Process Optimization and Monitoring Along Big Data Value Chain

Dalia Kriksciuniene[1], Virgilijus Sakalauskas[1(✉)], and Balys Kriksciunas[2]

[1] Department of Informatics, Vilnius University, Universiteto Str. 3, Vilnius, Lithuania
{dalia.kriksciuniene, virgilijus.sakalauskas}@khf.vu.lt
[2] Kaunas University of Technology, K. Donelaicio Str. 73, Kaunas, Lithuania
balysk@gmail.com

Abstract. The article deals with Big Data (BD), which is big not only by its volume, but also by velocity or variety - as the combined effect of specific characteristics, create different "portraits" of BD in various domains challenging value extraction from data. Big Data value chain means that the enterprises have to raise skills for dealing with data in all stages of its life cycle: starting from recognizing need to register and store data items, moving forward to their appropriate representation and visualization, processing data with the help of best-fit algorithms, applying methods in order to get insights, finding valuable decisions in uncertain situations, and elaborating tools for control of effectiveness of BD value chain processes. We will follow all the entirety of the processes used for BD monitoring along its value chain and optimize them for extracting highest possible value. The goal of paper is to describe innovative solutions in domain driven process optimization and monitoring along BD value chain. We analyse the specific characteristics of BD by suggested BD portrait concept, its impact for BD analysis along entire value chain, and transferring research results to other research domains. The presented case-study highlights the problems of practical big data value chain implementation.

Keywords: Big data · Artificial intelligence · Simulation · Fuzzy insights · Algorithms · Methods · Conceptual development · Applications in business · Numerical analysis · Optimization · Data mining · Modelling tools

1 Introduction

Big Data term refers to IT, yet the problem of BD is highlighted not only by IT market. As noticed by Gartner, through 2015, 85 % of Fortune 500 organizations will be unable to exploit Big Data for competitive advantage [1]. The European Big Data Value Strategic Research & Innovation Agenda (2014) states that Big Data has the strategic importance for the EU, that Europe must both face and exploit in a structured, aggressive and ambitious way to create value for society, its citizens, and its businesses in all sectors [2].

W. Abramowicz (Ed.): BIS 2015 Workshops, LNBIP 228, pp. 75–86, 2015.
DOI: 10.1007/978-3-319-26762-3_8

BD value in numbers: it is expected that by 2020 there will be more than 16 zettabytes of useful data which implies growth of 236 % per year from 2013 to 2020. The Big Data Value market can be also measured by the revenue that vendors earn from sales of related hardware, software and ICT services. According to IDC the Big Data market is growing six times faster than the overall ICT market, as compound annual growth rate (CAGR) of the Big Data market over the period 2013–2017 will be around 27 %, reaching an overall total of $50 billion. The increased adoption of Big Data will have positive impact on employment, and is expected to result in 3.75 million jobs in the EU by 2017; its impact to various sectors has socio-economic potential far beyond the specific Big Data market [2].

2 BD Dimensions and Value Chain Concept

Although it can be assumed that BD is expressed mostly by amount of data instances captured in various domains, the technological perspective for BD takes into account more specific dimensions. In Gartner IT glossary [1] Big Data is defined as high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. Summarizing [1–6] the main characteristics of BD can be specified by applying the "V"-based nominations. They include:

Volume	as descriptor of size of data set,
Velocity	describing data processing speed within allowed latency or real-time,
Variety	of information sources with different structure,
Variability	presence of data descriptors, such as variable field of the data base, which can be filled or half-empty leading to lack of data in BD settings,
Veracity	expressing the level of trust for the data content, which implies additional efforts for data cleansing, recognizing vicious content,
Value	characteristics, which means the level of efforts which is necessary for increasing value of information which can be extracted from BD analysis

The analysis reveals that specific BD problems are not only related to the adjective "big" but to different computational and technological approaches related to its dimensions as well.

The application of IT methods for Big Data cases are related not only to the effect of each separate dimension, but to their combinations as well. The computational methods effectively dealing with large amount of data (Volume) can be not suitable if high processing speed is required (Velocity) or the information source is not trustful (Veracity). The additional activities or methods will be required to successful application of IT tools and models.

The 6 V characteristics of BD are heavily dependent on domain areas emanating the sources of BD and the use-cases for decision making are emanating from. The requirements for precision, riskiness, trustfulness of the results of BD analysis can be very different in banking, medicine or in exploring customer behaviour and opinion mining. Same applies to the variety of sources of information which are used for analysis in these domain areas.

By summarizing existence and content of BD characteristics we suggest elaborating model which can be nominated as "BD portrait" and which could enable selection of analytic models leading to effective value extraction form data. Due to the unclear boundaries of problem areas and data sources used for value extraction, the fuzziness of BD portrait has to be admitted.

Data handling with the goal of efficient value extraction is not unified process, both in businesses and in science it is treated and handled in fragmented way with the unique and expensive outcome with low possibility of transferring to other domains. In order to ensure a coherent use of data, the concept of Data Value Chain was accentuated for facilitate cooperation of all stakeholders as European Public Private Partnership in Big Data (ICT 2013 event in Vilnius on 7 November 2013, speech by Kroes, 2013). Value chain is specified as steps of:

- Data generation, acquisition;
- Data analysis, processing;
- Data storage, curating;
- Data visualisation, usage& services;

The security, data protection, privacy, trust problems are general along the entire value chain (Fig. 1).

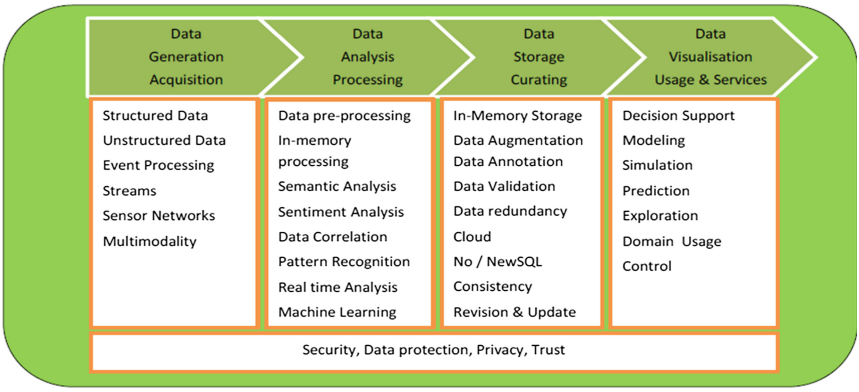


Fig. 1. Big Data Ecosystem along the Value Chain [2].

Extracting value as the result of information analysis, its optimization and enhancement of effectiveness is one of the ultimate goals of our research.

According to [2] the implementation of value chain rely on various factors: integrating technologies for data, developing appropriate data management skills in industry and academia; elaborating technologies, applications and solutions for the creation of value from Big Data to industry and the public sector, optimization of architectures for real-time analytics of both data at rest and in motion enabling data-driven decision-making, integrating BD services into private and public decision making systems such as ERP and marketing systems for optimising the functioning of

existing industries and potentially establishing entirely new business models, within cross-organisational, cross-sector, and cross-lingual innovation environments.

The main technical aspects include:

(1) data (making data sets and assets accessible, ensuring variety of domains, and including industrial, private and open data sources, ensuring their availability, integrity, and confidentiality).
(2) technology related to Big Data Value extraction, data-driven applications and business opportunities along the data value chain.
(3) application, in the demo and trial forms by allowing testing of technologies, applications, and business models, able to provide early insights on potential issues and will help to avoid failures in the later stages of commercial deployments.

Existing technologies of BD accessibility cover tools used in the initial stages of BD value chain that have been developed to collect, store, analyse, process, and visualize huge amounts of data. They can be used in various modes: for equipment of in-house solutions, used as cloud service or supercomputing networks. Many industrial solutions for storing and analysing BD are widely spread due to their performance and open access. It is important to notice that BD which has low volume but is big due to other characteristics of 6 V, such as variability or veracity do not require technologies related to massive or real-time data processing. Therefore advanced classical technologies (Matlab, SPSS, SAS, Statistica) and innovative BD-oriented solutions coexist in the market and are used for analytic purposes simultaneously. The advanced specialized Big data analytic technologies can be classified to ad hoc queries and analysis (e.g. Drill), batch processing (Hadoop, Hive, Pig), online analytical processing OLTP (Cassandra, HBase), stream processing (Storm, S4). The main concepts are further explained.

The Big data analytic tools have several technologies which have already taken strong place in the market. The Hadoop, MapReduce, Hive, Pig, Mahout, Cassandra are among them. Hadoop is an open source project framework that can store and process the unstructured data (MapReduce) in a cluster of computers (grid).

The MapReduce is a parallel computational batch processing frame for Hadoop where jobs are mostly written in Java; here the mapper function means that the larger problem is broken into smaller pieces of work, then the entire workload is distributed across the grid for simultaneous processing and followed by reducer function as a master job which collects all the interim results and combines them into the outcome of required analysis. Hive and Pig are most common for task formulation for big data analysis in MapReduce environments.

Mahout is an Apache project to create library of scalable, machine-learning algorithms for Hadoop mostly implemented in MapReduce. The data structures for optimized storage and retrieval serve as basis of new tools as well. Columnar database is a structure used for storing and optimizing of data by columns, it is particularly effective for analytic processing.

HBase is a distributed, columnar NoSQL database. Cassandra – open-source columnar database managed by Apache Software Foundation. Drill- is one of the most popular systems for interactive ad-hoc queries and analysis, applied for querying

different large-scale data sources. The advantage of the system is low latency of queries. The Drill performs as query layer, where the query written in human-readable syntax is transformed firstly into logical plan (platform-independent mode), then into physical plan (platform-specific way). The Drill executes queries for data sources stored in in Cassandra, HBase, MongoDB.

Main Big data computing platforms include clusters or grids, massive parallel processing (MPP), high performance computing (HPC). It refers to devices designed for high speed floating point processing, in-memory with some disc parallelization.

The performance of data processing technologies highly depend on data storage models. They can vary from Hadoop file system, called HDFS data, which serves as storage mechanism for Hadoop, to in-house solutions for data archives and cloud-general term used to refer to any computing resources- software, hardware or service-that is delivered as a service over a network [2, 7–9]. UVILNIUS HPC provides technologies completely in-line to the state-of the art, NoSQl, Hadoop, BigTable, S3, Open Nebula.

BD analytic tools have broad spectrum as related to the achieved value of extracted information from raw data. The aggregation, visualisation capabilities are followed by models and algorithms of increasing complexity from simple aggregation towards high level of insights, rules, decisions and optimized decision making procedures. The state of the art in BD analytic joins power of classical econometric, statistical methods followed by artificial intelligence and hybrid methods, and also specific BD related analytic approaches.

The goal of creating efficient algorithms for High-performance Big Data mining on parallel, distributed, and emerging platforms meet complex problems, deal with data with high dimensionality, large volumes, heterogeneous structure, multiple sources, various types of noise and outliers, redundant as well as incomplete records, and many other properties that make it very hard to analyse. As BD triggered a change of paradigm in data acquisition, storage, and fundamental analytics (business intelligence), a similar shift is required for advanced BD analytics and mining. Data mining (DM) and machine learning are efficient high–level approaches that can be used to address these objectives. Emerging hardware platforms represent an opportunity for such activities. Together with advanced storage (memory, permanent storage) and communication (interconnect architecture, wireless technology) systems, new types of massively parallel processing elements (multiprocessors, accelerators, GPUs) with affordable purchase, operating, and maintenance costs are readily available for BD processing.

Specialized BD mining algorithms, on the other hand, still need to be re–designed with respect to both, specific challenges of BD and architectures of emerging platforms. Hybrid approaches, based on the combination of two or more fundamental algorithms, are nowadays routinely used to solve data processing and machine learning tasks. The use of different high-performance computing methods together with advanced computing paradigms will allow efficient tackling of some of the challenges of BD.

It can be concluded that the state of the art in the areas analysed within BD value chain concept is best developed in the initial stages of “Data generation, acquisition”, “Data storage, curating”, “Data visualisation, usage & services”. The stage of “Data analysis, processing” is best served with BD-oriented algorithms and tools related to

aggregating. The tools for analysis of knowledge value, its performance evaluation and optimization is not the scope of largest providers of BD management tools. This part of BD value chain is not standardized, and is performed in specific business projects in unique (and most expensive) way.

The analysis of state of the art reveals a gap for research and implementation of solutions for the extended "Knowledge optimization" value chain step, suggested and to be solved by methodology of this proposal.

3 Methodological Approaches of Research and Innovation in BD

The proposed methodology can be expressed by main innovative challenges addressing gaps of state-of-the-art:

- The proposed model of "BD portrait" enables to analyse diverse cross-sector business use-cases available. It solves existing gap of fragmented and unique solutions for BD cases processed by separate organizations.
- The proposed model of BD value chain extension by "Knowledge value optimization" step solves existing gap in the BD research area and industry. This model can be fulfilled by high level research capacities, as it requires design and research of advanced artificial intelligence models for evaluating performance criteria, applying optimization methods.
- The visibility of experimental results both in industrial circles, business partner community and research publications enables transfer of solutions to other BD settings and new use-cases. Methods for evaluating efficiency of the applied models and the innovative solutions for BD analytics and process optimization along each stage of BD value chain can be applied for testing validity of the models.

The methodological approach of the research activities is mapped in Fig. 2.

The workflow of main Big Data value chain model processes is shown on Fig. 3. The all actions of this cycle is necessary and should be responsibly done to achieve the results and calculate BD value chain.

The process wheel on Fig. 3 shows all necessary steps for implementation in practice the big data value chain and optimisation of processes included in the network connecting for join work industrial and knowledge sectors.

4 Experimental Validations

In this chapter we present experimental research of BD value chain model applied for the business case of online portal 000webhost.com. The customer database of the online portal is used for BD value chain performance analysis.

000webhost.com is one of the largest portals providing free and paid services of hosting online content. According to the statistical information of www.alexa.com the 000webhost.com portal is among 2000 most visited pages with the 2.5 billion total

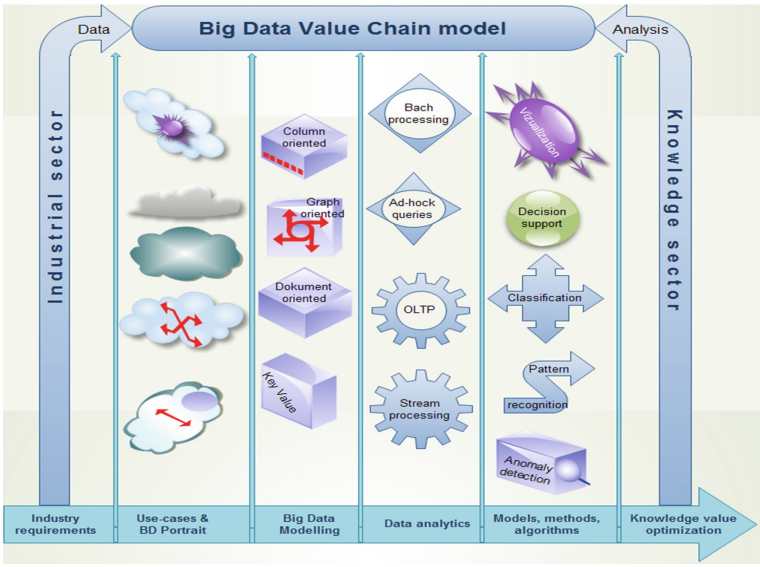


Fig. 2. Knowledge exchange network for implementing Big Data Value chain model

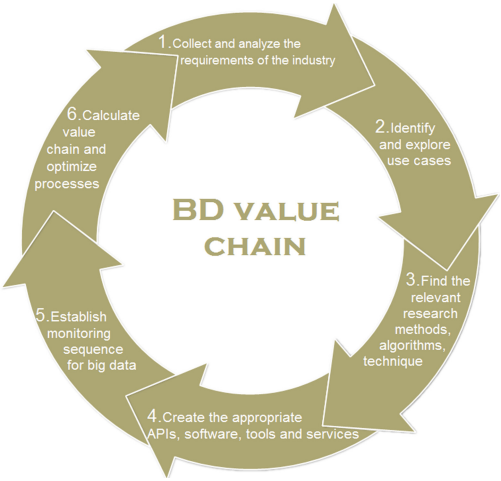


Fig. 3. The process wheel for implementing Big Data Value chain model

numbers of customers and 60000 daily numbers of visitors, who generate high number of instances which use all types of services and resources of hosting environment.

The main problem of portal management is related to number of customers which are registered by using automatic scripts for placing abuse content. These operations hinder performance of hosting services and reduce trust of the preferred customers. The registration form for 000webhost.com is constantly updated by adding various fields

for testing customer status, but it prolonged the registration process, and also led to only short-term improvement due to swift reaction of automated script generators. The proposed solution was to design two-stage process: short registration phase with minimal number of control operations and complex process of customer verification within 24 h after registration.

The verification of customers enables to filter customers for their further processing. The preferred customers are analysed for designing attractive incentives in order to strengthen their loyalty. The abuse customers are assigned the suspended status. The validation module not only eliminates harmful customers and their accounts, it also generated rules for recognizing bad customer cases, which are shifted to the library of rules for using them in the first (registration phase) stage. The learning cycle of the proposed system improves performance of the model and increases its sensitivity to constantly emerging new patterns of automated algorithms if abuse customer generators. The main requirements for validation module include its automated functioning, applying both deterministic and stochastic methods for the customer data existing in all information sources available for the hosting company. The process covers main steps of the value chain including analysis, monitoring and optimization:

1. Collecting customer data, parameters derived during customer history, behaviour pattern characteristics and other estimators in the formats suitable for their processing by intelligent methods.
2. Specification of rules for classifying existing customers. The rules are registered in PHP and MySQL programming languages, oriented to rapid performance speed, not overloading server by high complexity queries.
3. Integration of modules (PHP and MySQL format) to the monitoring system of 000webhost.com portal.
4. Optimization of customer verification process by applying indicators and constraints set for big data value chain model for calibrating optimal set of rules.

4.1 Data Collection

Essential part of customer data is collected in 000webhost database in two forms: the real-time primary data and the secondary - accumulated data consisting of the derived patterns characterizing history of customer activities.

The primary information consists of data inserted by customer in the registration window: name, password sequence, IP address, and registration time in UNIX time format, domain name and type and others.

The secondary data is derived by the methods suggested by experts based on their experience and analytic processing results. The secondary data consists of textual and numeric variables acquiring continuous, categorical, binary, aggregated values. Some variables are designed by using expert rules, which describe some patterns observed in control sequences of customer information: matching fragments, regular sequences of symbols occurring in passwords, domain names or other pieces of information which could indicate activities of automated algorithms and carry risk of abuse behaviour.

Table 1. Expert rules derived on the basis of secondary information

Probabilistic evaluation of domain extension (.tk, .com, .net). Obtained by estimating domain extension frequencies in the entire customer data base
Subdomain name length
Structure and symbol sequence of the subdomain name (e.g. regular repetition of vowels and consonant, patterns of number usage)
The extent of matching attributes: name / email / subdomain
Matching of attribute length name / email / subdomain
Correspondence of IP address and access country
Number of accounts acquired by the customer
Attribute indicating if the customer has written messages *(ticket)* concerning system disorders
Attribute indicating if the IP was already used for registering other accounts
Cases of CANCELED_ABUSE status assignment history: it is evaluated as relative weight by number of cases
Presence of current status CANCELED_ABUSE (nullification of validity)
Definition of parameter values, their interval ranges and median values according to frequencies of registered data types
Aggregated and proxy parameters, which are used for defining probability of risk that the account can be classified as CANCELED_ABUSE

In Table 1 the examples of expert rules created from secondary information are presented.

During creation of the secondary database the Big Data problems are concerned:

- The volume problem is the most urgent, as adding derived variables increases the customer data base (which is already of Big Data scale by volume).
- The variability problem emerges, as the number of expert-knowledge-based and computed variables is high and tends to increase, due to changing behaviour of customers of the webhosting portal. The behaviour changes are especially urgent in the cases of variety of automated scripts and their modifications in the abuse customer cases.
- The problem of the viability parameter, which is suggested in this article as urgent, because during the process of formalizing expert observations applied for creating new variables and meaningful rules it is not clear which variables have cause-effect relationships. The importance and causality of the variables can be detected and researched only by analysing longitudinal historical data. The effective selection of most important variables influencing customer validity procedures is very important; many of existing variables could be omitted if appropriate methods for measuring their influence were available.

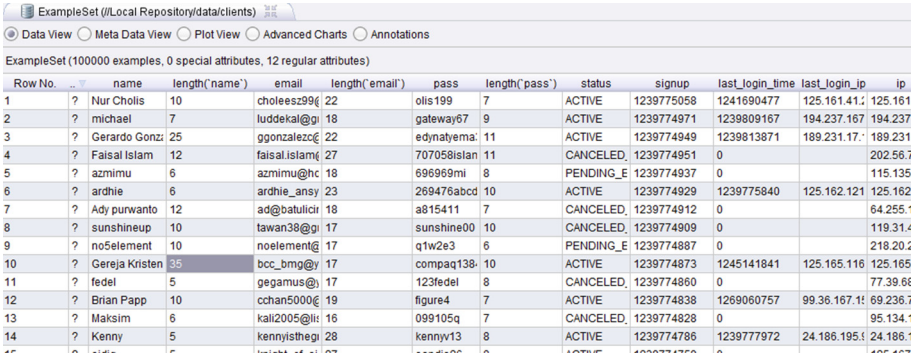
Effective recognition of characteristics enables to reduce duration of customer registration process and improve validation precision, bet also reduce system workload and cases of unfulfilled requests, system overloads, reduce requirement for hardware, storage, costs incurred for registering customer domain names in free service are, and also reduce risk for lawsuits due to abuse activities in the web portal.

4.2 Specification of Rules

At first we had done the 000webhost.com database copy, which is carried out further analysis. Local computer was running Web Server with MySQL database support module.

The data loaded into RapidMiner program should not be redundant. Some fields in the database obviously not represent a client in nature (customer generated access code), so we don't import them. For data format selection we will use the MySQL database queries. The fragment of selected data is presented in Table 2.

Table 2. The fragment of imported in RapidMiner data.


ExampleSet (//Local Repository/data/clients)

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (100000 examples, 0 special attributes, 12 regular attributes)

Row No.	..	name	length(`name`)	email	length(`email`)	pass	length(`pass`)	status	signup	last_login_time	last_login_ip	ip
1	?	Nur Cholis	10	choleesz99(	22	olis199	7	ACTIVE	1239775058	1241690477	125.161.41.2	125.161
2	?	michael	7	luddekal@gı	18	gateway67	9	ACTIVE	1239774971	1239809167	194.237.167	194.237
3	?	Gerardo Gonz	25	ggonzalezc@	22	edynatyema:	11	ACTIVE	1239774949	1239813871	189.231.17.'	189.231
4	?	Faisal Islam	12	faisal.islam(	27	707058islan	11	CANCELED_	1239774951	0		202.56.7
5	?	azmimu	6	azmimu@hc	18	696969mi	8	PENDING_E	1239774937	0		115.135
6	?	ardhie	6	ardhie_ansy	23	269476abcd	10	ACTIVE	1239774929	1239775840	125.162.121	125.162
7	?	Ady purwanto	12	ad@batulicir	18	a815411	7	CANCELED_	1239774912	0		64.255.1
8	?	sunshineup	10	tawan38@gı	17	sunshine00	10	CANCELED_	1239774909	0		119.31.4
9	?	no5element	10	noelement@	17	q1w2e3	6	PENDING_E	1239774887	0		218.20.2
10	?	Gereja Kristen	35	bcc_bmg@y	17	compaq138-	10	ACTIVE	1239774873	1245141841	125.165.116	125.165
11	?	fedel	5	gegamus@y	17	123fedel	8	CANCELED_	1239774860	0		77.39.68
12	?	Brian Papp	10	cchan5000@	19	figure4	7	ACTIVE	1239774838	1269060757	99.36.167.15	69.236.7
13	?	Maksim	6	kali2005@lis	16	099105q	7	CANCELED_	1239774828	0		95.134.1
14	?	Kenny	5	kennyistheg	28	kennyv13	8	ACTIVE	1239774786	1239777972	24.186.195.9	24.186.1

Expert rules were developed relying on customer classification by data attributes similarities. We found links between certain customer parameters, which define the exact nature of the customer.

The expert rules was created using PHP language and integrated into 000webhost.com registration form. The program code let us filter the database records, which satisfy the recognized rule. Expert Rules reliability coefficient is calculated by the following formula: $\beta = m/n$. Here n - number of all records, m - the number of records satisfying the rule. As an example, we stated that the number of don't activated accounts per day not exceed 20–30, and this expert rule validity is $\sim$ 98 %.

In total we have programmed and adapted 17 expert rules to the customer database. They are validated automatically - after filling the customer account data the expert rules are validated and in case it is found conformity to expert rule, client gets error message. The message don't disclose in details the expert rule content to the client, as then the customer can easily adjust the automatic generation of scripts and build unauthorized accounts.

4.3 The Analytical Findings

A large volume of data had a significant impact on the data analysis. Due to the long processing time of some methods the practical their implementations become complicated. For this purpose we use the data minimization. This problem is partly solved by SOM method, but in order to maintain the desired validation accuracy (> 95 %), it is recommended to apply more efficient methods.

In this paper, data minimization is understood as one of big data value chain. At this phase of application of intelligent algorithms the records are more precisely identified and filtered to get the more sophisticated authentication methods. By using k-Nearest Neighbors (k-NN) method analysis we found the values of parameters, which allow the fast enough calculations with the desired accuracy. We get final k-NN model accuracy 95.46 %. Also we identify well-functioning expert rules, which by using k-NN model validate the accuracy of client registration form with 99.62 % reliability. These rules are integrated in the BD value chain algorithm. Increasing number of database records, let us increase the effectiveness of existing expert rules and create new rules to improve the identification and assessment of client registration accuracy.

5 Conclusions

Research related to the BD sphere experiences change of scientific research scenario: from limited pilot experiments with the trial data set in small settings of (sometimes) simulated data it is shifting to the real settings which, first of all, involve condition of experimenting with big data of real scope. BD became the inseparable part and unreplaceable object of the research for exploring algorithms for detecting relationships and forecasting based on data.

Technological environment of the research and real-business problem solving mainly consists of storage & search, analytics and forecasting.

The tools for storage & search are quite developed and standardized in many cases (such as Hadoop, Cassandra). The research methods are heavily based on classical algorithm testing and their modifications along to the characteristics of business use-cases. They include statistics, data mining, and visualization methods. The forecasting problems are addressed in least standardized way, as various methods of soft, intelligent hybrid computing, which are applied for the decisions in the volatile business environment.

In addition to the specification of the business problem the profile of BD has to be explored, as it brings in requirements for analytical and knowledge extraction phases of the research.

The presented case-study let us to maintain that practical implementation of big data value chain let us achieve more valuable results than ignoring this attitude.

References

1. Gartner (2012). http://www.gartner.com/it-glossary/big-data/
2. European big data value partnership. European big data value cPPP, strategic research and innovation agenda, version 0.99 (2014)
3. Forrester (2012). https://www.forrester.com/Big-Data
4. IBM (2015). http://www-01.ibm.com/software/data/bigdata/
5. Hurwitz & Associates, Halper, F.: Four vendor views on big data and big data analytics (2012). https://fbhalper.wordpress.com/2012/01/30/four-vendor-views-on-big-data-and-big-data-analytics-ibm/
6. What is big data, IBM. http://www-01.ibm.com/software/in/data/bigdata/
7. Gandomi, A., Haider, M.: Beyond the hype: big data concepts, methods and analytics. Int. J. Inf. Manage. **35**, 137–144 (2015). Elsevier
8. Forbes, F.D.: The big data landscape (2012). http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/
9. The Apache Cassandra project. http://cassandra.apache.org/