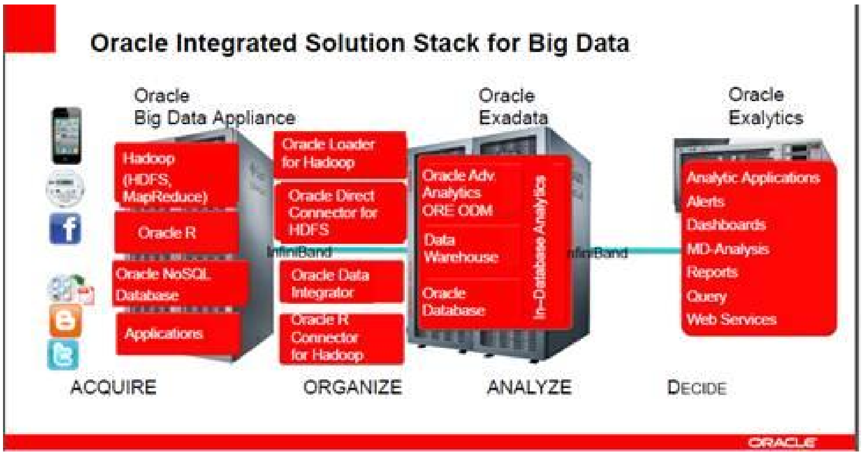


Fig.6. Microsoft Analytics Platform System

2.8. ORACLE

- **Analytical DBMSs:** Oracle Database, Oracle MySQL, Oracle Essbase.
- **In-memory DBMS:** Oracle TimesTen, Oracle Database 12c In-Memory Option (announced in 2013 without details, roadmaps, or release dates).
- **Stream-analysis option:** Oracle Event Processing.
- **Hadoop distribution:** Resells and supports Cloudera Enterprise.
- **Hardware/software systems:** Exadata, Exalytics, Oracle Big Data Appliance.



Oracle Integrated Solution Stack for Big Data

Oracle Big Data Appliance

Hadoop (HDFS, MapReduce)

Oracle R

Oracle NoSQL Database

Applications

Oracle Loader for Hadoop

Oracle Direct Connector for HDFS

InfiniBand

Oracle Data Integrator

Oracle R Connector for Hadoop

Oracle Exadata

Oracle Adv. Analytics ORE ODM

Data Warehouse

Oracle Database

In-Database Analytics

InfiniBand

Oracle Exalytics

Analytic Applications

Alerts

Dashboards

MD-Analysis

Reports

Query

Web Services

ACQUIRE

ORGANIZE

ANALYZE

DECIDE

ORACLE

Fig.7. Oracle Big Data Appliance

The **Oracle Big Data Appliance** consists of hardware and software from Oracle Corporation designed to integrate enterprise data, both structured and unstructured. It includes the Oracle Exadata Database Machine and the Oracle Exalytics Business Intelligence Machine, used for obtaining, consolidating and loading unstructured data into Oracle Database 11g. The product also includes an open source distribution of Apache Hadoop, Oracle NoSQL Database, Oracle Data Integrator with Application Adapter for Hadoop, Oracle Loader for Hadoop, an open source distribution of R, Oracle Linux, and Oracle Java Hotspot Virtual Machine [15]

Oracle Big Data Appliance (**Fig.7.** - Oracle Big Data Appliance) By combining the newest technologies from the Hadoop ecosystem and powerful Oracle SQL capabilities together on a single pre-configured platform, Oracle Big Data Appliance is uniquely able to support rapid development of new Big Data applications and tight integration with existing relational data. Oracle Big Data Appliance is pre-configured for secure environments leveraging Apache Sentry, Kerberos, both network encryption and encryption at rest as well as Oracle Audit Vault and Database Firewall.[16]

2.9. Pivotal

- **Analytical DBMS:** Pivotal Greenplum Database.
- **In-memory DBMS:** Pivotal GemFire and SQLFire. Pivotal HD used in combination with GemFire XD and HAWQ for in-memory analysis on top of Hadoop.
- **Stream-analysis option:** Pivotal is working a project aimed at integrating its GemFire (NoSQL) and SQLFire in-memory data grid capabilities with Pivotal Hadoop and Spring XD as a data-ingest mechanism to support scalable, streaming-data analysis.
- **Hadoop distribution:** Pivotal HD.

- **Hardware/software systems:** Pivotal Data Computing Appliance

Pivotal HD is 100% Apache Hadoop compliant and supports all Hadoop Distributed File System (HDFS) file formats. In addition, Pivotal HD supports Apache Hadoop-related projects, including Yarn (aka MapReduce 2.0), Zookeeper and Oozie (for resource and workflow management), Hive and HBase (for language and analytics support).[17]

Pivotal GemFire® stores all operational data compressed and in-memory to avoid disk I/O time lags. Nodes operate in a cluster, optimizing data distribution and processing, to ensure the highest speed and balanced utilization of system resources. Pivotal GemFire scales elastically and linearly – adding nodes increases capacity predictably.[18]

2.10. SAP

- **Analytical DBMSs:** SAP Hana, SAP IQ.
- **In-memory DBMS:** SAP Hana. **Stream-analysis option:** SAP Event Stream Processing.
- **Hadoop distribution:** Resells and supports Hortonworks, Intel; Hadoop integrations certified by Cloudera and MapR.
- **Hardware/software systems:** Multiple hardware configuration partners include Dell, Cisco, Fujitsu, Hitachi, HP, and IBM.

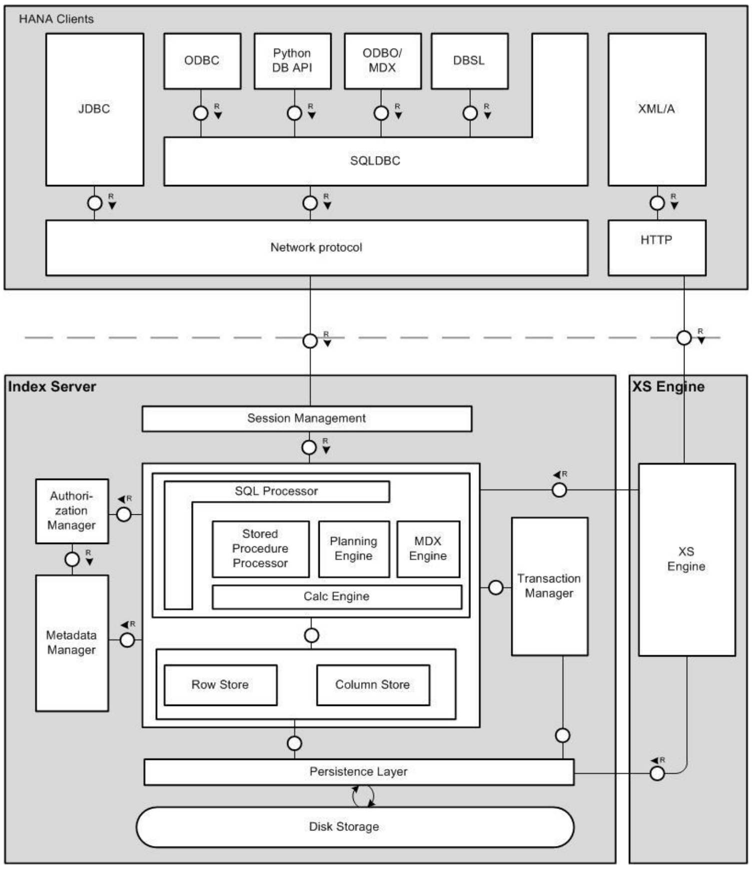


Fig.8 Architecture

SAP HANA is an in-memory, column-oriented, relational database management system developed and marketed by SAP SE. [19] HANA's architecture is designed to handle both high transaction rates and complex query processing on the same platform. SAP HANA was previously called SAP High-Performance Analytic Appliance [20]

The main process, called the index server, has a structure **Fig.8.** – Architecture.

The indexer performs session management, authorization, transaction management and command processing. Note that HANA has both a row store and a column store. Users can create tables using either store, but the column store has more capabilities. The index server also manages persistence between cached memory images of database objects, log files and permanent storage files.

The Authorization manager provides authentication and authorization services. The Authorization Manager can provide security based on SAML, OAuth or Kerberos authentication protocols.

The Extended Services (XS) Engine is a web server with privileged access to the database. Applications written with server-side JavaScript or as Java Servlets can be deployed to the XS Engine. These can either be HTML web applications or REST web service endpoints. Server-side JavaScript includes jQuery-based extensions for database access and to access HTTP request and response messages. The JavaScript engine is based on the Mozilla SpiderMonkey project. [21]

5. Conclusions

With data growing so rapidly and the rise of unstructured data accounting for 90% of the data today, the time has come for enterprises to re-evaluate their approach to data storage, management and analytics. Legacy systems will remain necessary for specific high-value, low-volume workloads, and complement the use of Hadoop -optimizing the data management structure in your organization by putting the right Big Data workloads in the right systems. The cost-effectiveness, scalability, and streamlined architectures of Hadoop will make the technology more and more attractive. In fact, the need for Hadoop is no longer a question. The only question now remaining is how to take advantage of it best. All of these tools provide a rich feature set ready for enterprise use. It will be up to the end user to do a thorough comparison and select either of these tools

References

[1] http://www.networkworld.com/article/2837779/big-data-business-intelligence/8-big-trends-in-big-data-analytics.html.

[2] http://www.datamation.com/applications/big-data-analytics-overview.html

[3] http://www.actian.com/solutions/#customer-analytics-content

[4] http://aws.amazon.com/

[5] http://en.wikipedia.org/wiki/Amazon_Web_Services

[6] http://en.wikipedia.org/wiki/Cloudera

[7] http://www.apache.org/foundation/sponsorship.html

[8] Vance, Ashlee (16 March 2009). "Bottling the Magic Behind Google and Facebook". *The New York Times*.

[9] Monash, C: "Are row-oriented RDBMS obsolete?" *DBMS2*, January 22, 2007

[10] Monash, C: "Mike Stonebraker on database compression – comments", *DBMS2*, March 24, 2007

[11] Gagliordi, Natalie. "HP adds scale to open-source R in latest big data platform". *ZDNet*.

[12] Prasad, Shreya; Fard, Arash; Gupta, Vishrut; Martinez, Jorge; LeFevre, Jeff; Xu, Vincent; Hsu, Meichun; Roy, Indrajit (2015). "Enabling predictive analytics in Vertica: Fast data transfer, distributed model creation and in-database prediction". *ACM SIGMOD International Conference on Management of Data (SIGMOD)*.

[13] http://www-01.ibm.com/software/data/puredata/analytics/nztechnology/analytics.html

[14] http://www.microsoft.com/en-us/server-cloud/products/analytics-platform-system/

[15] Darrow, Barb (2011-10-03). "Oracle BigData Appliance stakes big claim".

[16] http://www.oracle.com/technetwork/database/bigdata-appliance/overview/index.html

[17] http://pivotal.io/

[18] http://pivotal.io/big-data/pivotal-gemfire

[19] Jeff Kelly (July 12, 2013). "Primer on SAP HANA". *Wikibon*. Retrieved October 9, 2013

[20] http://en.wikipedia.org/wiki/SAP_HANA

[21] https://developer.mozilla.org/en-US/docs/Mozilla/Projects/SpiderMonkey

Mr. Ionuţ Ţăranu graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 1996, having its Master degree on "Database support for business". At present is in the process of getting his title of doctor in economy in the specialty of "Soft-computing methods for early medical diagnosis". He has been an Assistant Professor for 4 years at "Titu Maiorescu" University and also for 4 years at Academy of Economic Studies from Bucharest. He published a series of articles, from which the most important are Applying ABCD Rule of Dermatoscopy using cognitive systems and ABCDE Rule in Dermoscopy – Registration and determining the impact of parameter E for evolution in diagnosing skin cancer using soft computing alghorithms.

Mr. Taranu is currently the General Manager of Stima Soft company. He has more than 15 years of experience as a project manager and a business analyst with over 13 years of expertise in Software development, Business Process Management, Enterprise Architecture design and Outsourcing services. He is also involved in research projects, from which the most relevant are:

- Development of an Intelligent System for predicting, analyzing and monitoring performance indicators of technological and business processes in renewable energy area;
- Development of an eHealth platform for improving quality of life and the personalization of therapy at patients with diabetes;
- Development of an Educational Portal and integrated electronic system of education at the University of Medicine and Pharmacy "Carol Davila" to develop medical performance in dermatological oncology field;

Using Cloud Business Intelligence in competency assessment of IT professionals

Elena Alexandra TOADER
Bucharest University of Economic Studies
atoader22@yahoo.com

During the last years, the organizations and individuals have adopted Cloud Business Intelligence applications in order to provide access to BI related data such a dashboards, KPIs, analytics. Enterprises have increasingly implementing cloud computing models, improving their availability and reducing costs. The Cloud computing models and the Bi Cloud architecture were outlined, highlighting the advantages and disadvantages of adopting this solutions. The paper outlines the applicability of using the Oracle Business Intelligence Publisher reports in analyzing the results obtained from the competency assessment process of the IT professionals that are working in Romanian Software Organizations.

Keywords: *cloud computing, business intelligence, cloud business intelligence, competency assessment model, Bi Publisher*

1 Introduction

Nowadays, the organizations are generating large volumes of data as a result of business processes. Cloud Business Intelligence solutions are gradually gaining popularity among the companies, as many businesses are realizing the benefits of data analytics. In this changing business environment, there is a need for a more scalable and flexible information technology architecture that can show and process accurate data [1].

Cloud computing provide a competitive advantage to IT organizations by adding flexibility to the way IT resources are consumed and by enabling the users to pay only for services or resources used. Organizations are using Clouds in order to reduce the IT capital and to provide needed resources to run applications and services [2].

The evolution of cloud computing has revolutionized how the computing is abstracted and utilized on remote third party infrastructure. Cloud computing is attractive to companies and organizations as it eliminates the requirement for them to plan ahead for provisioning, and allows them to start with small resources. The most important benefit provided by Clouds are the resources offered in a pay-as-you-go manner, improving the availability and cost reduction. Clouds can help the organizations in saving money on the IT infrastructure [3].

This paper presents a practical solution implemented into a Business Intelligence product called BI Publisher and underlines the applicability of the BI solution in a competency assessment process of the IT professionals that are working in Romanian Software Organizations.

2. Cloud Computing

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [4]. Cloud computing represents a new paradigm model that allows the users to access any IT resource (hardware and software) over the Internet. This allows to the companies to increase the performance and reduce the costs [5].

2.1 Cloud Computing Deployment Models

Cloud deployment models indicate how the cloud services are made available to users. There are four deployment models associated with cloud computing [5]:

- Public cloud. This model supports all users that want to use a hardware resource (OS, memory, storage) or a software resource (application server, database) on a subscription basis. The most common public clouds are used for application development and testing, as well non-mission-critical tasks, such as file-sharing and email service
- Private cloud. This model represents a typically infrastructure used by a single organization. The infrastructure can be used by the organization to support many user groups or can be managed by a service provider. Private clouds are more expensive than public clouds due to the costs involved in acquiring and maintaining it.
- Hybrid clouds. In this model, organizations are using private clouds interconnected with public clouds. This infrastructure is used by many organizations in order to scale up his resources rapidly. Hybrid clouds are used into during the holiday's season, when the need for more computing resources is greater and can be helpful to access a private cloud to scale up the infrastructure.
- Community clouds are supported by multiple organizations that are sharing the computing resources. This infrastructure is used by the universities that are cooperating in different areas of research, police departments within a state or country, different organizations with the same domain of activity. The access into a community cloud is restricted to the members of the community.

2.2 Cloud Computing Service Models

Depending on the customer's needs, the cloud computing can provide different service models. Based on this models, it can be implemented different types of could solutions that are efficient and reliable. There are three types of cloud computing models [6]:

- Cloud Software as a Service (SaaS) – in this model, the consumer is using the provider's applications running on a cloud infrastructure. The applications are available through a web-based interface. The consumer has limited application configuration settings and does not has the possibility to administer or to control the cloud infrastructure (networks, servers, operating systems)
- Cloud Platform as a Service (PaaS) – in this model the consumer can deploy applications into the cloud infrastructure. The applications can be developed in different programming languages and tools. The consumer cannot manage or control the cloud infrastructure (network, servers, operating systems), but can manage the deployment of the applications and their configurations.
- Cloud Infrastructure as a Service (IaaS) – the consumer can manage the storage, network, operating systems, can deploy different software applications. Even though the consumer cannot manage the cloud infrastructure, he has limited control for selecting network components (such as host firewall) and for managing the operating system, storage and deployed applications.

2.3 The advantages and disadvantages of using Cloud Computing

Different studies has been conducted in order to outline the key advantages and disadvantages that cloud computing can offer to an organization [7][8][9]. In Table 1 are described the main advantages and disadvantages that cloud computing can offers to an enterprise.

Table 1. The main advantages and disadvantages of Cloud Computing

Advantages	**Disadvantages**
Lower costs of entry for smaller firms	Security at all levels (network, host, application and data level) and the privacy of users' information.
Lower IT barriers to innovation and can be helpful in many start-ups	Connectivity and open access given by the high speed of access for all users.
More easier for enterprises to scale up or down their services dynamically through software APIs and with minimal service provider interaction	Reliability of the applications that must be available to support every hour some failures or outages.
New classes of applications and delivers services can be used by an important number of users	Interoperability and portability between private clouds and public clouds. This leads to highly integrated connections between instances in order to produce reliable information
Rapid access to hardware resources without capital investments for users	Changes in IT organization: trainings must be provided and the changing of IT role within the organization

3. Business Intelligence integration into Cloud Computing

3.1 Business Intelligence definition

Business Intelligence (BI) represents a set of methodologies, processes, architectures and technologies that transform a raw data into meaningful and useful information used to enable more effective strategic, tactical and decision-making. [10]

BI is employed for monitoring the performance of the business processes through the analysis of multidimensional data taken from distributed transaction processing systems across the enterprise [11]. BI offers information analysis and information discovery technologies as Data Warehouse, On-line Analytical Processing (OLAP), and Data Mining.

BI is integrated with different system types as enterprise resource planning (ERP), customer relationship management (CRM), supply chain, marketing and other databases. BI involves intelligent reporting on top of existing data exported from the systems described above, which helps in business decision making [12].

A BI system has four components: a data warehouse (data source), business analytics (collection of tools for data mining and analysing the data from the warehouse), business performance management (for monitoring the performance) and a user interface (a web-browser interface) [13]. Operational BI is providing more functions in the organization with role-specific dashboards and scorecard and is related to the performance management and the business process management. The BI data resulted must be consistent and reliable [14].

BI involves intelligent reporting on top of existing data which helps in decision making process. BI has evolved over time, but the main key components still exists: the factual data needs to be aggregate from various data sources in order to involve the required transformation.

BI systems help the organization in obtaining useful, correct and in-time information taken from different data sources. The BI systems close the gap between the report analyses and the huge amount of data available for the decision factor. In this way, BI systems support the decision making process [15].

There are a few factors considered necessary when an organization has adopted a BI system: the BI solution should be business-oriented, rather than technology-oriented, act towards reaching the goals of the organization; a truthful partnership between management and informatics within the organization should be realized and the entire organization should be evaluated as a whole [15].

3.2 Cloud Business Intelligence

Cloud BI is a concept of delivering business intelligence capabilities “as a service” using clouded computing architecture [10]. It represents ways for reporting and analysis solutions in order to be developed installed and consumed more easily due to its lower cost and easier deployment. Cloud BI represents a new way to do Business Intelligence: the BI software is running in the Cloud instead of implementing complex and expensive software on-site [16].

A Cloud BI platform uses an infrastructure-as-a-service (IaaS), complements and extends a platform-as-a-service (PaaS), utilizes and on-demand, virtualized, software and hardware environment and delivers functionalities as software-as-a-service (SaaS) [16]. The BI system should be easily deployed and migrated to the cloud by providing Web-based flexibility that is specific to the new platform architecture.

The basic BI Architecture is containing three layers: hardware layer (the cloud infrastructure-as-a-service – IasS), the software layer (the cloud platform-as-a-service – Paas), the data layer (cloud software-as-a-service – Saas), the web client [10].

The software and hardware layers are containing the elements given by the cloud computer provider. Hardware layer is referring to storage, processing and networks. Software layer is referring to the operating systems and drivers that handle the hardware.

The data layer refers to the tools necessary to perform all the data processes. The data layer is composed by: data integration, database, data warehousing tools, BI tools. It is important to have a relational or a multidimensional database that administer all the data within the organization. The data warehousing tools are a set of applications that allow creation and maintenance of the data warehouse.

The top of the application architecture is the web client, since all the elements will be accessed over the Internet. There is no need for the clients to have installed any application because all the content and configuration is reached through the internet browsers.

The BI Cloud architecture is described in Figure 1.

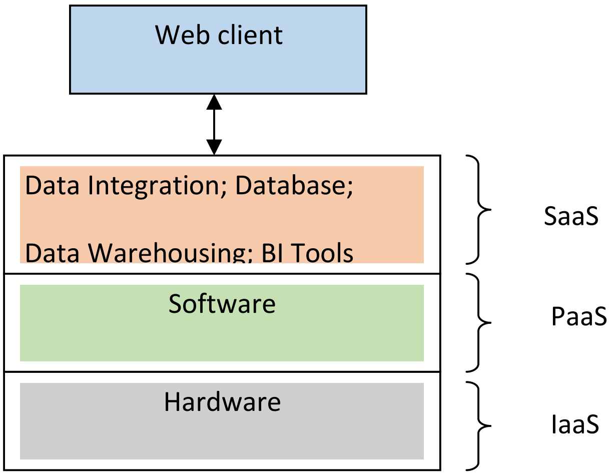


Fig 1. Bi Cloud Architecture (Source [10])

Some of benefits of Cloud Computing for BI were outlined by [5]:

- Lower costs. Under the cloud, the organization does not need to invest money in acquiring hardware, software, licenses.
- Multiple redundant sites. Most of cloud computing providers have sites geographically dispersed, so the possibility to have multiple sites can be redundant.
- Scalable provision of resources. With the cloud computing, the resources are scaled out and scaled in. This depends on the workload during the day.
- On-demand performance improvements. Under a cloud computing, the expanding of the data is made transparently for the users.

- Usage billing. The organizations are paying for a service monthly or yearly, depending on its business needs.
- Fast deployment. The platforms are up and running in few minutes for installing and deploying new features.
- Easy maintenance. The cloud computing provider makes the maintenance for hardware and software.

Since the traditional technologies were not fast enough to accomplish the business needs, and the situations in which the data warehouses and the application servers are reached their limits, the Cloud BI technological approach was needed to be achieved. Cloud computing is transforming the economics of BI and opens up the opportunity for enterprises to compete using the insight that BI provides.

3.3 BI Publisher and Enterprise Manager Cloud Control 12c

Oracle Business Intelligence (BI) Publisher is an Oracle primary reporting tool for authoring, managing, and delivering reports and documents in a faster manner than traditional reporting tools. With BI Publisher it can be viewed thousands of documents per hour with minimal impact to transactional systems. BI Publisher ships standard with Enterprise Manager Cloud Control 12c.

Oracle Enterprise Manager is an Oracle integrated enterprise IT management product line, which provides a complete, integrated and business-driven enterprise cloud management solution. In addition, it can be used to patch, monitor, and scale up/down resources in the cloud [17].

Oracle BI Publisher (formerly known as XML Publisher) is based on a very versatile open source language: XML. It can access relational, OLAP, and other data sources and enables the creation, management, and delivery of all kinds of operational reports, financial reports, and any other customer-facing documents.

The architecture of BI Publisher tool is described in Figure 2.

The Oracle BI Publisher Repository can be: Data Warehouses, Exadata, and OLAP, different applications: EBusiness Suite, SAP, XML file. The layout templates can be developed in Word, XSL, Excel, Flash, and Acrobat. The layout tools are: Word, JDeveloper, Excel, Flex and Acrobat. The report output format consisting of high fidelity and highly formatted documents is delivered in a wide diversity of formats such as: pdf, html, excel, power point. The results can be viewed online, saved for further processing, can be e-mailed, can be sent over FTP or scheduled for a delivery by, and for, a wide range of users and destinations.

The benefits of using BI Publisher tool are: optimize data extraction and document generation process, better report maintenance, easy to understand and debug the reports, advantages in using large reports.

The cloud feature of BI Publisher is enabling the users to create reports and dashboard layouts over the Internet, to deliver the right information to the right people at the right time, to make in a useful manner reports by dragging business indicators into blank sheets [19].

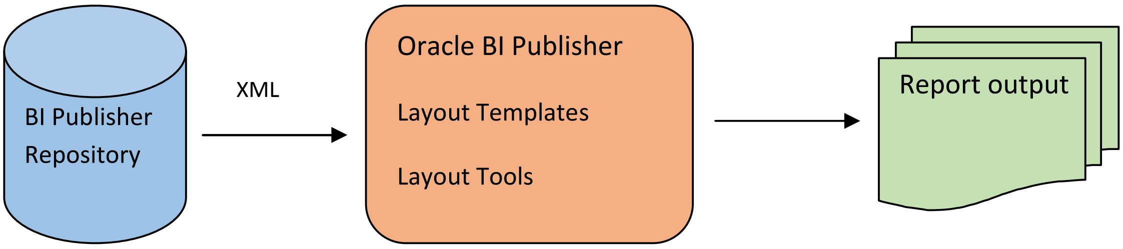


Fig 2. BI Publisher architecture (Adapted after [18])

4. Methodology

4.1 Research Methodology

Using a BI Cloud tool called BI Publisher integrated into Enterprise Management Cloud Control 12c, there can be outlined the competencies assessment results of the IT professionals that are working within a Romanian software organization.

According to [20], the best approach for assessing the competencies of IT professionals is an online assessment tool which tries to value each competency from the competency model developed by [21]. The online assessment tool contains 15 questions, each of them being linked to a competency. Each question is considered to be an assessment item. The IT professional must answer to 15 questions related to each competency from the competency model. The response on each question is having a score, and a final assessment score is computed at the end of the assessment process.

Six IT professionals that are working within a Romanian software organization have assess their competencies using the competency assessment tool. The results were exported into an XML data model that has been imported into the BI Publisher tool. The idea to the current research was to develop some reports using the BI Publisher tool and the XML data model exported from the online assessment tool. The appraisal results highlights in an easy and attractive manner the performance level of the IT professionals.

4.2 Results and Discussion

The database schema of the online assessment tool is described in Figure 3.

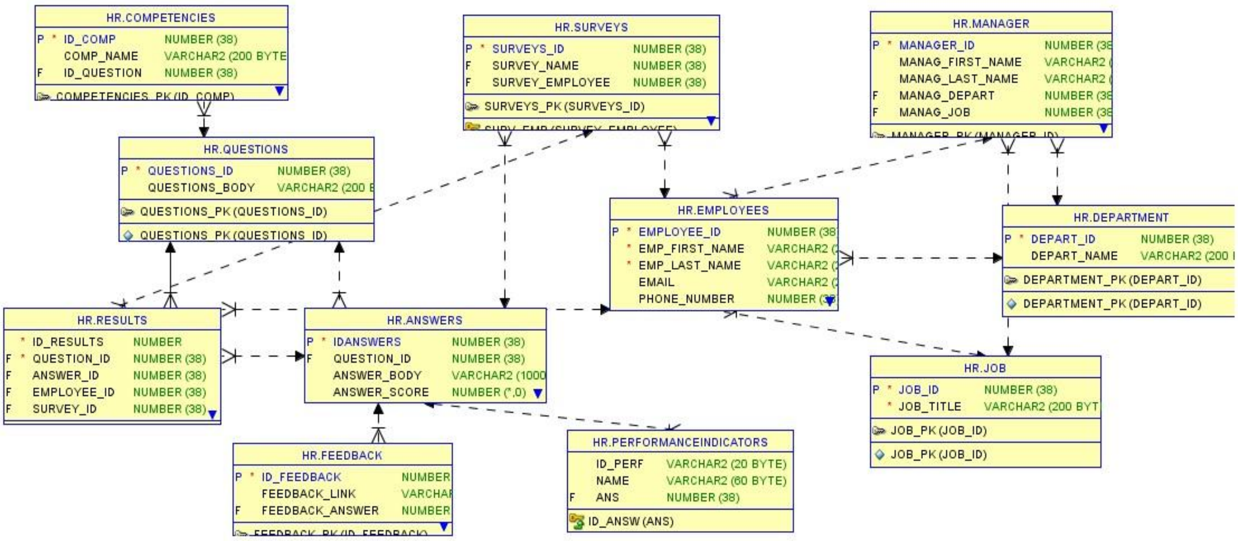

HR.COMPETENCIES
P * ID_COMP NUMBER (38)
COMP_NAME VARCHAR2 (200 BYTE
F ID_QUESTION NUMBER (38)
COMPETENCIES_PK (ID_COMP)
HR.SURVEYS
P * SURVEYS_ID NUMBER (38)
F SURVEY_NAME NUMBER (38)
F SURVEY_EMPLOYEE NUMBER (38)
SURVEYS_PK (SURVEYS_ID)
HR.MANAGER
P * MANAGER_ID NUMBER (38
MANAG_FIRST_NAME VARCHAR2
MANAG_LAST_NAME VARCHAR2
F MANAG_DEPART NUMBER (38
F MANAG_JOB NUMBER (38
HR.QUESTIONS
P * QUESTIONS_ID NUMBER (38)
QUESTIONS_BODY VARCHAR2 (200
QUESTIONS_PK (QUESTIONS_ID)
QUESTIONS_PK (QUESTIONS_ID)
HR.EMPLOYEES
P * EMPLOYEE_ID NUMBER (38
* EMP_FIRST_NAME VARCHAR2
* EMP_LAST_NAME VARCHAR2
EMAIL VARCHAR2
PHONE_NUMBER NUMBER
HR.DEPARTMENT
P * DEPART_ID NUMBER (38)
DEPART_NAME VARCHAR2 (200
DEPARTMENT_PK (DEPART_ID)
DEPARTMENT_PK (DEPART_ID)
HR.RESULTS
* ID_RESULTS NUMBER
F * QUESTION_ID NUMBER (38)
F ANSWER_ID NUMBER (38)
F EMPLOYEE_ID NUMBER (38)
F SURVEY_ID NUMBER (38)
HR.ANSWERS
P * IDANSWERS NUMBER (38)
F QUESTION_ID NUMBER (38)
ANSWER_BODY VARCHAR2 (1000
ANSWER_SCORE NUMBER (*,0)
HR.JOB
P * JOB_ID NUMBER (38)
* JOB_TITLE VARCHAR2 (200 BYT
JOB_PK (JOB_ID)
JOB_PK (JOB_ID)
HR.FEEDBACK
P * ID_FEEDBACK NUMBER
FEEDBACK_LINK VARCHA
F FEEDBACK_ANSWER NUMBER
HR.PERFORMANCEINDICATORS
ID_PERF VARCHAR2 (20 BYTE)
NAME VARCHAR2 (60 BYTE)
F ANS NUMBER (38)
ID_ANSW (ANS)

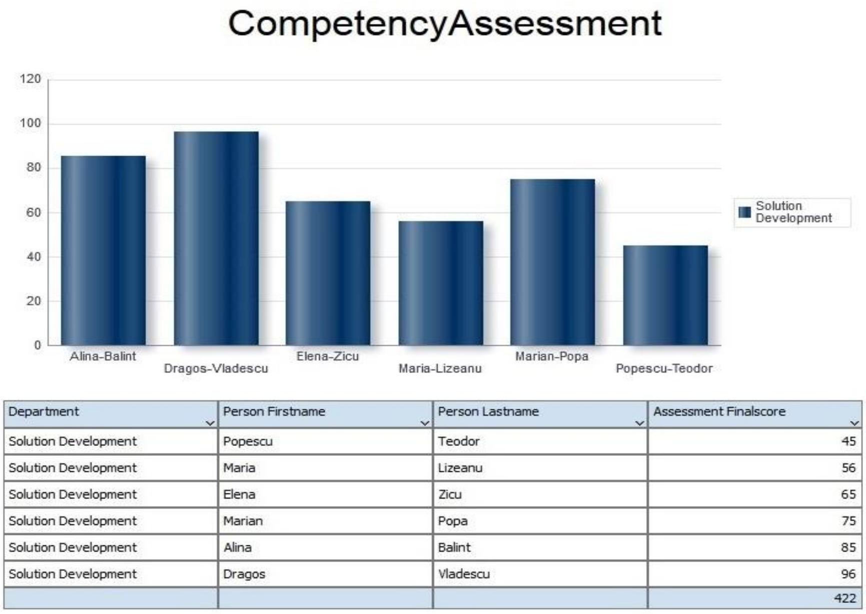
Fig 3. Database schema of the competency assessment tool

The competencies table contains all the 15 PM competencies from the competency model defined by [21]. The Questions table contains the text of the questions, each of them being linked to a competency from the Competencies table. Each question is considered to be an assessment item and has 4 possible responses that are stored in the Answers table. Each answer contains a performance indicator taken from the Performance Indicator table and a feedback containing a web resource (taken from the Feedback table). The Results table stores all the answers given by the employees to all assessment items as well the final assessment score and the Employees table stores information related to the employees: first name, second name, hire date, email, phone, manager, department and position. The Manager Table stores information related to the employee hierarchical manager, the Job table stores information related to the position and the Department table stores information related to the departments within the organization.

The xml data model that is exported from the database is containing information related to the employee: employee id, the first name, last name, the department, manager and the job position, as well information related to the score of each competency item assessed, as well the final assessment score computed. The xml associated with an employee assessment score is described in source code below:

```xml
<?xml version="1.0" encoding="UTF-8"?>
<DATA><CompetencyElements>
<Person_ID>5633</Person_ID>
<Person_FirstName>Elena</Person_FirstName>
<Person_LastName>Zicu</Person_LastName>
<Department>Solution
Development</Department>
<Manager>Danila Cristian</Manager>
<Job>Developer</Job>
<Competency_Assessment>
<Q1_Assessment_Score>0.3</Q1_Assessment_Score>
<Q2_Assessment_Score>0.2</Q2_Assessment_Score>
<Q3_Assessment_Score>0.3</Q3_Assessment_Score>
<Q4_Assessment_Score>0.2</Q4_Assessment_Score>
<Q5_Assessment_Score>0.1</Q5_Assessment_Score>
<Q6_Assessment_Score>0.3</Q6_Assessment_Score>
<Q7_Assessment_Score>0.1</Q7_Assessment_Score>
<Q8_Assessment_Score>0.2</Q8_Assessment_Score>
<Q9_Assessment_Score>0.4</Q9_Assessment_Score>
<Q10_Assessment_Score>0.1</Q10_Assessment_Score>
<Q11_Assessment_Score>0.3</Q11_Assessment_Score>
<Q12_Assessment_Score>0.2</Q12_Assessment_Score>
<Q13_Assessment_Score>0.2</Q13_Assessment_Score>
<Q14_Assessment_Score>0.4</Q14_Assessment_Score>
<Q15_Assessment_Score>0.3</Q15_Assessment_Score>
<Assessment_finalScore>65</Assessment_finalScore>
</Competency_Assessment>
</CompetencyElements>
</DATA>
```

In Figure 4 it can be observed the assessment final score showed in a bar chart and a table containing the information related to the first name, the last name, the department and the assessment final for each employee.



Department	Person Firstname	Person Lastname	Assessment Finalscore
Solution Development	Popescu	Teodor	45
Solution Development	Maria	Lizeanu	56
Solution Development	Elena	Zicu	65
Solution Development	Marian	Popa	75
Solution Development	Alina	Balint	85
Solution Development	Dragos	Vladescu	96
			422

Fig 4. Final scores associated with each IT professional

As it can be seen, Dragos Vladescu has the highest finale assessment score (96%), followed by Alina Balint (85%). The lowest scores have been achieved by Popescu Teodor (45%) and Maria Lizeanu (56). The table showing the final scores is very helpful in data analysing.

The results for the assessment of question number three, related to the assessment of the competency item: the automation and optimization of work processes are showed in Figure 5.

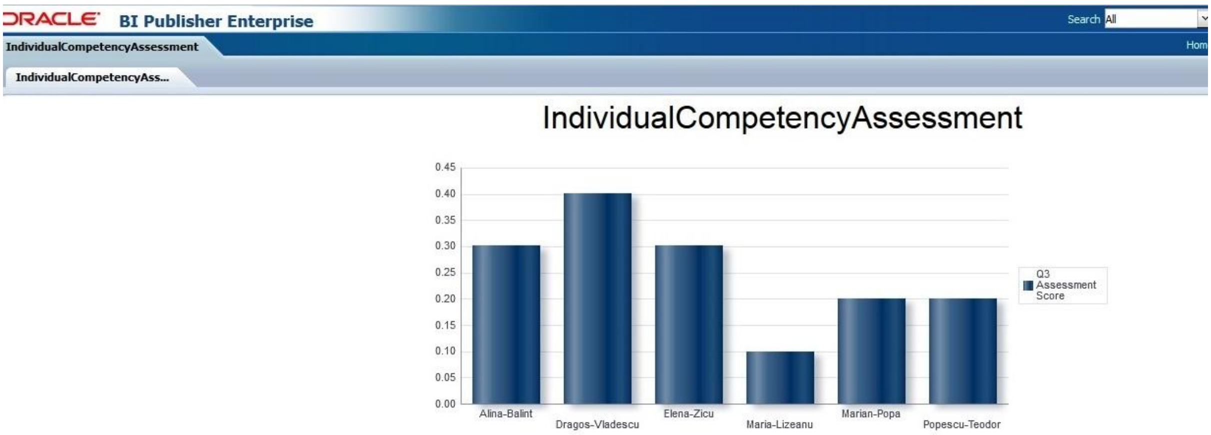


Fig 5. Assessment results for the question Q3

As it can be observed, the highest assessment score related to the evaluation of the question number three is achieved by Dragos Vladescu, followed by Alina Balint and Elena Zicu. The lowest assessment score is obtained by Maria Lizeanu which is having one of the lowest assessment final scores as we can see from the Figure 11.

The two reports presented in Figure 10 and Figure 11 are useful for the managers of the department in order to see which employee is performant and how the employees can improve their actual degree of the competency. As well there are useful for the employees which can compare their assessment results with the colleagues' results in order to raise the competitiveness between the employees from the same department.

The results of the individual assessment for the employee: Elena Zicu is showed in Figure 6.

As it can be seen, the highest score were achieved on the assessment of the competencies efficiency (question number 9) and health, security, safety and environment (question number 14). The lower assessment scores are related to the competency implementation of the maintenance technique (question 5), motivation (question 7) and creativity (question 10). This report is useful as well for the manager that can analyse the individual assessment score in order to improve the current performance level and as well it is useful for the employee who made the assessment in order to be aware of the individual competency level that possesses.

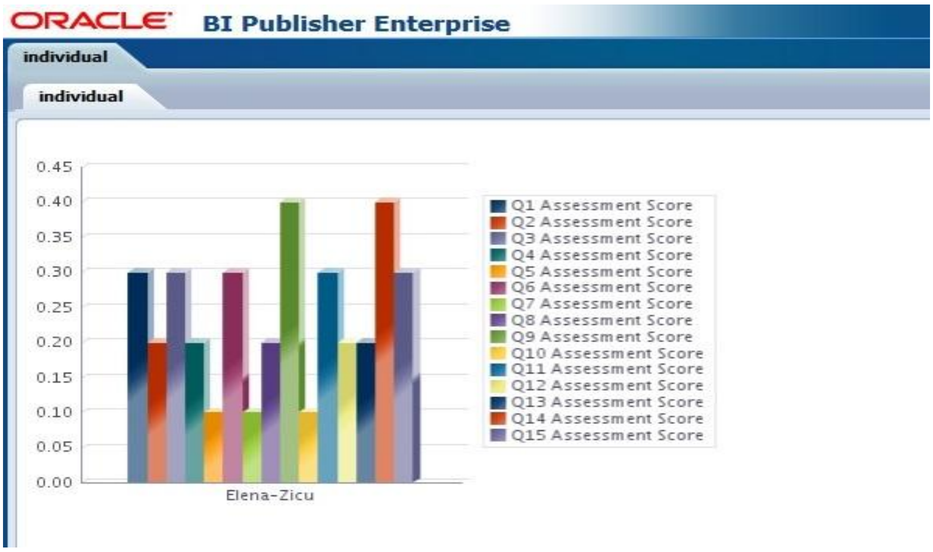


Fig 6. Individual results of the competency assessment

5. Conclusions

The aim of the study was to underline the applicability of the BI Cloud solution BI Publisher applicability of using the BI Publisher reports in analysing the results obtained from the competency assessment process of the IT professionals that are working in Romanian Software Organizations. First, Cloud computing concepts were presented revealing the cloud computing deployment models, the cloud computing service models as well a short analysis presenting the advantages and disadvantages of using Cloud Computing. The traditional Business Intelligence solutions has been presented and their integrations into BI Cloud solutions have been highlighted. At the end, the BI Publisher tool integrated into Enterprise Manager Cloud Control 12c has been showed, as well its architecture and the advantages of developing reports. There has been developed some reports that point out the assessment results of the six IT professionals. The reports are useful for the manager and as well for the professionals in order to improve the performance level achieved.

Acknowledgment

„This paper was co-financed from the European Social Fund, through the Sectorial Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/138907 "Excellence in scientific interdisciplinary research, doctoral and postdoctoral, in the economic, social and medical fields -EXCELIS", coordinator The Bucharest University of Economic Studies”.

References

[1] Bharat Chandra and Meena Iyer, BI in the Cloud – Defining the Architecture for Quick Wins. Available: http://www.infosys.com/infosys-labs/publications/Documents/BI-in-a-cloud.pdf

[2] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, M. Zaharia, Above the Clouds: A Berkeley View of Cloud Computing, Technical report UCB/EECS-2009-28, Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, USA (February 2009).

[3] Farrukh Shahzad, State-of-the-art Survey on Cloud Computing Security Challenges,
Approaches and Solutions, Procedia Computer Science 37 (2014) 357 – 362

[4] National Institute of Standards and Technology. NIST Definition of Cloud Computing v15. October, 2009

[5] Eumir P. Reyes (2010), A Systems Thinking Approach to Business

Intelligence Solutions Based on Cloud Computing, Massachusetts Institute of Technology

[6] Mell Peter, Grance Tim, The NIST Definition of Cloud Computing, National Institute of Standards and Technology, Information Technology Laboratory, available at:

[7] Marston S, Li Z, Bandyopadhyay S, Zhang J, Ghalsasi A. Cloud computing — The business perspective, Elseviewer, 2010

[8] Vaquero LM, Rodero-Merino L, Caceres J, Lindner M. A break in the clouds: Towards a cloud definition, SIGCOMM Computer Communications Review, 39:50 55, 2009.

[9] Avram M. G. (2013) Advantages and challenges of adopting cloud computing from an enterprise perspective, The 7th International Conference Interdisciplinarity in Engineering (INTER-ENG 2013), Procedia Technology, 12 (2014) 529 – 534

[10] Yuvraj Singh Gurjar, Vijay Singh Rathore, Cloud Business Intelligence – Is What Business Need Toda, International Journal of Recent Technology and Engineering (IJRTE) Vol 1(6), January 2013

[11] Marenakos L. (2005), Leveraging business intelligence, ABI/INFORM Glob. (2005), pp. 8–11

[12] W. H. Inmon, "Building the Data Warehouse," John Wiley Sons, Inc., New York (NY, USA), 2005.

[13] A. Butuza, I. Hauer, C. Muntean, A. Popa, ”Increasing the Business Performance using Business Intelligence”, Analele Universităţii "Eftimie Murgu" Reşiţa, anul XVIII, nr.3, 2011, pp. 67-72, available

[14] Shimaa Ouf, Mona Nasr, Business Inteligence in the cloud, 2011 International Conference on Computer and Network Engineering (ICCNE 2011), vol 2(279)

[15] B. Ghilic-Micu, M. Mircea, M. Stoica, ”The Audit of Business Intelligence Solutions”, Informatica Economică vol. 14, no. 1/2010, available at http://revistaie.ase.ro/content/53/07%20Ghilic,%20Mircea,%20Stoica.pdf

[16] J. Dibbern, T. Goles, R. Hirschheim, and B. Jayatilaka, "Information Systems Outsourcing. A Survey and Analysis of the Literature," 2004.

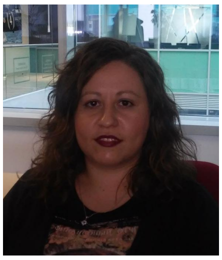
[17] Leslie Grumbach Studdard (2013), Oracle Fusion Middleware Report Designer's Guide for Oracle Business Intelligence Publisher, 11g Release 1(11.1.1), Oracle, E22254-03, Available: http://docs.oracle.com/cd/E28280_01/bi.1111/e22254.pdf

[18] Wiggins Ike (2009), Learn how this free utility can make running BI Publisher reporting components incredibly easy, Available: http://www.oracle.com/technetwork/articles/wiggins-bipublisher-085427.html

[19] Daniela Bozdoc (2011) Oracle BI Publisher 11g: A Practical Guide to Enterprise Reporting. Available: https://www.packtpub.com/sites/default/files/3180EN-Chapter-3-Multiple-Data-Sources.pdf

[20] Bodea C. N, E-A. Toader, "Project management competency assessment methods for IT professionals", 13th International Conference on Informatics in Economy (IE 2014), Education, Research and business Technologies, Bucharest, May 2014

[21] Bodea C. N., E-A. Toader, "Development of the PM competency model for IT professionals, base for HR management in software organizations", 12th International Conference on Informatics in Economy (IE 2013), Education, Research and business Technologies, Bucharest, April 2013



Elena Alexandra TOADER has graduated the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest Academy of Economic Studies in 2008. She has graduated the Professional Master's Program in Economic Informatics at the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest Academy of Economic Studies in 2010. She is a PHD Student at the Economic Informatics PHD School and the main field of interest is competency assessment methods.

Cloud Computing and its Challenges and Benefits in the Bank System

Bogdan NEDELCU, Madalina-Elena STEFANET, Ioan-Florentin TAMASESCU,
Smaranda-Elena TINTOIU, Alin VEZEANU
University of Economic Studies, Bucharest, Romania
bogdannedelcu@hotmail.com, mada.stefanet@gmail.com ,
ioan.tamasescu@gmail.com , alin_v12@yahoo.com , smarandat138@gmail.com

The purpose of this article is to highlight the current situation of Cloud Computing systems. There is a tendency for enterprises and banks to seek such databases, so the article tries to answer the question: "Is Cloud Computing safe". Answering this question requires an analysis of the security system (strengths and weaknesses), accompanied by arguments for and against this trend and suggestions for improvement that can increase the customers confidence in the future.

***Keywords**: Cloud Computing, Bank System, Security*

1 Introduction

Cloud computing has experienced a fast growth during the last years, and it is expected to keep developing more and more. Cloud services will be profitable in business application, which will transform services in cloud-based services. This change is needed especially for application like ERP (enterprise resource planning) or CRM (customer relationship management).

Banks are an important segment of business area that cloud computing is targeting in the next few years. Due to this type of business needs, cloud services must be similar with a "silver bullet". There are many advantages that cloud provides for banks as customers. First of all, cost savings, using cloud-servers instead of personal servers, will save a lot of money. Moreover, cloud provides: usage-based billing, business continuity, business agility, green IT.

Thus, nowadays cloud computing services has some disadvantages that stops banks to adopt the cloud, such as security, confidentiality of the data, and also quality of services.

2. Security strengths in cloud computing

Security is one of the biggest arguments used against the actual cloud computing system. However, cloud computing systems are often safer than mainframe systems managed at the local level, at least for small and medium companies (banks). This may list the strengths of cloud computing systems: private cloud, data centralization, multi-factor authentication, sharing security, economy of scale and others.

Private cloud is probably the most important argument in favor of using cloud computing systems by organizations (banks). An interesting comparison is between the current situation of internet banking and cloud computing. Security issues were also an inhibitor to adoption of internet banking[1] (about mid 90's), which can be considered a precursor of cloud computing. Similarly, as cloud computing providers who continue to address market concerns relating to safety, economy and convenience of cloud computing will become a commonplace like online banking and other online financial transactions today.

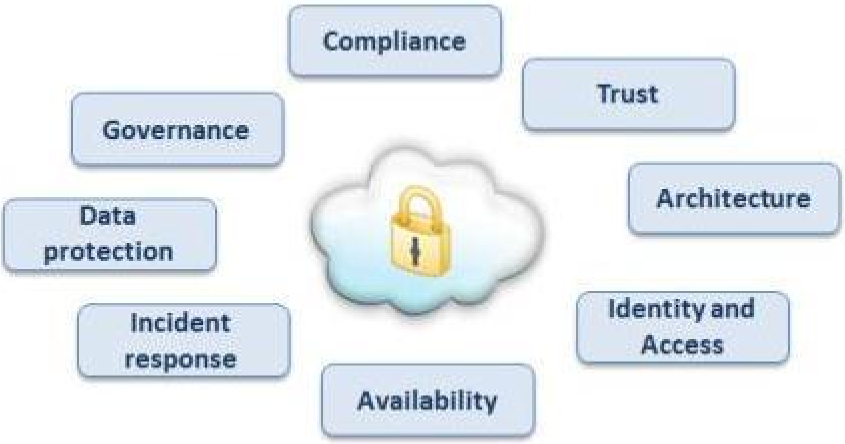


Fig.1.Cloud Security [1]

Despite the conventional and economical benefits, cloud computing may not be for everyone. For example, a security and risky perspective, public cloud computing may not appeal to organizations with missions like extreme advertisement and / or highly sensitive data. However, for most, cloud computing security advantages described above along with the ability to create private cloud (which allows customers to control who is in the cloud, where data is stored, who has access etc.) should provide the necessary security guarantees to satisfy most organizations.[2]

- *Centralization of data* falls into two categories: preventing leak of data and monitoring. Using back-up systems is inefficient in terms of time and at high risk of data loss through the physical degradation of the backup devices that visibly reduce clouds computing efficency while saving data and its potential.[2]
- *Multi-factor Authentification*: A sizable part of the cloud computing providers mainframe systems combines elements like passwords, hard token elements, biometric elements, increasing the security level. For many companies it is more profitable to resort to such a system than to implement its own cloud security system with these benefits.[1]
- *Security patching*: Cloud computing offers this concept and also offers the possibility of testing. There are organizations that do not have the resources to implement such a concept or that implementation would result in huge consumption of time, so the existence of the cloud system is a plus.[1]
- *Economy of scale*: IT services centrally managed and maintained to improve services and reduce operating costs. Cloud computing providers have the ability to invest in staff, resources and facilities that allow customers to pay only for what they use rather than invest in the resources to be managed and maintained over time. Thus, without repeating the Cloud features mentioned above, providing IT Cloud offers economies of scale, as the IT system must be scalable, fair and secure.[2]
- *Security certifications*: Many industries require IT systems and facilities to maintain a certain type of information security and/or privacy certification. For example, compliance with the Federal Information Security Management Act, or FISMA, is required for the federal government while Health Insurance Portability and Accountability Action (HIPAA) compliance is required for the

healthcare industry. These certifications can be prohibitively expensive for smaller organizations to achieve; however, many cloud vendors provide access to systems and facilities that are already certified. Even if your business does not require a certification, it may be comforting to engage with vendors who offer them as it demonstrates mature business practices as it relates to information security.[2]

- *Physical security*: Reputable cloud computing vendors often host their systems in facilities that have much stronger physical security controls with meaningful certifications that many small-to-midsize companies cannot provide on their own.[1]
- *Reduce cost of testing security*: a SaaS provider only passes on a portion of its security testing costs. By sharing the same application as a service, you don't foot the expensive security code review and/or penetration test. Even with Platform as a Service (PaaS) where developers get to write code, there are potential cost economies of scale (particularly around use of code scanning tools that sweep source code for security weaknesses).[2]
- *Pre-hardened, change control builds*: this is primarily a benefit of virtualization based on Cloud Computing. Now it is the chance to start 'secure' (by your own definition) – create your Gold Image - VM and clone away. There are ways to do this today with bare-metal OS installs but frequently these require additional 3rd party tools, which are time consuming to clone or add another agent to each endpoint.
- *Reduce exposure through patching offline*: Gold images can be kept up to date securely. Offline VMs can be conveniently patched "off" the network.[2]
- *Easier to test impact of security changes:* Spin up a copy of your production environment, implement a security change and test the impact at low cost, with minimal startup time. This is a big deal and removes a major barrier to implement security in production environments.[2]
- Convenience and continuous availability: Public clouds offer services that are available wherever the end user might be located. This approach enables easy access to information and accommodates the needs of users in different time zones and geographic locations. As a side benefit, collaboration booms since it is now easier than ever to access, view and modify shared documents and files. Moreover, service uptime is in most cases guaranteed, providing in that way continuous availability of resources. The various cloud vendors typically use multiple servers for maximum redundancy. In case of system failure, alternative instances are automatically spawned on other machines. [3]
- Resiliency and Redundancy: A cloud deployment is usually built on a robust architecture thus providing resiliency and redundancy to its users. The cloud offers automatic failover between hardware platforms out of the box, while disaster recovery services are also often included. [3]

Other advantages of Cloud computing security, mention are: reliable access, automatic data backup and encryption features that are unique to each client.[1]

Cloud computing, as we have seen, can be a wonderful business enabler. However it is for a small business owner to calculate if it's the right fit for your current environment. If the limited risks can be properly managed, though, it promises cheaper, faster and more efficient ways of working which could help your business achieve stellar performance. [4]

3. Security issues in cloud computing

The intention to adopt cloud computing has increased rapidly in many organizations. Cloud computing offers many potential benefits to small and medium enterprises such as fast deployment, pay-for-use, lower costs, scalability, rapid provisioning, rapid elasticity, ubiquitous network access, greater resiliency, and on-demand security controls. Despite these extraordinary benefits of cloud computing, studies indicate that organizations are slow in adopting it due to security issues and challenges associated with it. In other words, security is one of the major issues which reduce the cloud computing adoption. [5]

Organizations use the Cloud in a variety of different service models (SaaS, PaaS, IaaS) and deployment models (Private, public, hybrid, and community). There are a number of security issues or concerns associated with cloud computing, but these issues fall into two broad categories: security issues faced by cloud providers and security issues faced by their customers (companies or organizations who host applications or store data on the on the cloud.

For resolving these problems, the responsibility goes both ways, however: the provider must ensure that their infrastructure is secure and that their clients' data and applications are protected while the user must take measures to fortify their application and use strong passwords and authentication measures. [6]

According to Takabi et al. (2010), cloud service providers and customers are responsible for security and privacy in cloud computing environments but their level of responsibility will differ for different delivery models. Infrastructure as a Service (IaaS) serves as the foundation layer for the other delivery models, and a lack of security in this layer affects the other delivery models. In IaaS, although customers are responsible for protecting operating systems, applications, and content, the security of customer data is a significant responsibility for cloud providers. In Platform as a service (PaaS), users are responsible for protecting the applications that developers build and run on the platforms, while providers are responsible for taking care of the users' applications and workspaces from one another.

In System as a Service, cloud providers, particularly public cloud providers, have more responsibility than clients for enhancing the security of applications and achieving a successful data migration. In the SaaS model, data breaches, application vulnerabilities and availability are important issues that can lead to financial and legal liabilities.

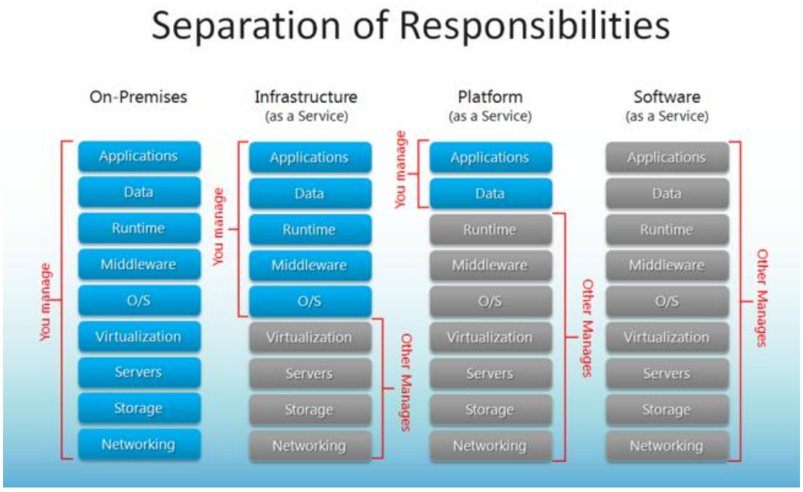


Fig 2. Separation of Responsibilities [5]

When an organization chose to store data or host applications on the public cloud, it loses its ability to have physical access to the servers hosting its information. As a result, potentially business sensitive and confidential data is at risk from insider attacks. According to a recent Cloud Security Alliance Report, insider attacks are the third biggest threat in cloud computing. Therefore, cloud service providers must ensure that use background checks are conducted for employees who have physical access to the servers in the data center. Additionally, data centers must be frequently monitored for suspicious activity.

In order to conserve resources, cut costs, and maintain efficiency, cloud service providers often store more than one customer's data on the same server. As a result there is a chance that one user's private data can be viewed by other users. To handle such sensitive situations, cloud service providers should ensure proper data isolation and logical storage segregation.

Over time, cloud providers are increasingly concerned about the security services, therefore, have implemented various ways of protection within the cloud community. [7]

Data Security - Data security is a common concern for any technology, but becomes a major problem when users SaaS providers must rely on the security itself. In SaaS, organizational data are often processed and stored in clear text in the cloud. SaaS provider is directly responsible for their safety as long as they are stored and processed. The best way is to encrypt their data security, the client, or that the client alone can be clearly read in the data. But the most commonly used method of data security backup is increasingly used lately. Also, the backup can raise large problems when providers of cloud backup services contracted to third persons or companies. But encryption itself is not a complete solution, because at some point, the data will be decrypted to be processed or to carry out the specific tasks.

A truly viable solution for data security is also more detailed filtering sites especially web content - this of course requires high costs for implementation of cloud providers.

There are complex data security challenges in the cloud:

- The need to protect confidential business, government, or regulatory data
- Cloud service models with multiple tenants sharing the same infrastructure
- Data mobility and legal issues relative to such government rules as the EU Data Privacy Directive
- Lack of standards about how cloud service providers securely recycle disk space and erase existing data
- Auditing, reporting, and compliance concerns
- Loss of visibility to key security and operational intelligence that no longer is available to feed enterprise IT security intelligence and risk management

A new type of insider who does not even work for your company, but may have control and visibility into your data

But encryption itself is not a complete solution, because at some point, the data will be decrypted to be processed or to carry out the specific tasks.

A truly viable solution for data security is also more detailed filtering sites especially web content - this of course requires high costs for implementation of cloud providers.

There are complex data security challenges in the cloud:

- The need to protect confidential business, government, or regulatory data
- Cloud service models with multiple tenants sharing the same infrastructure

- Data mobility and legal issues relative to such government rules as the EU Data Privacy Directive
- Lack of standards about how cloud service providers securely recycle disk space and erase existing data
- Auditing, reporting, and compliance concerns
- Loss of visibility to key security and operational intelligence that no longer is available to feed enterprise IT security intelligence and risk management
- A new type of insider who does not even work for your company, but may have control and visibility into your data

Specific security challenges pertain to each of the three cloud service models—Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).

- SaaS deploys the provider's applications running on a cloud infrastructure; it offers anywhere access, but also increases security risk. With this service model it's essential to implement policies for identity management and access control to applications. For example, with Salesforce.com, only certain salespeople may be authorized to access and download confidential customer sales information.
- PaaS is a shared development environment, such as Microsoft™ Windows Azure, where the consumer controls deployed applications but does not manage the underlying cloud infrastructure. This cloud service model requires strong authentication to identify users, an audit trail, and the ability to support compliance regulations and privacy mandates.
- IaaS lets the consumer provision processing, storage, networks, and other fundamental computing resources and controls operating systems, storage, and deployed applications. As with Amazon Elastic Compute Cloud (EC2), the consumer does not manage or control the underlying cloud infrastructure. Data security is typically a shared responsibility between the cloud service provider and the cloud consumer. Data encryption without the need to modify applications is a key requirement in this environment to remove the custodial risk of IaaS infrastructure personnel accessing sensitive data.[8]

Techniques for Protecting Data in the Cloud

Traditional models of data protection have often focused on network-centric and perimeter security, frequently with devices such as firewalls and intrusion detection systems. But this approach does not provide sufficient protection against APTs, privileged users, or other insidious types of security attacks. Many enterprises use database audit and protection (DAP) and Security Information and Event Management (SIEM) solutions to gather together information about what is happening. But monitoring and event correlation alone do not translate into data security. At a time when regulation and compliance issues are at an all-time high, it's dangerous to assume that monitoring, collecting, and storing logs can protect the organization from security threats, as they are reactive controls. In today's environment, both data firewalls and data security intelligence are essential to adequately protect the enterprise from new and different types of attacks. Best practices should include securing sensitive data, establishing appropriate separation of duties between IT operations and IT security, ensuring that the use of cloud data conforms to existing enterprise policies, as well as strong key management and strict acces policies.

"It is important to utilize security controls that protect sensitive data no matter where it lives, as point solutions by their very nature provide only limited visibility," says Tumulak. He emphasizes that an effective cloud security solution should incorporate three key capabilities:

- Data lockdown
- Access policies
- Security intelligence

Security applications - Applications are typically delivered through a web browser. Any application vulnerabilities can create problems in the SaaS services. These security issues are no different than any other web technology, but traditional solutions have proven ineffective. Therefore, The Open Web Application Security Project (OWASP) has identified the top 10 problems and tried to offer viable solutions.

Studies indicate that most websites are secured at the network level while there may be security loopholes at the application level which may allow information access to unauthorized users. Software and hardware resources can be used to provide security to applications. In this way, attackers will not be able to get control over these applications and change them. XSS attacks, Cookie Poisoning, Hidden field manipulation, SQL injection attacks, DoS attacks, and Google Hacking are some examples of threats to application level security which resulting from the unauthorized usage of the applications.[9]

4. Changes needed in Cloud Computing

Cloud computing security or, more simply, cloud security is an evolving sub-domain of computer security, network security , and, more broadly, information security.It refers to a broad set of policies, technologies, and controls deployed to protect data, applications, and the associated infrastructure of cloud computing. Cloud security architecture is effective only if the correct defensive implementations are in place. An efficient cloud security architecture should recognize the issues that will arise with security management. The security management addresses these issues with security controls. These controls are put in place to safeguard any weaknesses in the system and reduce the effect of an attack. While there are many types of controls behind a cloud security architecture, they can usually be found in one of the following categories:

Deterrent controls

These controls are intended to reduce attacks on a cloud system. Much like a warning sign on a fence or a property, deterrent controls typically reduce the threat level by informing potential attackers that there will be adverse consequences for them if they proceed.

Preventive controls

Preventive controls strengthen the system against incidents, generally by reducing if not actually eliminating vulnerabilities. Strong authentication of cloud users, for instance, makes it less likely that unauthorized users can access cloud systems, and more likely that cloud users are positively identified.

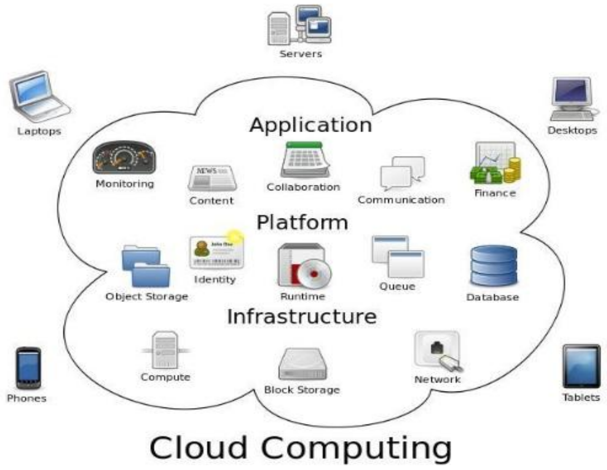


Fig. 3. Cloud Computing [9]

Detective controls

Detective controls are intended to detect and react appropriately to any incidents that occur. In the event of an attack, a detective control will signal the preventative or corrective controls to address the issue. System and network security monitoring, including intrusion detection and prevention arrangements, are typically employed to detect attacks on cloud systems and the supporting communications infrastructure.

Corrective controls

Corrective controls reduce the consequences of an incident, normally by limiting the damage. They come into effect during or after an incident. Restoring system backups in order to rebuild a compromised system is an example of a corrective control.[10]

The For Steps to a Secure Cloud Deployment

Most IT executives think cloud computing is a way to reduce capital expenditures by using virtualization technology. Many vendors tack the word “cloud” onto any Internet service. For our purposes here, we’re using the GARTNER Inc. description of how the cloud came to be so important to business: “the commoditization and standardization of technologies, in part to virtualization and the rise of service-oriented software architectures, and most importantly, to the dramatic growth in popularity of the Internet.” This is important in four specific areas:

1. *Centralized data management, using SharePoint as an example*
2. *Centralized application management, using Exchange as an example*
3. *Federated identity management, using Active Directory Federation Services (ADFS) as an example*
4. *Additional assistance for migrating to the cloud*

Centralized Data Management

In 2007, Gartner began telling security conferences it was time to abandon the hardened perimeter boundary between the enterprise and the Internet. Even at that time, experts were arguing that enterprise boundaries were already porous. Perimeters had become irrelevant to the task of keeping out intruders, so access control was required with every IT service. Security de-perimeterization is the current reality. To be truly secure, only the server that contains data can ultimately control access.

Still, it isn’t rational to manage access at every server, because many deployments contain hundreds or even thousands of servers. IT can’t really determine data rights and access rules. IT can, however, establish a role-management system with which business owners can permit or deny access relevant to business objectives.

The regulatory environment has become increasingly stringent both for data modification and data access. This requires a new paradigm: one that will allow data to migrate to whichever server is best able to service access requests, while ensuring compliance at reasonable cost. Here are some requirements to consider for data management in a cloud environment:

- Fast access to data for which the user is authorized, and when and where it is required
- Access not compromised by a natural or business catastrophe
- Data discovery by legal governmental requests, assuming the enterprise can provide the data needed
- Data Loss Prevention (DLP) is an integral part of the service offering
- A service-oriented architecture (SOA) should enable easy data migration back and forth to the cloud
- Identity of data must not include its physical location, so that the data can easily be moved
- Location tags for data should be the logical country of origin, not the data's physical location
- Data backup and recovery operations need to be based on the data identity, not its location
- Data-access rules can be created and maintained by the business owner of the data
- Access permissions can be viewed by compliance auditors
- Sensitive data can have audit controls for both modification and access
- Separation of duties prevents the same administrator from modifying data and audit logs
- Service Level Agreements (SLAs) need to spell out everyone's expectations and responsibilities

Starbucks Corp. found that the cost and delay of physical (paper-based) distribution of current pricing, business analysis and news was not cost-effective. As a result, it now supports SharePoint for its network of 16,000 locations. That SharePoint site has become a business-critical communications channel through which employees can get current information, with the ability to search quickly for the information that they need when they need it.

Availability and reliability is tracked with Microsoft System Center of Operation Manager (SCOM) and other analytic tools. Because SharePoint supports both internal and external network connections, the server locations can adapt to suit the current network topology without concern for local, cloud or mixed environments. This deployment has enabled Starbucks to realize the following benefits:

- Supporting store growth and capacity needs by improving system stability with effective monitoring and reporting tools
- Allowing store partners to work more efficiently and effectively with an intuitive portal interface and easy access to information across the enterprise
- Maintaining data security with enhanced document management and privacy functionality
- Aligning store priorities with company objectives by integrating trends and growth reports with partner communications

Integrity Protection

Any data store must be prevented from becoming an infection vector for viruses or spyware. Data types, like executables and compressed or encrypted files, can be blocked for a variety of integrity and compliance concerns. Microsoft employee David Tesar blogged about some of the business reasons to protect SharePoint using ForeFront Protection 2010 for SharePoint, which was released in May 2010.

Data Loss Protection and Detection

To ensure full protection, data from one customer must be properly segregated from that of another. It must be stored securely when “at rest” and able to move securely from one location to another (security “in motion”). IT managers must ensure that cloud providers have systems in place to prevent data leaks or access by third parties. This should be part of an SLA. Proper separation of duties should ensure that unauthorized users can’t defeat auditing and/or monitoring—even “privileged” users at the cloud provider. [11]

5. Banks and Cloud Computing

To maintain and achieve better performance in the future, banks around the world will have to adopt and master two fundamental changes:

1. The first transformation consists in changing product offerings, customer service should reflect the fact that the consumer is in control. Nowadays, the consumer is impatient and wants to be fully in control after interaction with the internet where he can do what he wants.

2. A second transformation consists in reshaping and reinventing core banking operations to enable a model Economic Competitiveness, efficient and sustainable business.

The reasons for banks to move to the cloud are many but the most important reason is the multitude of applications. Using Cloud, banking institutions pay only for the services they use .This is easier and more efficient to test new applications on the Cloud to traditional infrastructure. [12]

Models in Cloud

Cloud offers three different service models:

Platform-as-a-Service (PaaS)

In the PaaS model, cloud providers deliver a computing platform which can be accessed via web browsers, typically including an operating system, programming language execution environment, databases, and web servers. [13][14]

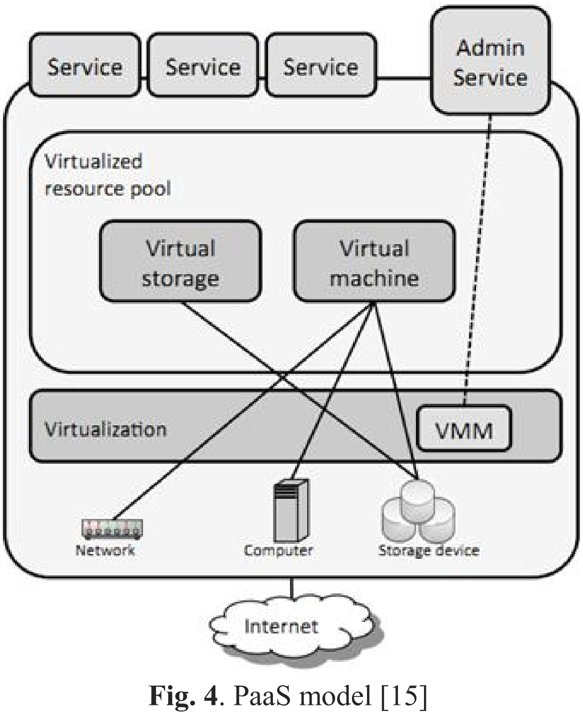


Fig. 4. PaaS model [15]

Software-as-a-Service (SaaS):

SaaS is a cloud service that provides some data and software that are accessed via a web connection. Types of software that can be delivered this way include accounting, customer relationship management, enterprise resource planning, invoicing, human resource management, content management, and service desk management.[13][14]

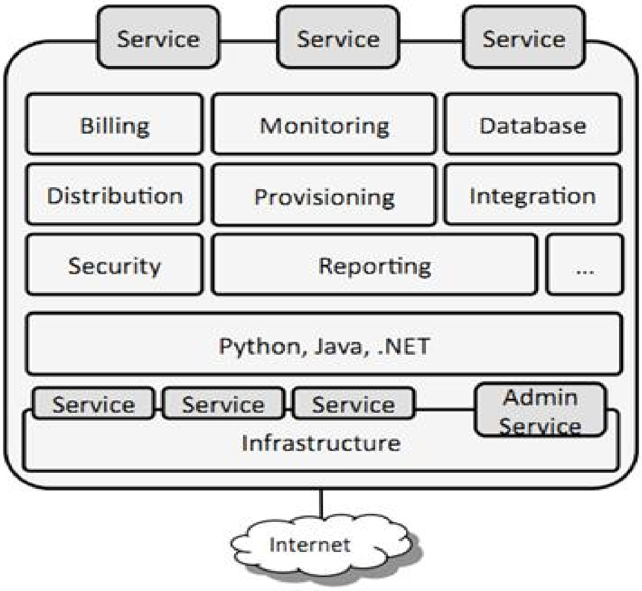


Fig. 5. SaaS model [15]

Infrastructure-as-a-Service (IaaS)

This cloud model offers a full service outsourced packages without the necessity of buying a server, software, data center space or any network equipment.[13][14]

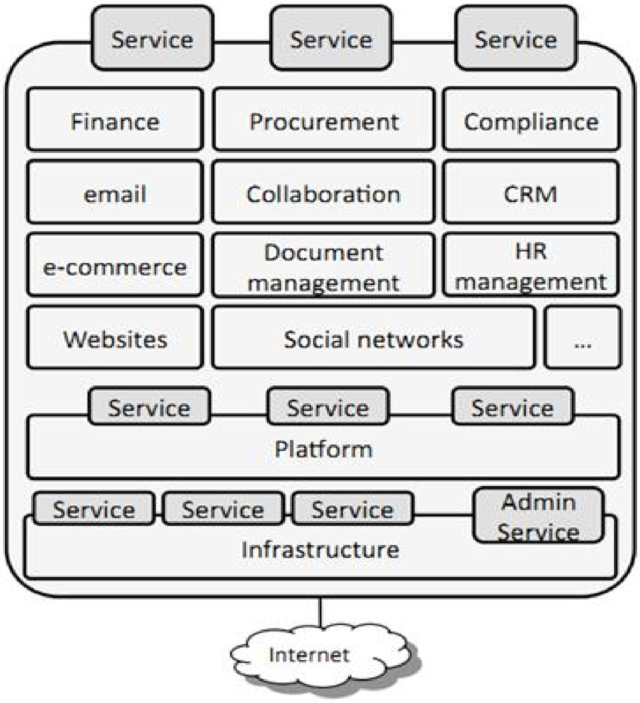


Fig. 6. IaaS model [15]

Business Process-as-a-Service (BPaaS)
BPaaS combines all the other service and it is used for billing, payroll and human resources.

Banking Industry Trends:
What the banks do not understand, and it is difficult to switch to another way to run things, is that the customer can't be controlled. This is known by the new banks using online platforms for its customers.
Banks need to look at the things from the outside, from the point of view of the consumer, to change something in the banking system.
Offering services that best fits the character and needs of each individual, banks want to replace the information offices and queues. [12]
Private cloud has come to dominate core banking. Banks are aware of the potential security breach or disruption in areas such as transactions and withdrawals by utilizing ATM. Banks must keep its core banking processes under control in their database to know when data are used. [12]

5.1. Pros for actual level

1. *Reduced costs*: using cloud banks do not have to invest so much in the software, hardware and labor. [16]
2. *Highly flexible*: Cloud Platform provides the ability to respond quickly to market changes, customer needs but also to respond quickly technological. Capacity will be an important competitive advantage. [16]
3. *Faster customer service*: Free Cloud services and products developed and released easily. Banks will be able to increase computing power to meet peak demand without having to improving technology. [16]
4. *It strengthens the relationship between customer and bank*: The combination of Big Data and unlimited computing power will allows to banks to develop systems that will make better decisions for their clients. [16]
5. *It brings customers closer to the banks*: transactions between buyers and sellers will be done easily at the moment. La more work needed to process the payments are inefficient because they use different technologies. This deficiency can be remedied in the Cloud. [16]

5.2 Cons for actual level

1. *Security and Privacy* - These two concepts are the most important when it comes to date. Keeping in mind that your client entrusts his personal data, shows a very high confidence and the bank must ensure the security and confidentiality of date. Cloud does not stay in this chapter, and this was mentioned by the founder of the cloud, Rajat Bhargav: "When you use cloud, you have a network that is open to the rest of the world. Cloud is more insecure than it was the repository data to headquarters." [17]
2. *Downtime* outages and downtime data, this is due to internet connection. [19] A disconnect from the network or from the Internet for a few minutes has a huge impact on a bank because, most likely, in time occur certain transactions or other exchanges of information that are lost in a year time.

The types of losses [18]:

- *Loss of the application service*: very much depends on the application and the bank branch.
- *Data loss*: if such an incident data may be lost and this has a financial impact but also legal

3. Vulnerability: We have to consider that any system is vulnerable to cyber-attacks, and banks in turn are not protected from hackers. Public and private clouds can be affected by both malicious attacks and infrastructure failures such as power outages. Such events can affect Internet domain name servers, prevent access to clouds or directly affect cloud operations. [19]

6. Opportunities

Since there are many doubts and suspicions about the security of cloud computing systems, opportunities for them to be integrated into banking systems (or in large organizations) appeard, once new methods of increasing the level of security were developped. Some of the possibilities of increasing security are: Kerberos authentication servers, firewall, VPN systems and virtualization.

The most powerful and widely used authentication service is Kerberos Authentication Server world. It was created in the project Athena, Get MIT (Massachusetts Institute of Technology). It allows users to communicate on the network to disclose their identity and to authenticate, preventing Iiniilor listening situations. Kerberos performs data security through encryption. The Kerberos is an authentication protocol based on a trusted authority called trusted third party (trusted third party). Kerberos works by providing the users or service vouchers, which are used for identification, and some cryptographic keys required for secure communication network. The Kerberos system is a relatively inexpensive option, but that ensures a better level of security.[21]

In general, a firewall (sometimes called bridge security) is a system that requires access control policy between two networks. A firewall is to implement this policy in terms of network configuration, one or more host systems and routers with special functions, other security measures such as customer authentication cryptographic methods. In other words, a firewall is a mechanism used to protect a trusted network from the point of view of security unsecure, we can not trust. Typically, one is the internal networks of organizations / banks (safe, reliable), while the other is network Cloud (they do not have confidence in terms of security) .[22]

VPNaaS (Virtual Private Network as a Service) is a solution of different market requirements: companies / banks want their mobile employees can access their internet network through a solution managed and controlled. For economic reasons desired VPN Cloud outsourcing providers by this system. NCP (Next Generation Network Access Technology) is a VPN solution which complies with the requirements and desires of suppliers and users, bringing benefits to both parties. Another plus for using VPNs in Cloud, considering the parallel made above with Internet banking is that banks offering Internet Banking systems used internally VPN systems to ensure greater data security. [23]

Another opportunity for cloud computing is virtualization. Virtualization is software that separates physical infrastructures to create various dedicated resources. It is the fundamental technology that powers cloud computing. "Virtualization software makes it possible to run multiple operating systems and multiple applications on the same server at the same time," said Mike Adams, director of product marketing at VMware, a pioneer in virtualization and cloud software and services. "It enables businesses to reduce IT costs while increasing the efficiency, utilization and flexibility of their existing computer hardware." In contrast, with virtualization, companies can maintain and secure their own "castle", Rick Philips said. This "castle" provides the following benefits**:**

- *maximize resources* — Virtualization can reduce the number of physical systems you need to acquire, and you can get more value out of the servers. Most traditionally built systems are underutilized. Virtualization allows maximum use of the hardware investment;
- *multiple system* — With virtualization, you can also run multiple types of applications and even run different operating systems for those applications on the same physical hardware;
- *IT budget integration* — When you use virtualization, management,

administration and all the attendant requirements of managing your own infrastructure remain a direct cost of your IT operation. [20]

7. Conclusions

In recent years, cloud computing has grown considerably and services offered increasingly better, this development will not stop.

Expanded areas where most is the bank has expanded greatly in this area because it offered many advantages as a customer.

The advantages are: cost saving, using cloud servers and applications and platforms made available instead of using personal servers and software purchased from specialty companies in banking will save a lot of money

But unfortunately like any tool it has drawbacks, the most common drawback in cloud computing is security and downtime. Considering the fact that we are in the 21st century nothing is safe as long as it is in a database, it can be broken easily by people specialized in IT, cloud computing has therefore created a model , private cloud, especially for banking institutions to avoid problems with security.

And about the downtime this is very difficult to avoid because not only refers to the banking system but in all market areas by using a network.

Cloud computing and cloud infrastructure have become a great ally for some areas, very popular after market, these areas are the banking and mobile networks as well as that of small and medium enterprises.

References

[1] “Security Advantages of Cloud Computing “, John Wood & Rick Tracy, 25.01.2011, unpublished

[2] “Assessing the Security Benefits of Cloud Computing”, Craig Balding, 21.07.2008, unpublished

[3] “Advantages and Disadvantages of Cloud Computing – Cloud computing pros and cons”, Ilias Tsagklis, 23.04.2013

[4] “Cloud computing & small businesses security -pros and cons”, Dan Conlon, Trend Micro, 2011

[5]”Cloud Computing Security – Network and Application Levels" - CloudTweaks.com

[6] From Wikipedia, the free encyclopedia - article "Cloud computing security"

[7] "Cloud Computing Security – Network And Application Levels" - Mojgan Afshari - senior lecturer in the Department of Educational Management, Planning and Policy at the University of Malaya

[8] "Cloud computing and application security: Issues and risks" - Kevin Beaver, CISSP-independent information security consultant

[9] "Data Security in the Cloud" - Vormetric si Custom Solution Grup

[10] “Cloud computing security" - wikipedia

[11] "Cloud Security: Safely Sharing IT Solutions" - Dan Griffin, Tom Jones

[12] A new era in banking - Cloud computing changes the game - Accenture

[13] Cloud Computing in Banking - Capgemini

[14] Cloud computing - Wikipedia

[15] WHAT ARE SERVICE MODELS IN CLOUD COMPUTING? - CLOUD COMPETENCE CENTER

[16] Six reasons why cloud computing will transform the way banks serve clients – and the five hurdles to overcome – bankingtech.com

[17] Cloud computing is the future but not if security problems persist- tech time

[18] Downtime, Outages and Failures - Understanding Their True Costs - Martin Perlin

[19] Advantages and Disadvantages of Cloud Computing – Cloud computing pros and cons - Illias Tsagklis

[20] “Virtualization vs. Cloud Computing: What's the Difference?” - Sara Angeles, Business News Daily, 2014

[21][22] „Securitatea arhitecturilor de tip Cloud” - Cloud Computing, Clubul Informaticii Economice - Cyber Knowledge Club, pg. 57-58 , 62-63

[23] „NCP's Cloud VPN Solution - Provider and Users on Cloud Nine", NPC Network Communications Products engineering GmbH, 2014

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.

Madalina-Elena STEFANET studies at Academy of Economic Studies from Bucharest, Faculty of Cybernetics, Statistics and Economic Informatics since 2013.

Ioan-Florentin TAMASESCU studies at Academy of Economic Studies from Bucharest, Faculty of Cybernetics, Statistics and Economic Informatics since 2013.

Smaranda-Elena TINTOIU studies at Academy of Economic Studies from Bucharest, Faculty of Cybernetics, Statistics and Economic Informatics since 2013.

Alin VEZEANU studies at Academy of Economic Studies from Bucharest, Faculty of Cybernetics, Statistics and Economic Informatics since 2013.

In-memory databases and innovations in Business Intelligence

Ruxandra BĂBEANU, Marian CIOBANU
University of Economic Studies, Bucharest, Romania
babeanu.ruxandra@gmail.com, marianciobanu6@yahoo.com

The large amount of data that companies are dealing with, day by day, is a big challenge for the traditional BI systems and databases. A significant part of this data is usually wasted because the companies do not own the appropriate capacity to process it. In the actual competitive environment, this lost data could point up valuable information if it was analyzed and put in the right context. In these circumstances, in-memory databases seem to be the solution. This innovative technology combined with specialized BI solutions offers high performance and satisfaction to users and comes up with new data modeling and processing options.

Keywords*: Business Intelligence, in-memory database, SAP HANA, data modeling, text processing*

1 Introduction

Business Intelligence solutions are not a novelty any more for the big companies and not only for these ones. The competitive advantages given by the usage of BI tools are significant. Having at the right time a report concerning a certain region or a certain product sales, or a report predicting the way a new product launch on the market will influence the company, brings a great advantage against competition. This will materialize quickly in revenues and in a higher trust level of the consumer.

Lots of companies, even though they have implemented Data Warehouses and Business Intelligence solutions, have big problems in obtaining the data in time. There are cases where the execution of a report takes between a few hours to several days, and the time lost for this execution influences, directly or indirectly, the profit. Therefore, taking the right decision at the right time, meaning before the competitors, definitely represents an important move on the market.

The hard work in obtaining the data at the right moment is, in most of the cases, due to a combination of hardware support and inappropriate software. Standard databases store the data on disc and the I/O operations are very slow compared to those made in RAM memory.

The need of processing large amounts of data very fast was the main reason that led to the development of in-memory databases. These databases revolutionized the entire Business Intelligence area and managed to bring outstanding results exactly where the classic databases failed.

2. General characteristics of in-memory databases

As their name suggest, the in-memory databases stop storing the data at disk level, moving instead the storage to the memory. This change comes with some great advantages, but also brings a set of risks.

Considering that the main characteristic of these DBMSs is the memory storage, which can lead to amazing processing speeds, there is also a big risk to manage, and that is the risk in loosing data in case of an unplanned restart of the system or in case of problems with the power. Therefore, for these type of databases the ACID (Atomicity, Consistency, Isolation, Durability) concept is not fully accomplished [1]. The problem appears at the last property, *durability*, which is not fully satisfied. There are several solutions that can assure the *persistency of data,* some of them being already implemented by the main in-memory databases producers [1].

From a *hardware perspective*, the following solutions can be implemented:

- use of NVDIMM (non-volatile dual in-line memory module) memory, a special type of memory that has the property of keeping the data in case of an unexpected incident, such as an unexpected system restart or an electric power failure;

- use of NVRAM (non-volatile random access memory) memory, which is a special type of memory having similar properties as NVDIMM memory type.

From a *software perspective*, there are also several solutions that can be implemented to assure the durability of data:

- use of *a journal file*, which store a log of all transactions in the database;
- use of "*high availability*" implementation, which suppose replication of the database on a secondary database, and, in case of major issues of the primary database, the standby database will replace automatically the primary database; it is recommended to use this method along with any other method mentioned above, in order to assure complete recovery in case of system crash;
- use of *snapshot files*; these files keep the state of the database at different moments of time, but the major disadvantage in this case is that the most recent changes will not be captured. The recommended approach in this case is to use this method concurrent with any other method mentioned above.

The advantages brought by an in-memory database are, first of all, due to the fact that the slow I/O disk operations don't exist anymore. Here a few outlines regarding the in-memory databases:

- very *fast processing* of large amounts of data;
- *reduced time of blocking the records*, because all the insert, update and delete operations are made in real time;
- *fast data queries*.

Regarding the ***data access method***, there are defined different indexing methods, based either on hash functions or on several tree types. *T-tree* is a special tree type, created particularly for these kind of databases. It is a balanced tree and it is especially optimized for indexing the data stored exclusively in memory. The main characteristic of this tree is that the *index is not storing any effective data*, is storing only a pointer to the data, because all is kept in memory.

In-memory databases use for ***data representation*** the column store model. This leads to an efficient data store, because for building the relational tuples they use the pointers. This way the effective value is stored only once in a dictionary and, instead of storing the same value many times in the memory, the tuples will be built by storing only the pointer to that value. If we have to deal with large values appearing many times in the database, the storage space will be this way significantly improved.

The ***performance*** is the key characteristic for the in-memory databases. This is not any more influenced by the slow I/O operations from the memory to the disk and backwards, it is just based on the processing speed. Therefore, an appropriate hardware will certainly assure an optimum response time of the database.

In order to assure the ***physical integrity*** of data, the *backup files* must be saved on a safe storage support. In case of a major incident, the database has to be recovered from the last backup file and afterwards has to be updated using the *journal file,* which logs all the transactions in the database.

3. Short history

The need of performance in the IT domain combined with the advantages of in-memory computing are the main factors that influenced the appearance of in-memory databases. An in-memory DBMS uses the memory as the main storage support, compared to the classic DBMSs that use the disk as the main storage place.

The first in-memory databases appeared in the early '90s. In 1992, White Cross Systems Limited Company offered the first version of this type of DBMS, sold nowadays as "Kognitio Analytical

Platform", which was firstly delivered on proprietary hardware. In1993, the company Perihelion Software launched the in-memory DBMS called "Polyhedra" [1],[3].At that time, Polyhedra was designed to "keep the working copy of the data in memory" and it was based on a client- server architecture [1]. Polyhedrais now sold by ENEA AB.

In 2001 was released eXtremeDB, belonging to McObject company. It was an advanced DBMS, conceived for in-memory storage and designed to use minimal CPU and memory resources. It was designed especially for the real-time embedded systems [1].

The main in-memory database producers nowadays are:

- Oracle, with Oracle Database 12c – In Memory option, released in 2014;
- SAP, with SAP HANA (High - Performance Analytical Appliance), launched in 2010;
- IBM, with DB2 BLU, released in 2013;
- Microsoft, with SQL Server 2012 (actual SQL Server 2014), released in 2012.

4. Main characteristics of SAP HANA

SAP HANA is a complex technology, developed in C++ and running on SUSE Linux Enterprise Server, capable of fast processing of large amounts of data in real time. Also, it is a *platform* used to support the developing of real time applications and, not least, it is an in-memory database, which can sustain a large data warehouse.

SAP HANA, as an in-memory database, comes with some special features, using an innovative hardware and software. It offers the possibility of analysing big volumes of data and the flexibility in analysing various types of data, not only the traditional ones.

SAP HANA must be seen as a complex platform, designed to sustain and improve all the business processes. It can handle not only the Analytics and Business Intelligence processes, but also the OLTP transactions, as it can be delivered as support for SAP ERP.

Regarding the architecture, in *Figure.1 – HANA Architecture*, we can observe the main components of SAP HANA platform [5].

HANA is a system that has the in-memory database as the foundation. To this, it can be added several components and add-ons to support various functionalities: real-time data replication, monitoring instruments, release management instruments etc.

HANA database includes four dedicated servers, as follows [5], [7]:

- *IndexServer* – the main component because it is the place where the data is effectively stored; it also contains the data processing engines;
- *Preprocessor* – the server used in case of text processing and text analysis;
- *NameServer* – holds the information about the system landscape and in a distributed system its main responsibility is to locate the data;
- *StatisticsServer* – keeps information about the performance parameters: system status, CPU usage, memory usage etc.

To be noticed that, in case of a multi-node cluster configuration, the four servers can be found on each node.

The main capabilities HANA offers are stated below:

- *optimized analytics data models*: Analytic View, Attribute View, Calculation View etc.
- *removal of data caching*;
- taking over the *complex operations pushed at database level* by the specially optimized BI client tools which have as source data HANA database;
- including the *SAP HANA predictive* functionality.

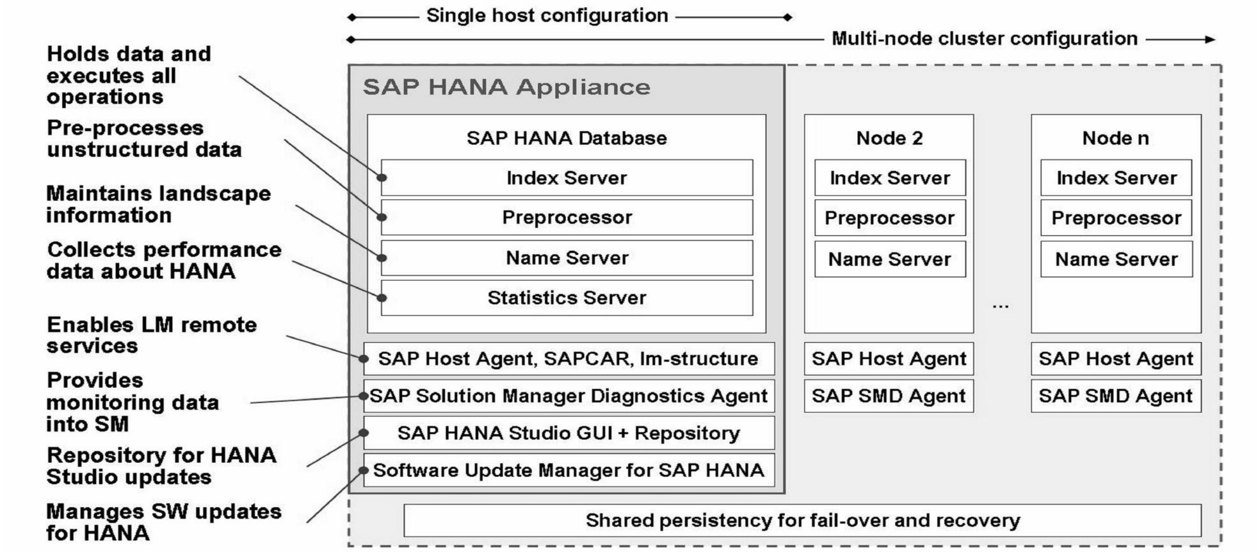

Single host configuration
Multi-node cluster configuration
Holds data and executes all operations
Pre-processes unstructured data
Maintains landscape information
Collects performance data about HANA
Enables LM remote services
Provides monitoring data into SM
Repository for HANA Studio updates
Manages SW updates for HANA
SAP HANA Appliance
SAP HANA Database
Index Server
Preprocessor
Name Server
Statistics Server
Node 2
Index Server
Preprocessor
Name Server
...
Node n
Index Server
Preprocessor
Name Server
SAP Host Agent, SAPCAR, lm-structure
SAP Solution Manager Diagnostics Agent
SAP HANA Studio GUI + Repository
Software Update Manager for SAP HANA
SAP Host Agent
SAP SMD Agent
SAP Host Agent
SAP SMD Agent
Shared persistency for fail-over and recovery

Fig. 1. Hana Architecture

5. Business Intelligence with Sap HANA

Considering the capabilities HANA possesses, the performance of the BI solutions which have as source data this database are remarkable. SAP created some special functionalities for the BI solutions, some of them specially optimized for SAP HANA.

The *interface* through which the user interacts with the system is SAP HANA Studio. This working environment is based on Eclipse and offers the user the possibility of working in several perspectives (Information modeling, Application development, Administration, Monitoring and security etc.).

Before to get to the data modeling step, first we have to bring the data in HANA. An overview regarding the data acquisition in HANA can be found in *Figure 2 – Different ways for data acquisition in HANA* [5].

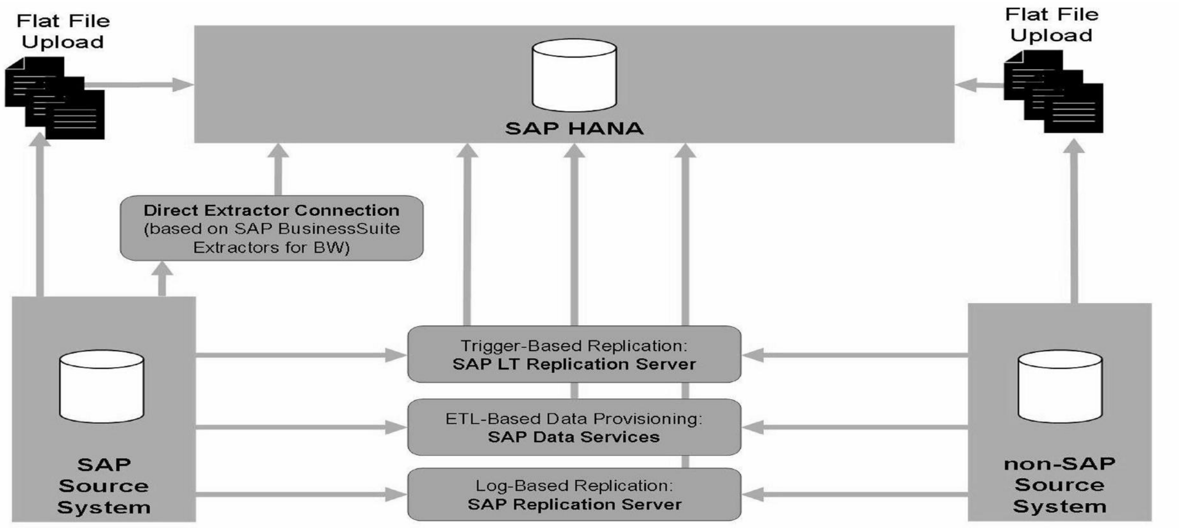

Flat File Upload
Flat File Upload
SAP HANA
Direct Extractor Connection (based on SAP BusinessSuite Extractors for BW)
Trigger-Based Replication: SAP LT Replication Server
ETL-Based Data Provisioning: SAP Data Services
Log-Based Replication: SAP Replication Server
SAP Source System
non-SAP Source System

Fig. 2. Different ways for data acquisition in Hana

Loading the data into HANA can be done in one of the following options:

- Direct load from *flat files* (*xls*,*.xlsx* or *.csv)* through an interface accessible from HANA Studio;
- *SAP Data Services*–software solution for data integration from various sources, for data quality management and for text analysis;

- *SAP Landscape Transformation* – tool used for real-time replication of data, either from SAP source system or from non-SAP source system;
- *SAP Direct Extractor connection* –tool used to import data from SAP Business Suite using specific connectors.
- *SAP Replication Server* – advanced tool used for transferring and synchronizing the data across the organization.

The data can be stored in "column-based" tables or in "row-based" tables; SAP's recommendation is to use the tables in the first category in order to assure high level of performance.

For the data modeling, in HANA Studio the Modeller Perspective must be accessed. The modeling objects available in HANA are:

- ***Attribute Views***– reusable objects, used for *describing the data*; attribute views *"are used to give context. This context is provided by text tables, which give meaning to data*" [6]. They can contain dimensions, hierarchies and calculated attributes. For example, in the banking domain, if we have a fact table containing the IDs of the customers, we could use an attribute view to make available more information about the customer, such as name, age, monthly revenue etc. In this way, the IDs acquire a meaning and are framed within a context;
- ***Analytic Views*** – they offer the possibility of data modelling using measures (numeric values) and aggregations. These views can contain *Attribute views*, which are linked to the fact table according to the *star schema mode.* In SAP HANA, these are the fastest modeling objects, but the limitation is that an analytic view can be created using only one fact table [6];
- ***Calculation Views*** – objects that offer the greatest flexibility in data modeling. These are similar to Analytic views and offer the possibility to define complex calculation. The main difference as against analytic views is that the calculation views offer the option to *join various fact tables* for using different measures. Furthermore, there are two ways of creating this type of objects: *graphical modeling* or *scripting modeling* in case there are complex operations that cannot be covered by the graphical model.
- ***Decision Tables*** – objects that allow business rules modeling regarding decisions, in a certain context. These objects use the decision trees logics. A decision table can be created using an existing table in HANA, which will represent in this case the data source of the object. Next step is to define the attributes that will be considered in the decision, along with their associated conditions and the action that must be taken when the defined conditions are fulfilled.

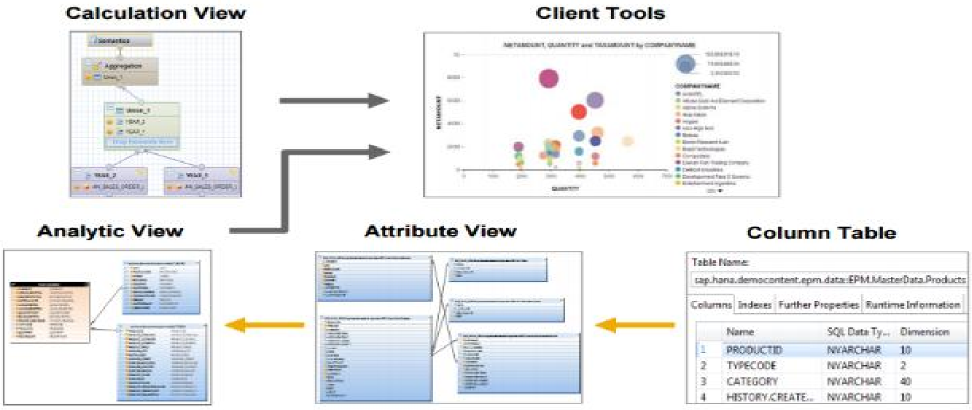


Table Name:
sap.hana.democontent.epm.data::EPM.MasterData.Products

Columns | Indexes | Further Properties | Runtime Information

	Name	SQL Data Ty...	Dimension
1	PRODUCTID	NVARCHAR	10
2	TYPECODE	NVARCHAR	2
3	CATEGORY	NVARCHAR	40
4	HISTORY.CREATE...	NVARCHAR	10

Fig. 3. Query views in Hana

Figure 3- Query views in HANA offers a better view of how all of these analytic objects can be used [6].

If the user' requirements are not fulfilled by the graphic modeling options, HANA offers the possibility to create stored procedures to define the desired logic, using SQLScript languange.

SAP HANA has 3 different engines used for processing the queries, each one used depending on the queried object, as can be seen in *Figure 4- Processing engines in HANA* [8]:

- **Join engine** – used when an Attribute View is queried;
- **OLAP engine** – used when a query based on an Analytical View is executed;
- **Calculation engine** – used when Calculation Views or Analytic Views with calculated columns are queried.

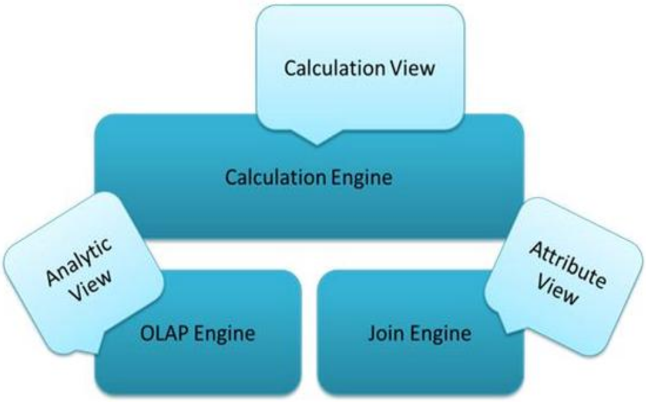


Fig. 4 Processing engines in HANA

A functionality implemented in HANA and frequently used in BI reporting is **the capacity of processing geospatial data**. A report revealing the nearest delivery point to the shop or which shop doesn't have enough storing space to assure the stocks are very important and is definitely more suggestive if this information is presented on a map. In this respect, in HANA there is a special engine called the *Spatial Engine* that is designed to support geospatial data and is integrated as an extension of the Calculation engine. The geospatial data can be accessed and operated through a standard set of methods [6].

The text analysis capacity implemented in HANA is designed to meet the actual challenges of the companies. Synthetizing or extracting the relevant and important information from the immense and various amount of data a company faces, is an extremely important and useful functionality HANA owns [6].

There are two types of analysing the text [6]:

- "**Fuzzy search**" – allows searching of strings that only follow a defined pattern. For example, if we have a table designed for storing city names and we wish to search for "Munchen" in the specified column, there will be returned also the records storing names like "Munich", "Munhcen", "Munhen" etc.;
- "**Text Analysis**" – offers the option of complex text analysis on various strings, by defining some analysis rules for different industries and various languages. This allows *extraction of important information from unstructured data*, such as:

- ***identifying parts of speech*** (nouns, verbs etc.);

-***named entity recognition***: locations, persons, currency, dates etc.;

. ***sentiment analysis,*** meaning that HANA offers the possibility to analyse a text and extract information that suggest if the communicated feelings are positive or negative.

In *Figure 5 –Text Processing in HANA* we can observe the entire flow of text processing in HANA [6]. Therefore, starting with an unstructured text data source and using dedicated algorithms, HANA analyses and extracts the most important features from the text. The result is afterwards stored in a table, as a structured text and can be used in reporting or analyses [6].

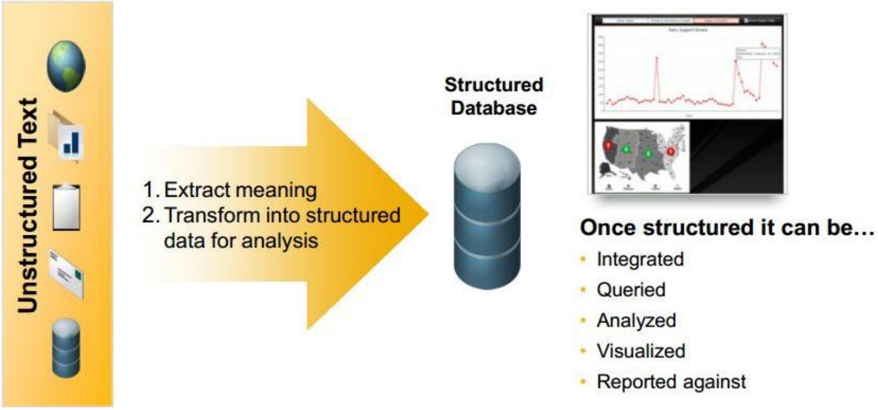


Fig.5. Text Processing in HANA

The Business Intelligence tools, as presented in *Figure 6 – BI tools* [6], can use the result of different types of data modeling available in HANA. In order to improve the quality of the querying process, the dedicated BI tools have been optimized to *push for execution at database level most of the complex operations*. This ways is promoted the concept of doing a small amount of operations at client level (application), many of them being optimized for execution at database level.

HANA data models can be a data source for the following tools, some of them not necessarily being specialized for the Business Intelligence domain:

- *Microsoft Excel* – the connection is made through ODBC connectors;
- *SAP Crystal Reports* – specialized tool used for "pixel perfect" designing of the reports;
- *SAP Business Objects Explorer* – allows exploring data in a fast way; it offers advanced options for viewing the data, providing also various graphic types;
- *Analysis for Office* – is an add-on installed on Microsoft Excel that allows data displaying as a datasheet and it offers special options for viewing the data, filtering and quick accessing;

- *Analysis for OLAP* – used for analysing and displaying the data in a Web environment;
- *Web Intelligence* – tool used for creating Web, Desktop or Mobile reports;
- *SAP Business Objects Dashboards* – used for creating dashboards.

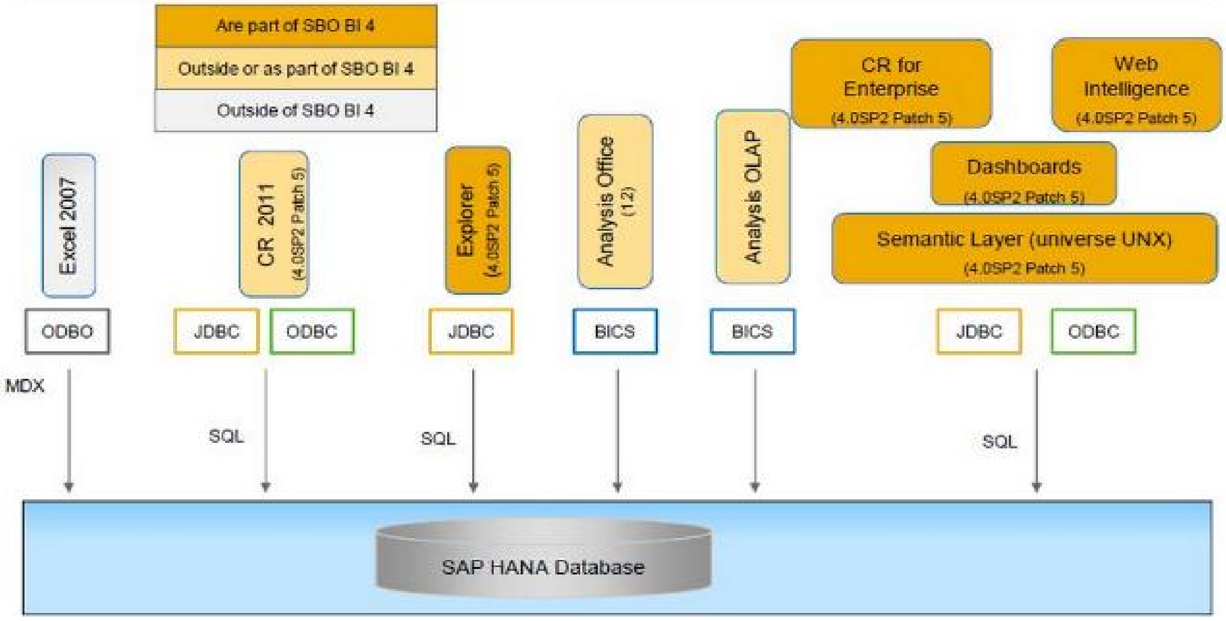

Are part of SBO BI 4
Outside or as part of SBO BI 4
Outside of SBO BI 4
Excel 2007
CR 2011 (4.0SP2 Patch 5)
Explorer (4.0SP2 Patch 5)
Analysis Office (1.2)
Analysis OLAP
CR for Enterprise (4.0SP2 Patch 5)
Web Intelligence (4.0SP2 Patch 5)
Dashboards (4.0SP2 Patch 5)
Semantic Layer (universe UNX) (4.0SP2 Patch 5)
ODBO
JDBC
ODBC
JDBC
BICS
BICS
JDBC
ODBC
MDX
SQL
SQL
SQL
SAP HANA Database

Fig. 6. BI Tools

6. Conclusions

In-memory databases are a technology that can bring tremendous value in a company. We must outline that this DBMSs are *column oriented*, assuring a high level of data compression this way. Additionally, *for indexing it is used a special algorithm,* usually based on T-tree. These systems offer **high performance**, even though they imply certain risks regarding the loss of data. There are many mechanisms that can be implemented in order to assure a safe recovering of data, such as: *high-availability* implementations, usage of *backup* copies and *journal* files etc.

SAP HANA is a top technology in this area and innovates the BI domain by bringing a variety of data modeling options and fast data processing.

References

[1] "In-memory database -Unabridged Guide" Walter Jennings, 2012

[2] "Main memory database systems: an overview" http://pages.cs.wisc.edu/~jhuang/qual/main-memory-db-overview.pdf

[3] "List of in-memory databases" http://en.wikipedia.org/wiki/List_of_in-memory_databases

[4] "In- memory database" http://en.wikipedia.org/wiki/In-memory_database

[5] "HA100 - SAP HANA Introduction", SAP SE (training material)

[6] "HA300 – SAP HANA Implementation and Modeling", SAP SE (training material)

[7] "An insight into SAP HANA Architecture" http://sapHANAtutorial.com/an-insight-into-sap-HANA-architecture/

[8] "Understanding SAP HANA Engine" http://sapHANAtutorial.com/sap-HANA-engine/

Ruxandra BĂBEANU graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2013. She is a currently graduating from the master program Databases-Business support at the Bucharest University of Economic Studies.

Marian CIOBANU graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2013. He is a currently graduating from the master program Databases – Business support at the Bucharest University of Economic Studies.

Applying BI Techniques To Improve Decision Making And Provide Knowledge Based Management

Alexandra Maria Ioana FLOREA
Bucharest University of Economic Studies
alexandra.florea@ie.ase.ro

The paper focuses on BI techniques and especially data mining algorithms that can support and improve the decision making process, with applications within the financial sector. We consider the data mining techniques to be more efficient and thus we applied several techniques, supervised and unsupervised learning algorithms The case study in which these algorithms have been implemented regards the activity of a banking institution, with focus on the management of lending activities.

Keywords: *Business Intelligence, Data Mining, Naïve Bayse, Support Vector Machine,*

1 Introduction

Business Intelligence refers to information systems for identifying, extracting and analyzing data available in enterprise systems whose purpose is to provide real support for business decisions [1].

[2] identifies a various number of BI techniques such as: predictive modelling through which we can predict value for a specific data item attribute; characterization and descriptive data mining used for data distribution, dispersion and exception; classification, used to determine to which class a data item belongs; clustering and outlier analysis which partitions a set into classes, whereby items with similar characteristics are grouped together; OLAP (OnLine Analytical Processing) with tools that enable users to analyze different dimensions of multidimensional data (for example, it provides time series and trend analysis views). Other techniques we can mention are Association, correlation, causality analysis (Link Analysis), Temporal and sequential patterns analysis, Model Visualization, Exploratory Data Analysis (EDA).

Machine learning can offer a set of tools that could be useful in summarizing various types of unliniar connections between data. [3]

A new concept that is quickly making its way in the knowledge management efforts is the use of Big Data. As mentioned in [4] Big Data found its way quickly in online shopping. For example we can identify the behavior of each customer, even by correlating his logins with IP addresses for tracking views when he is not authenticated. With the help of such analyses we can identify the products or the class of products that even though are being viewed and/or added to the cart they aren't eventually bought as much as other products.

There are almost unlimited options for using BI techniques to support the decision process in any type of business, from energy power plants to financial institutions.

In our research, the knowledge discovered by applying data mining techniques will enable representatives of a financial institution to asses the lending activity of the institution. The solution we are presenting in this paper represents a case study whose aim is to develop an integrated solution which automates the sales processes of a banking institution, with focus on the management of lending activities.

2. Algorithms used for determining the likelihood of contracting the loan

In order to implement the functionalities regarding the decisions to grant or reject loan applications it is necessary to develope performant algorithms to determine the values of the pre-scoring and scoring processes, to estimate the probability of granting the loan and to determine the maximum threshold for the amount awarded.

The objectives are:

- creating a scoring model, which contains characteristics that are relevant and have impact on lending decision as well as financial information, the relationship with the bank, areas of interest;
- creating a predictive model which is based on data mining techniques such as logistic regression, the Naïve Bayes classifier and the Support Vector Machines algorithm;
- data analysis through clustering methods in order to obtain a profile of the customer.

The algorithms will take into account the financial indicators and socio-demographic data that characterize a customer and include the following:

Table 1. Financial and socio-demographic indicators related to customers

Attribute name	Variable type	Explanation
CNP	Varchar2	Identity number used to identify customers
ID_CERERE	Varchar2	Loan application ID
ID_PRODUS	Varchar2	Loan product ID
NUME_CLIENT	Varchar2	Client’s name and surname
PROFESIA	Varchar2	Client’s profession
SEX	Varchar2	Genul persoanei (feminin/masculin)
MONEDA	Varchar2	Currency in which the loan is requested. (RON, EUR, USD)
GARANTIEVALOARE	Number	Total value of collaterals offered by the client.
VALOARE_ESTIMATA_BUNURI	Number	Total value of assests owned by the client
VENIT_ANUAL_RON	Number	Total annual income
SUMA_INDATORARE_RON	Number	Total amount of debt
DATA_CERERE	Date	Date of the lending application.
VARSTA	Number	Client’s age
CATEGORIE	Varchar2	Type of credit.
DESCRIERE	Varchar2	Description of the requested loan.
PRESCORING/SCORING	Number	Value of calculated scor after applying the scoring algorithm (maximum 100)
SUMA_DEPOZIT	Number	Total amount of deposits owned by the client.
FIDELITATE	Number	Customer history relationship with the bank (0 - client without historical relationship with the bank, 1 - client with other products such as current account and / or other loans 2 - client with products including deposit)
STARE_CIVILA	Varchar2	Customer status (unmarried , married , divorced, widow)
SUMA_SOLICITATA	Number	The loan amount requested by the client

The algorithm is based on data mining techniques with supervised learning mechanism (classification algorithms using Bayes classifier, regression algorithms, significant attributes identification) and unsupervised learning mechanism (clustering). The following algorithms have been used

- *Naïve Bayes classification algorithm* is a supervised learning technique which enables the relationship between each independent variable and the dependent variable, by calculating a conditional probability for each of these relationships [5]. Thus the prediction is achieved by determining the effects of independent variables on the dependent variable

- The *SVM method* allows binary classification and is based on the structural risk minimization principle. The algorithm involves the following steps [6] separating classes using linear programming to obtain linear and nonlinear patterns of discrimination between data points; determining overlapping classes; application of kernel techniques to eliminate nonlinearity; determining the optimal solution.
- *Regression* is a method for determining the relationship between a dependent variable Y (response type variable) and one or more independent variables X1, ..., Xn (predictors or explanatory variables). Regression allows the determination of Y variation when the independent variables Xi values change. A value or range of values of the dependent variable for certain values of the independent variables can be estimated
- Determining significant attributes (Attribute Importance) is a method that is based on the "Minimum Description Length" (MDL) algorithm to classify a set of attributes depending on their usefulness in making predictions. This method significantly reduces the time and resources necessary in the calculation of prediction models by selecting a subset of meaningful attributes through the elimination of redundant, irrelevant or informal attributes and identification og those attributes useful in making predictions.
- *Clustering* is a method of determining the similarities and dissimilarities between elements of a set, in order to group them into distinct and homogeneous classes. The method involves a classification with unsupervised learning, in which it is not a priori known either the number of possible classes or the inclusion of objects in certain classes

The Naïve Bayes and SVM classification algorithms will be applied to determin the likelihood of contracting a credit, regression to determine the maximum amount that can be awarded based on the financial situation of clients and clustering will be used for grouping customers into clusters representing financial and socio-demographic profile thereof.

3. The analysis and reporting module

In the initial phase we identified two specific reporting requirements: an operational level of reporting that regards current lending applications and a tactical reporting level, which refers to the analysis of activity on a higher level in various areas, territories, sales agents, and product categories, types and groups of customers

In this regard two reporting modules are required. A module within the operating system, which will be built along with the process management application and a multidimensional analysis module will be based on a Data Mart that could be integrated with existing decision support system within the organization.

For designing the data mart it is necessary to identify objects of type: sizes, hierarchies, tables facts, mappings, this being modeled using UML stereotypes presented in Table 2.

Table 2. Stereotypes defined for multidimensional modeling

Object	Stereotype	Meaning
Class	<<Dimensiune>>	Dimension class
Class	<<Tabela_Fapte>>	Facts class
Attribute	<<O_ID>>	Identifying attribute
Attribute	<<O_DESC>>	Descriptive attribute
Attribute	<<ID_Parinte>>	Parent attribute (dimension class)
Attribute	<<Atrib_Dim>>	Attribute (dimension class)
Attribute	<<Masura_Baza>>	Measure attribute (facts class)
Attribute	<<Masura_Derivată >>	Calculated attribute (facts class)

For multidimensional analysis the following entities are designed:

- *Clienti* dimension with information regarding name, status, sex, adress, county, group and client type;
- *Zone* dimension with information regarding the unit, region and area;
- *Produse* dimension, with information on the name, category and type of product, associated fees and amount limits;
- *Utilizatori* dimension, with information on salespeople, managers and their roles;
- *Cereri* facts table, in which the data in the loand application is summarized;
- *Punctaj_scoring* facts table, in which the scorecards obtained on the basis of loan applications are mantained.

The class diagram for the multidimensionl objects is shown in Figure 1.

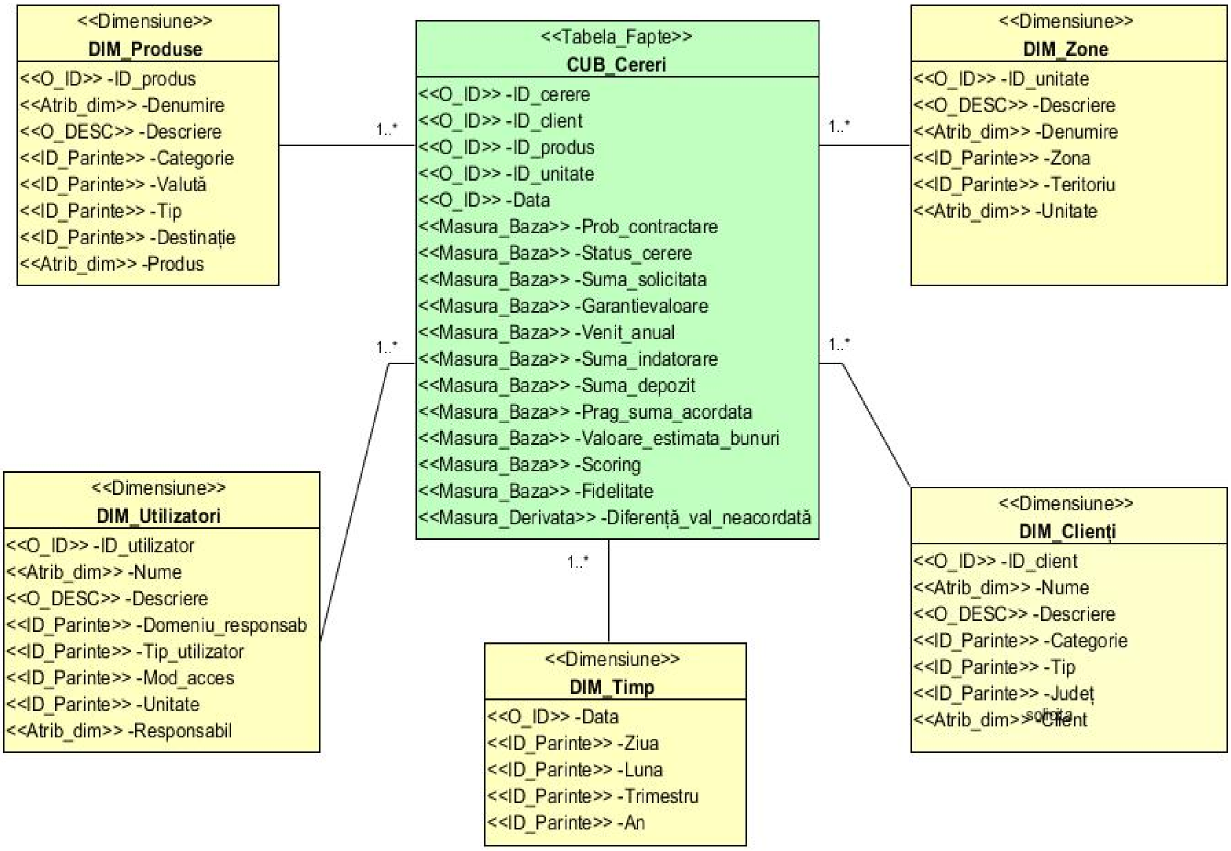


Fig. 1. Class Diagram for the Data Mart

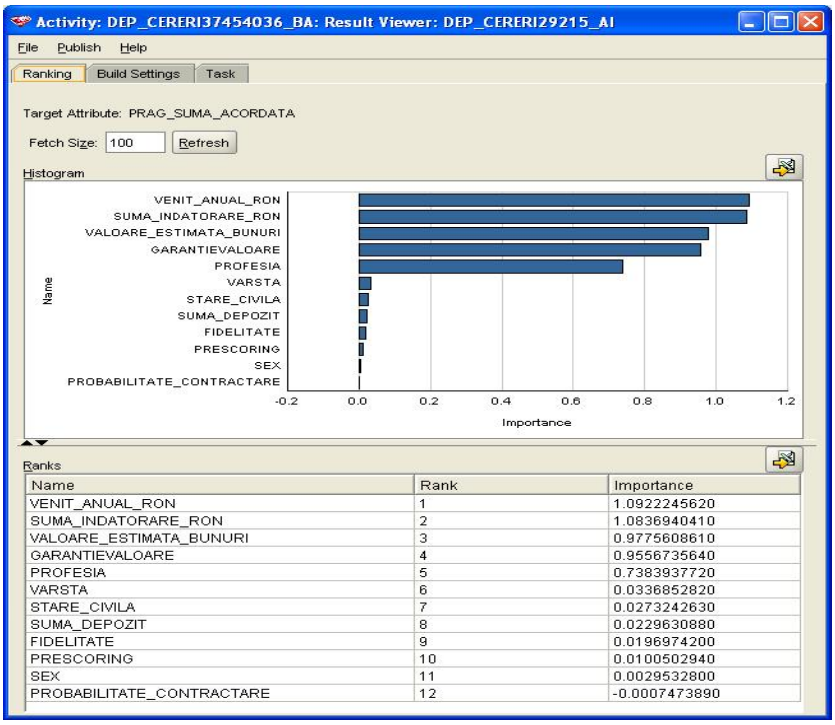
The objects of the multidimensional model can interconnect with the objects of the organizational data warehouse through XSD schemas.

4 Implementing the Data Mining algorithms

The full implementation of the system involves the realization of the prescoring / scoring algorithm used to calculate scores for each loan applications, implementing data mining algorithms to determine the likelihood of contracting a loan, the maximum amount that can be granted and grouping customers in clusters based on their financial and socio-demografic profile and building a data mart that can be used for multivariate analysis of the lending activity.

In the next section of the paper we will present how we implemented the Data Mining algorithms to determine the likelihood of contracting a loan.

We applied the algorithm for determining the important attributes and determine the maximum amount that can be awarded as shown in Figure 2.



Target Attribute: PRAG_SUMA_ACORDATA

Fetch Size: 100

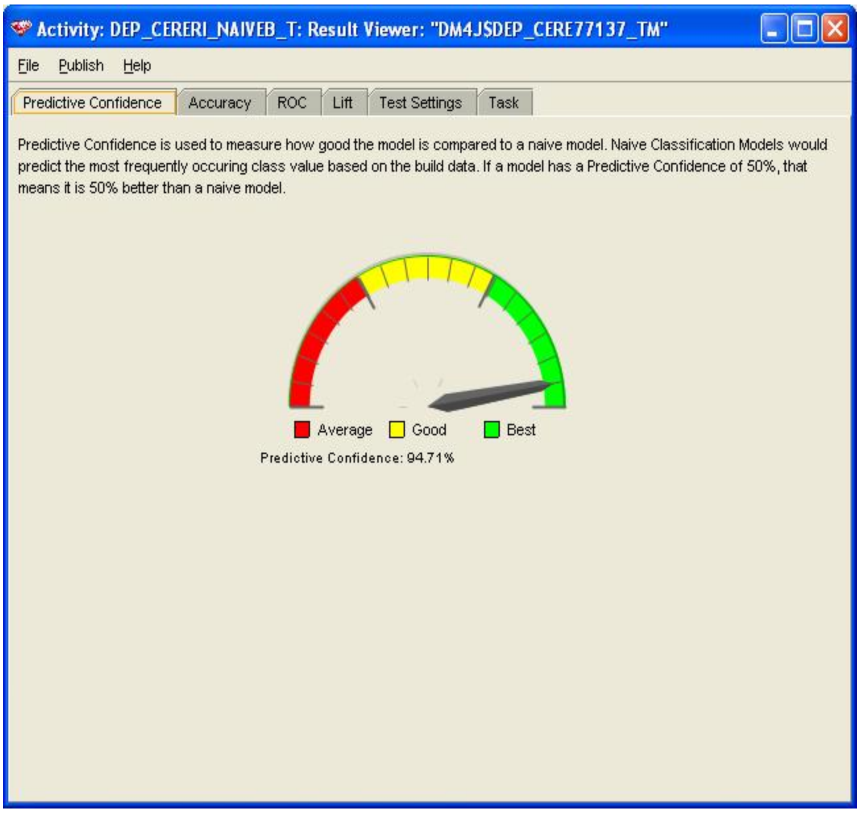
Name	Rank	Importance
VENIT_ANUAL_RON	1	1.0922245620
SUMA_INDATORARE_RON	2	1.0836940410
VALOARE_ESTIMATA_BUNURI	3	0.9775608610
GARANTIEVALOARE	4	0.9556735640
PROFESIA	5	0.7383937720
VARSTA	6	0.0336852820
STARE_CIVILA	7	0.0273242630
SUMA_DEPOZIT	8	0.0229630880
FIDELITATE	9	0.0196974200
PRESCORING	10	0.0100502940
SEX	11	0.0029532800
PROBABILITATE_CONTRACTARE	12	-0.0007473890

Fig. 2. The determination of significant attributes for the maximum amount that can be awarded to a client

We can observe from the analysis results that to determine the maximum amount that can be awarded to a particular customer, the attributes with the highest degree of relevance are: annual income, the total amount of debt, the total value of goods held by the client, the guarantees provided for the loan, profession, age, marital status, amount in the customer's deposits, if they exists, loyalty to the bank, prescoring/scoring value. These attributes will be used in the regression model that will determine the maximum allowable amount for lending.

5. Determining the lending likelihood

Next we determined the likelihood of granting the loan through two classification methods: Naive Bays and SVM. The target attribute is the probability of granting the loan with values 0/1, where 0 - loan is given 1 - no loan is given. After selecting the relevant attributes that are used in the algorithms and implementing the corresponding steps it is observed that models have a high accuracy of 94.71% and 94.34% for Naive Bays for SVM (Figure 3).



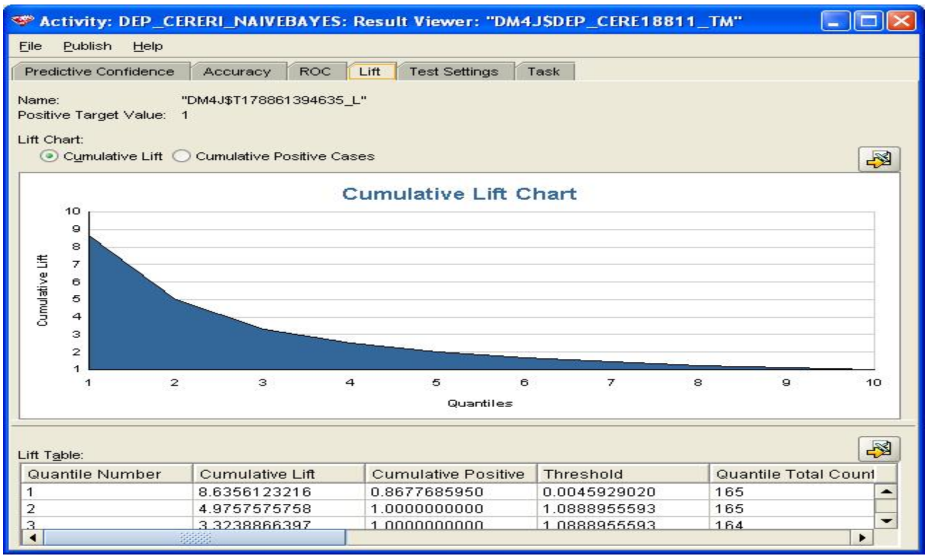
Predictive Confidence is used to measure how good the model is compared to a naive model. Naive Classification Models would predict the most frequently occuring class value based on the build data. If a model has a Predictive Confidence of 50%, that means it is 50% better than a naive model.

Predictive Confidence: 94.71%

Fig. 3. Naïve Bayes algorithm accuracy

To analyze the accuracy of the algorithm we analyzed a series of values that characterize the determination of the dependent attributes. Such a set of values is represented by the LIFT matrix which represents the learning rate of the model.

From the graph in Figure 4 we can observe a high rate of learning the model in the first three quintiles. Basically, the matrix is the ratio between the percentage of correct classification carried out and the percentage of real positive model classifications.



Name: "DM4J$T178861394635_L"
Positive Target Value: 1

Lift Chart:
● Cumulative Lift ○ Cumulative Positive Cases

Lift Table:

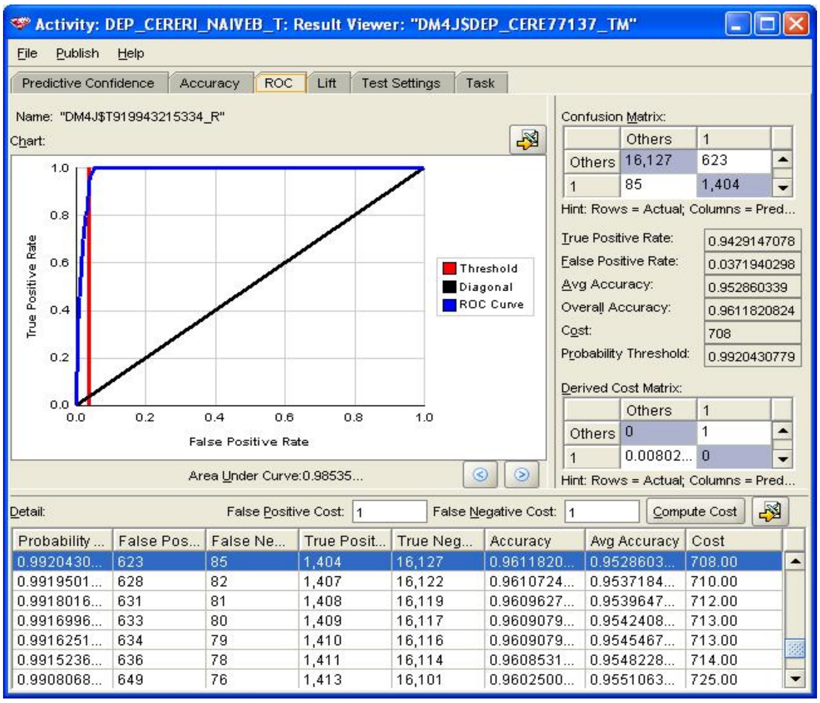
Quantile Number	Cumulative Lift	Cumulative Positive	Threshold	Quantile Total Count
1	8.6356123216	0.8677685950	0.0045929020	165
2	4.9757575758	1.0000000000	1.0888955593	165
3	3.3238866397	1.0000000000	1.0888955593	164

Fig. 4. LIFT Matrix for the Naïve Bayes algorithm

For the Naïve Bayes algorithm we can observe the ROC matrix (Figure 5) which represents a metric for comparison of existing (real) values with those predicted by the built model.

.

The ROC matrix, as well as the LIFT matrix, applies classification models and can be used to gain insight into the ability of the model to determine the values.



Name: "DM4J$T919943215334_R"

Chart:

Area Under Curve:0.98535...

Confusion Matrix:

	Others	1
Others	16,127	623
1	85	1,404

Hint: Rows = Actual; Columns = Pred...

True Positive Rate: 0.9429147078

False Positive Rate: 0.0371940298

Avg Accuracy: 0.952860339

Overall Accuracy: 0.9611820824

Cost: 708

Probability Threshold: 0.9920430779

Derived Cost Matrix:

	Others	1
Others	0	1
1	0.00802...	0

Hint: Rows = Actual; Columns = Pred...

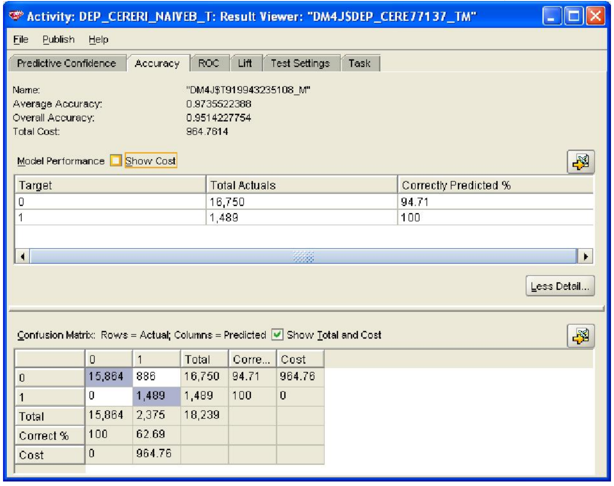
Detail: False Positive Cost: 1 False Negative Cost: 1

Probability ...	False Pos...	False Ne...	True Posit...	True Neg...	Accuracy	Avg Accuracy	Cost
0.9920430...	623	85	1,404	16,127	0.9611820...	0.9528603...	708.00
0.9919501...	628	82	1,407	16,122	0.9610724...	0.9537184...	710.00
0.9918016...	631	81	1,408	16,119	0.9609627...	0.9539647...	712.00
0.9916996...	633	80	1,409	16,117	0.9609079...	0.9542408...	713.00
0.9916251...	634	79	1,410	16,116	0.9609079...	0.9545467...	713.00
0.9915236...	636	78	1,411	16,114	0.9608531...	0.9548228...	714.00
0.9908068...	649	76	1,413	16,101	0.9602500...	0.9551063...	725.00

Fig.5. ROC Matrix for the Naïve Bayes algorithm

We analyzed the cost matrix for both classification models to compare the results and accuracy obtained.

For the Naive Bayes model there are 16750 instances associated with the value 0 (loan can be granted) of which 94.71% were correctly predicted and 1489 cases of default, corresponding to the value 1, from which the model has correctly predicted a percentage of 100 %



Name: "DM4J$T919943235108_M"

Average Accuracy: 0.9735522388

Overall Accuracy: 0.9514227754

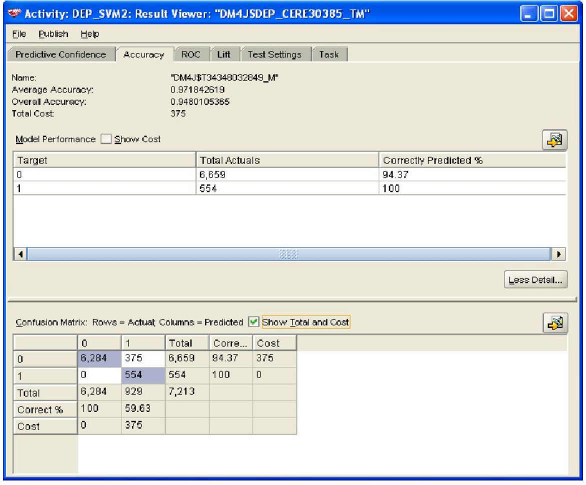
Total Cost: 964.7614

Model Performance □ Show Cost

Target	Total Actuals	Correctly Predicted %
0	16,750	94.71
1	1,489	100

Confusion Matrix: Rows = Actual; Columns = Predicted ☑ Show Total and Cost

	0	1	Total	Corre...	Cost
0	15,864	886	16,750	94.71	964.76
1	0	1,489	1,489	100	0
Total	15,864	2,375	18,239		
Correct %	100	62.69			
Cost	0	964.76			



Name: "DM4J$T34348032849_M"

Average Accuracy: 0.971842619

Overall Accuracy: 0.9480105365

Total Cost: 375

Model Performance □ Show Cost

Target	Total Actuals	Correctly Predicted %
0	6,659	94.37
1	554	100

Confusion Matrix: Rows = Actual; Columns = Predicted ☑ Show Total and Cost

	0	1	Total	Corre...	Cost
0	6,284	375	6,659	94.37	375
1	0	554	554	100	0
Total	6,284	929	7,213		
Correct %	100	59.63			
Cost	0	375			

Fig. 6 –Cost matrix for :

a. Naïve Bayes algorithm

b. SVM algorithm

The total cost is 964.76 u.m. The confusion matrix indicates 0 false-negative predictions when the real value is 1 and 886 false pozitive predictions predictions when the real value is 0.

For the SVM model we used a smaller training set of 6659 instances related to the value 0, out of which 94.37% were correctly predicted and 554 cases related to the value 1 from which the model correctly predicted 100% . The total cost is 375 u.m. The confusion matrix indicates 0 false-negative predictions when the real value is 1 and 375 false-pozitive predictions when the real value is 0.

6. Evaluating the system

To evaluate the system there must be organized a series of working session with the users involved in order to validate the system's functionalities, identified its main evaluation criteria and developed a report on the degree of fulfillment of these criteria. We will present the identified criteria and theri degree of completion below.

1. *Integrating data from multiple sources - the degree of fulfillment: High.* Due to the implementation of the CRM system, data sources are multiple and diverse, being integrated into two functional modules namely the CRM system and data mart. For data integration various techniques and methods are used such as JDBC and bridge connectors, XML/XSD schemas and data warehouses. No data is taken from outside the bank and as such it does not influence the performance of the solution.
2. *Integration Service – the degree of fulfillment: High.* Service integration is done in the Microsoft CRM and SharePoint via plug-in components and Web services
3. *Portal integration – the degree of fulfillment: Average* – Portal integration is only performed for processing document of two types: client and collateral documents. There is no portal structure provided for integrating applications within the bank.
4. *Business process interoperability – the degree of fulfillment: High.* In the system we ensure the implementation of all business processes identified in the analysis phase through software components described previously. The processes are fully automated and run fluently and seamlessly and the interconnection of heterogeneous platforms.
5. *Flexibility – degree of fulfillment: High.* From both a technical and functional perspective the solution is flexible and can be easily adapted by adding new features such as integration with an ERP system or a full organizational portal. It is also The interconnection with objects in the Data Mart is also provided through an organizational data warehouse.
6. *Scalability – degree of fulfillment: High.* Thanks to the CRM, SharePoint portal and Data Mart system the resizing of solution can be achieved according to the changes in the organization without affecting overall system performance.
7. *Maintenance - degree of fulfillment: Average.* The solution was developed using five homogeneous servers in terms of the operating system but heterogeneous in terms of platforms and installed software products. Therefore maintaining the solution may cause problems if not properly monitored and specialists in database administration and management of operating systems are not involved.
8. *Decision support - degree of fulfillment: Average.* The system provides reporting tools for the current operations in the CRM system and business intelligence tools to realize analytical dashboards or reports dedicated to senior managers for analysis of the lending activity on different time periods, different types of products and depending on the customer profile.

9. *Performance – High.* Due to advanced technologies and next-generation database management systems used the solution offers a lower response time. Even during the multidimensional analyzes we can obtain a high performance in developing analytical reports. Also, the data processing is carried out consistently and rapidly, being used a single database to manage current activities.
10. *Friendly interface - High* - From the perspective of the end user using the solution through the use of mobile devices such as laptop, PDA, tablet or smartphone is a welcome feature in the easy conduct of activities. The interfaces present the information both graphically and in spreadsheet form via videoformats and exports can be made to spreadsheets or PDF documents. To use the system users do not need advanced knowledge in IT.

The analysis of these criteria by the managers involved in decision making can conclud that the system's functional and technical requirements are met and the solution is suitable for broad deployment and in other financial and banking institutions.

7. Conclusions

In the case study discussed in the paper we developed three data miming models in order to establish the likelihood of granting a loan, the maximum amount that can be granted and the grouping of customers based on their profile, presenting in detail the model of determining the maximum amount that can be granted. The values so determined will be associated to the loan applications to develop the final decision on granting or not granting the loan.

We have also built a Data Mart based in the entities of the CRM system in order to enable integration with an organizational data warehouse and also multidimensional analysis of current activities. The Data Mart is used for analytical reports which are summarized information on the volume of loans granted in different time intervals, by category, by product and by customer typology.

Finally, we evaluated the proposed solution on a series of criteria such as the degree of integration, interoperability, flexibility, performance, maintenance, scalability and decision support offered.

Acknowledgment: This paper was co-financed from the European Social Fund, through the Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/138907 "Excellence in scientific interdisciplinary research, doctoral and postdoctoral, in the economic, social and medical fields - EXCELIS", coordinator The Bucharest University of Economic Studies

References

[1] S. Chaudhuri, U. Dayal, V. Narasayya, An Overview Of Business Intelligence Technology, *Communications of the ACM* 54 (8): 88–98, 2011.

[2] J. Ranjan - Business intelligence: concepts, components, techniques and benefits, *Journal of Theoretical and Applied Information Technology*, Vol 9, No 1, 2009.

[3] Vlad Diaconita, Procesarea volumelor mari de date folosind HADOOP Yarn, *"Studii si Cercetari de Calcul Economic si Cibernetica Economica",* Nr. Special 1-2, pg. 43-51, ISSN:1843-0112

[4] Vlad Diaconita, Big Data and Machine Learning for Knowledge Management, *Proceedings of the 9th International Conference On Business Excellence*, Economica Printing Press, Bucharest 2014, pp 244-248, ISBN 978-973-709-738-5

[5] O.L Mangasarian. - Linear and non-linear separations of patterns by linear programming, Operations Research, Volume 13, Issue 3, 1965, pg. 444-452

[6] N. Friedman, N, D. Geiger, D, M.Goldszmidt - Bayesian Network Classifiers, *Machine Learning*, Volume 29 (2-3), p. 131.

Alexandra Maria Ioana FLOREA has graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2007 and also from the Faculty of Marketing in 2008. She has a PhD in Economic informatics and since February 2014 she is a lecturer. She teaches Databases, DBMS and Software Packages seminars and courses at the Economic Cybernetics, Statistics and Informatics Faculty. She is co-author of 3 books, has 8 articles published in prestigious journals included in international recognized databases (SCOPUS, Elsevier, EBSCO, ProQuest, or DOAJ) and also 22 papers in the volumes of national and international scientific manifestations, of which 5 are indexed Thomson ISI Web of Science and her fields of interest include integrated information systems, information system analysis and design methodologies and database management systems.

Approaches for parallel data loading and data querying

Vlad DIACONITA
The Bucharest Academy of Economic Studies
diaconita.vlad@ie.ase.ro

This paper aims to bring contributions in data loading and data querying using products from the Apache Hadoop ecosystem. Currently, we talk about Big Data at up to zettabytes scale (10^{21} bytes). Research in this area is usually interdisciplinary combining elements from statistics, system integration, parallel processing and cloud computing.

Keywords: *Hadoop, loading data, Sqoop, Tez*

1 Introduction

Contributing to this growth in data volume are people interacting with different applications as computerization is incorporated in many appliances such as watches, sport belts, cars, airplanes or even in fridges and toasters. This lead to the expansion of Internet of Things (IoT). As shown in [2] The Internet of Things, also called the Internet of Everything or the Industrial Internet, is a new technology paradigm envisioned as a global network of machines and devices capable of interacting with each other. The real value of the IoT for enterprises can be fully realized when connected devices are able to communicate with each other and integrate with vendor-managed inventory systems, customer support systems, business intelligence applications, and business analytics. In [3] it's shown that by 2020 there will be 26 billion devices communicating in IoT.

The classical characteristics of Big Data are volume, velocity, and variety as discussed in many works such as [1] or [6]. As shown in [4] although the three Vs are used to define and differentiate consumer Big Data from large-scale data sets, two more Vs are critical in collecting, analyzing, and extracting insights from Big Data: veracity and value. These two Vs underline the importance of data quality and usefulness. Summarizing, in [4] it's shown that compared to traditional data, the features of big data can be characterized by five Vs, namely, huge Volume, high Velocity, high Variety, low Veracity, and high Value. The authors argue that the real challenges are not only in the vast amount of data but center around the diversified data types (Variety), timely response requirements (Velocity), and uncertainties in the data (Veracity).

Not all data is accurate so particular attention must be given to eliminating faulted or irrelevant data so real value can be extracted from relevant data. Also, valuable insights can be missed when data is simplified so it can fit a model.

By the means of advanced statistical modeling, optimization techniques, and data mining, organizations have at hand the right solutions to quickly mine for value in their data, being it structured, semi-structured or unstructured.

Modern visualization models work well with Big Data approaches. Chord diagrams that can show directed relationships among a group of entities, Voronoi diagrams can be used to display the most similar objects and Parallel Sets to exhibit intuitively and explore multi-dimensional categorical data.

2. Big Data in different areas

It's a field that started with the web search industry, but it's now touching various industries that are using devices which are generating logs (cars, airplanes, buildings, wearable devices, medical devices even home appliances) or that store digitalized records of people or companies (governments, insurance companies, banks, stock markets).

Big Data it's with no doubt an area that raises a lot of eyebrows in regards to privacy, but it’s a field that cannot be ignored and that is already shaping our future. Hidden in large volumes of data are valuable information, patterns, which in the past could be harder identified and understood because of the resources needed to extract them by running sophisticated machine learning algorithms. More and more firms attempt to obtain relevant data using modern techniques such as speech analytics tools on top of more “classical” approaches such as social media mining or tracking viewing or purchasing habits. In some cases, there is also access to geographical data that shows the movement of a particular customer. Based on these multitude of data, some businesses try to predict what the level of income of a client is, what are his TV viewing preferences, what is the best holiday destination for a family or even if they are expecting a new member in the family. Big Data also enables a better dynamic prices policies in a multitude of areas such as the hospitality industry. In [13] the authors discuss how large volumes of data can be used in developing climate and energy targets and in [14] how data mining techniques can be utilized in the analysis of KPIs.

Big Data techniques are being used in universities for a better understanding of a student’s profile, identifying patterns to predict and guide the student’s academic performance, plagiarism detection and attracting new students. Using admission data and clustering algorithms (e.g. k-means), we can identify patterns in the way students choose faculties and specializations.

Infrastructure

As shown in [6] and [7] Big data is data too big to be handled and analyzed by traditional database protocols such as SQL.

Many enterprises and institutions are storing data into HDFS and expect to be able to process it in many ways (data mining, real-time SQL type querying, batch processing with or without machine learning algorithms, etc.). As shown in [8] HDFS is a distributed file system designed to run on large clusters of commodity hardware based on Google File System (GFS) usually dedicated to batch processing.

Originally used for web search index MapReduce is the primary programming model and associated implementation for processing and generating large datasets. As shown in [9], in this model the users specify the computation in terms of a map and a reduce function, and the underlying runtime system automatically parallelizes the computation across large-scale clusters of machines, handles machine failures, and schedules inter-machine communication to make efficient use of the network and disks.

As shown in [10], until recently MapReduce was the only programming model in Hadoop. But in 2012 the Hadoop v2.0 was released as a beta version and the YARN resource manager was introduced so that now MapReduce is just one framework that can execute under a YARN-managed cluster (Figure 1). The different parallel computing frameworks and paradigms that can be implemented using Hadoop YARN are encouraged, on the infrastructure side by the faster networks and internet connections, more and more cores on a CPU, larger memories and faster storage using SSD.

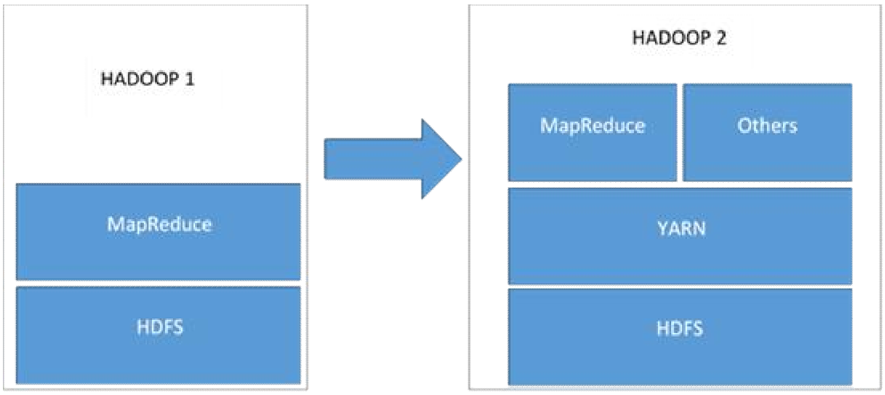


Fig. 5. Hadoop Yarn Architecture, source: adaptation from [15]

The different parallel computing frameworks and paradigms that can be implemented using Hadoop YARN are encouraged, on the infrastructure side by the faster networks and internet connections, more and more cores on a CPU, larger memories and faster storage using SSD.

3. Loading data

Sqoop is an open source Apache project and it is designed to transfer data between Apache Hadoop and other data stores such as relational databases. It has various connectors that can be used with most database and data warehousing systems. In Figure 2 it's shown how an import or export command is being processed.

We loaded data from the products table stored in MySQL to Hive using this command:

% sqoop import --connect jdbc:mysql://localhost/world --username sqoop --password sqoop --table exchange_rates -m 1 --target-dir /user/hdfs/sqoop-mysql-import/exchange_rates

The actual data loading is done using map and reduce tasks as shown in Figure 3.

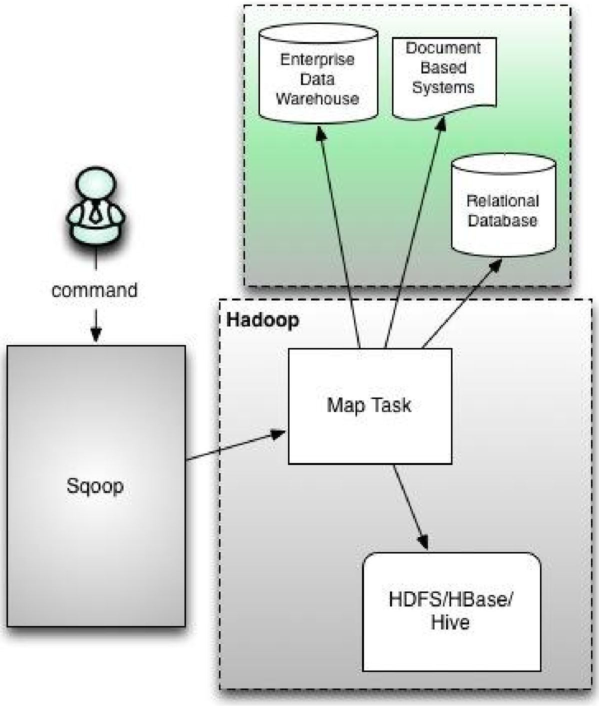
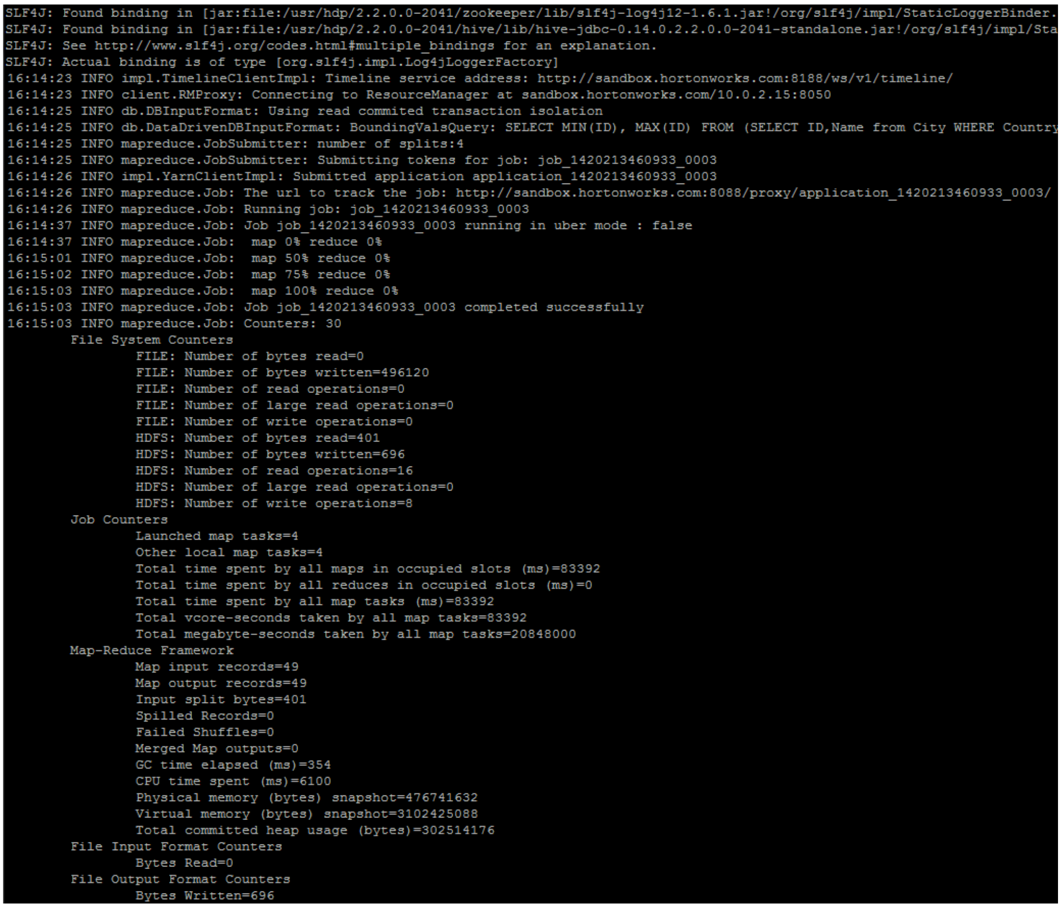


Fig. 6. Apache Sqoop, source: http://hortonworks.com/hadoop/sqoop/



```
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/zookeeper/lib/slf4j-log4j12-1.6.1.jar!/org/slf4j/impl/StaticLoggerBinder.
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/hive/lib/hive-jdbc-0.14.0.2.2.0.0-2041-standalone.jar!/org/slf4j/impl/Sta
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16:14:23 INFO impl.TimelineClientImpl: Timeline service address: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
16:14:23 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
16:14:25 INFO db.DBInputFormat: Using read commited transaction isolation
16:14:25 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(ID), MAX(ID) FROM (SELECT ID,Name from City WHERE Country
16:14:25 INFO mapreduce.JobSubmitter: number of splits:4
16:14:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1420213460933_0003
16:14:26 INFO impl.YarnClientImpl: Submitted application application_1420213460933_0003
16:14:26 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1420213460933_0003/
16:14:26 INFO mapreduce.Job: Running job: job_1420213460933_0003
16:14:37 INFO mapreduce.Job: Job job_1420213460933_0003 running in uber mode : false
16:14:37 INFO mapreduce.Job:  map 0% reduce 0%
16:15:01 INFO mapreduce.Job:  map 50% reduce 0%
16:15:02 INFO mapreduce.Job:  map 75% reduce 0%
16:15:03 INFO mapreduce.Job:  map 100% reduce 0%
16:15:03 INFO mapreduce.Job: Job job_1420213460933_0003 completed successfully
16:15:03 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=496120
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=401
                HDFS: Number of bytes written=696
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Job Counters
                Launched map tasks=4
                Other local map tasks=4
                Total time spent by all maps in occupied slots (ms)=83392
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=83392
                Total vcore-seconds taken by all map tasks=83392
                Total megabyte-seconds taken by all map tasks=20848000
        Map-Reduce Framework
                Map input records=49
                Map output records=49
                Input split bytes=401
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=354
                CPU time spent (ms)=6100
                Physical memory (bytes) snapshot=476741632
                Virtual memory (bytes) snapshot=3102425088
                Total committed heap usage (bytes)=302514176
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=696
```

Fig. 7. Loading data with Apache Sqoop

4. Querying data in Hive

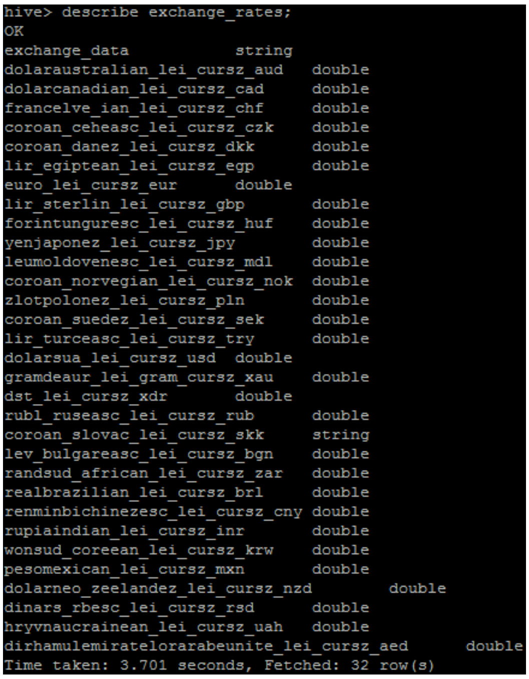
As shown in [11] Hive is an open-source data warehousing solution built on top of Hadoop. Data in Hive is organized in Tables, Partitions and Buckets. It supports primitive data types, nestable collection types and user defined types. Most important, it implements an SQL type querying language: HiveQL.

As shown in [12], Apache Tez is an extensible and scalable framework that improves the MapReduce paradigm by dramatically improving its speed. It's used by Apache Hive, Apache and by third party data access applications developed. It enables data access applications to work with petabytes of data over thousands of nodes. The Apache Tez component library allows developers to create Hadoop applications that integrate natively with Apache Hadoop YARN and perform well within mixed workload clusters.

Vectorization is a feature is used that fetches 1000 rows so the processing speed can be up to 3X faster with the same CPU time.

After we had loaded the data with Sqoop we tried to optimize the processing time using Tez, Query Vectorization and CBO.

We can use the SQL describe command to see the structure of the table that was imported as shown in Figure 4.



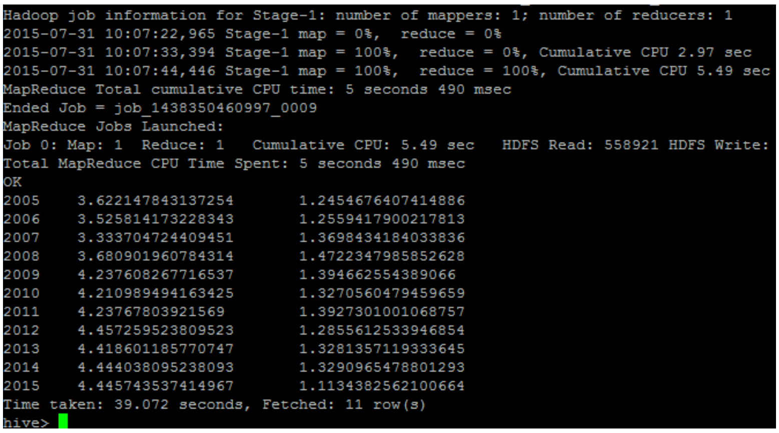
```
hive> describe exchange_rates;
OK
exchange_data               string
dolaraustralian_lei_cursz_aud   double
dolarcanadian_lei_cursz_cad     double
francelve_ian_lei_cursz_chf     double
coroan_ceheasc_lei_cursz_czk    double
coroan_danez_lei_cursz_dkk      double
lir_egiptean_lei_cursz_egp      double
euro_lei_cursz_eur          double
lir_sterlin_lei_cursz_gbp       double
forintunguresc_lei_cursz_huf    double
yenjaponez_lei_cursz_jpy        double
leumoldovenesc_lei_cursz_mdl    double
coroan_norvegian_lei_cursz_nok  double
zlotpolonez_lei_cursz_pln       double
coroan_suedez_lei_cursz_sek     double
lir_turceasc_lei_cursz_try      double
dolarsua_lei_cursz_usd  double
gramdeaur_lei_gram_cursz_xau    double
dst_lei_cursz_xdr           double
rubl_ruseasc_lei_cursz_rub      double
coroan_slovac_lei_cursz_skk     string
lev_bulgareasc_lei_cursz_bgn    double
randsud_african_lei_cursz_zar   double
realbrazilian_lei_cursz_brl     double
renminbichinezesc_lei_cursz_cny double
rupiaindian_lei_cursz_inr       double
wonsud_coreean_lei_cursz_krw    double
pesomexican_lei_cursz_mxn       double
dolarneo_zeelandez_lei_cursz_nzd        double
dinars_rbesc_lei_cursz_rsd      double
hryvnaucrainean_lei_cursz_uah   double
dirhamulemiratelorarabeunite_lei_cursz_aed      double
Time taken: 3.701 seconds, Fetched: 32 row(s)
```

Fig. 4. The exchange_rates table

We will run the same query using different optimization techniques:

hive> set hive.execution.engine=mr;
hive> select substr(exchange_data,-4), avg(euro_lei_cursz_eur),avg(euro_lei_cursz_eur/dolarsua_lei_cursz_usd) from exchange_rates group by substr(exchange_data,-4);

Running with MapReduce it takes 39.072 seconds on a single node cluster as shown in Figure 5.



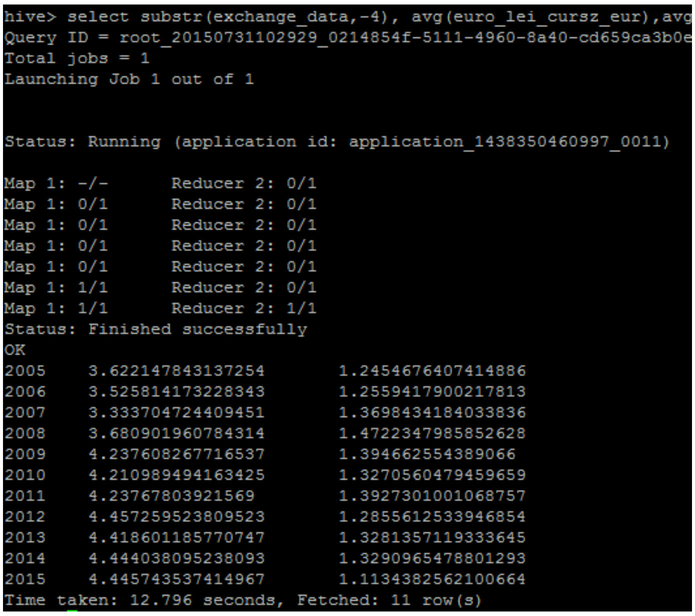
```
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2015-07-31 10:07:22,965 Stage-1 map = 0%,  reduce = 0%
2015-07-31 10:07:33,394 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.97 sec
2015-07-31 10:07:44,446 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.49 sec
MapReduce Total cumulative CPU time: 5 seconds 490 msec
Ended Job = job_1438350460997_0009
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 5.49 sec   HDFS Read: 558921 HDFS Write:
Total MapReduce CPU Time Spent: 5 seconds 490 msec
OK
2005    3.622147843137254       1.2454676407414886
2006    3.525814173228343       1.2559417900217813
2007    3.333704724409451       1.3698434184033836
2008    3.680901960784314       1.4722347985852628
2009    4.237608267716537       1.394662554389066
2010    4.210989494163425       1.3270560479459659
2011    4.23767803921569        1.3927301001068757
2012    4.457259523809523       1.2855612533946854
2013    4.418601185770747       1.3281357119333645
2014    4.444038095238093       1.3290965478801293
2015    4.445743537414967       1.1134382562100664
Time taken: 39.072 seconds, Fetched: 11 row(s)
hive> 
```

Fig. 5. Running the query with MapReduce

Tez activation can be done in Hive with the following command:

hive> set hive.execution.engine=tez;

Running the same query takes at first run with Tez 24.826 seconds. As shown in Figure 5, if we run the query again in the same session it takes only 12.796 seconds to complete because it uses the hot containers previously produced.



```
hive> select substr(exchange_data,-4), avg(euro_lei_cursz_eur),avg
Query ID = root_20150731102929_0214854f-5111-4960-8a40-cd659ca3b0e
Total jobs = 1
Launching Job 1 out of 1


Status: Running (application id: application_1438350460997_0011)

Map 1: -/-      Reducer 2: 0/1
Map 1: 0/1      Reducer 2: 0/1
Map 1: 0/1      Reducer 2: 0/1
Map 1: 0/1      Reducer 2: 0/1
Map 1: 0/1      Reducer 2: 0/1
Map 1: 1/1      Reducer 2: 0/1
Map 1: 1/1      Reducer 2: 1/1
Status: Finished successfully
OK
2005    3.622147843137254       1.2454676407414886
2006    3.525814173228343       1.2559417900217813
2007    3.333704724409451       1.3698434184033836
2008    3.680901960784314       1.4722347985852628
2009    4.237608267716537       1.394662554389066
2010    4.210989494163425       1.3270560479459659
2011    4.23767803921569        1.3927301001068757
2012    4.457259523809523       1.2855612533946854
2013    4.418601185770747       1.3281357119333645
2014    4.444038095238093       1.3290965478801293
2015    4.445743537414967       1.1134382562100664
Time taken: 12.796 seconds, Fetched: 11 row(s)
```

Fig. 5. Running the query with Tez

To use query vectorization we need to create another table:

*hive> create table exchange_rates_orc stored as orc as select * from exchange_rates;*
hive>set hive.vectorized.execution.enabled;
The query is run using the new table:
select substr(exchange_data,-4), avg(euro_lei_cursz_eur),avg(euro_lei_cursz_eur/dolarsua_lei_cursz_usd) from exchange_rates_orc group by substr(exchange_data,-4);

The query time is now of only 10.192 seconds.
Going one step further we can use stats and cost based optimization (CBO) running the following commands:

hive> analyze table exchange_rates compute statistics;
hive> analyze table exchange_rates compute statistics for columns euro_lei_cursz_eur, dolarsua_lei_cursz_usd;
hive> set hive.compute.query.using.stats=true;
hive> set hive.stats.fetch.partition.stats=false;
hive> set hive.cbo.enable=true;
hive> set hive.stats.fetch.column.stats=true;

The query time is of 10.098 seconds. Even better gains can be obtained if we use a much larger dataset than the one we are working with.

Conclusions

In this paper, we discussed the main characteristics of Big Data and we analyzed how data can be imported from relational databases. We also discussed several approaches to optimize parallel data loading and querying by using multiple mappers, Tez, query vectorization and CBO.

Acknowledgements

This paper was co-financed from the European Social Fund, through the Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/138907 "Excellence in scientific interdisciplinary research, doctoral and postdoctoral, in the economic, social and medical fields - EXCELIS", coordinator The Bucharest University of Economic Studies.

References

[1] Emani, C. K., Cullot, N., & Nicolle, C. (2015). Understandable Big Data: A survey. *Computer Science Review*.

[2] Lee, I., & Lee, K. (2015). The Internet of Things (IoT): Applications, investments, and challenges for enterprises. *Business Horizons*.

[3] Gartner. (2014, March 19). *Gartner says the Internet of Things will transform the data center*. Retrieved from http://www.gartner.com/newsroom/id/2684616

[4] Erevelles, S., et al. (2015) , Big Data consumer analytics and the transformation of marketing, Journal of Business Research, http://dx.doi.org/10.1016/j.jbusres.2015.07.001

[5] Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. *Big Data Research*, *2*(2), 59-64.

[6] Davis, K. (2012). *Ethics of Big Data: Balancing risk and innovation*. O'Reilly Media, Inc

[7] Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

[8] White, T. (2012). Hadoop: The Definitive Guide 3rd edition, O'Reilly, 630 p, ISBN 978-1-4493-1152-0

[9] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

[10] Zafar, H., Khan, F. A., Carpenter, B., Shafi, A., & Malik, A. W. (2015). MPJ Express meets YARN: towards Java HPC on Hadoop systems. *Procedia Computer Science*, *51*, 2678-2682.

[11] Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., & Murthy, R. (2009). Hive: a warehousing solution over a map-reduce framework.*Proceedings of the VLDB Endowment*, *2*(2), 1626-1629.

[12] http://hortonworks.com/hadoop/tez/

[13] Florea, A.M.I, *National And International Policies Towards Europe's Climate And Energy Targets Until 2020.,* The Ninth International Conference On Economic Cybernetic Analysis: Positive And Negative Effects Of European Union And Eurozone Enlargement - PONE2014, Bucharest (Romania), 30 October – 1 November, 2014

[14] Florea, A.M.I, *Technologies for the Development of a Decision Support System For Business Process Modelling and Analysis of Key Performance Indicators*, Special Number 1-2/2015, Studii si Cercetari de Calcul Economic si Cibernetica Economica", pp 43-51, ISSN print:0585-7511 ISSN on-line:1843-0112

[15] http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.1.3/bk_using-apache-hadoop/content/yarn_overview.html

Vlad DIACONITA graduated in 2005 the Economic Informatics undergraduate program from The Bucharest Academy of Economic Studies, Romania. Since 2010 he holds a Ph.D. in the domain of Cybernetics and Statistics in Economics.
Since July 2014 is pursuing post-doctoral research financed by EU through the Excelis program with the project: "Distributed analysis of large volumes of data for decision support".
His interests are mainly in the domain of databases, data warehouses, big data, system integration, decision support, cloud computing.
He is a member of INFOREC professional association and a member of IEEE Computer Society.