Journal of Theoretical and Applied Computer Science
ISSN 2299-2634 (printed), 2300-5653 (online)
Vol. 10, No. 2, 2016, pp. 41-52
http://www.jtacs.org

Case Studies on Big Data

Larysa Globa[1], Ievgeniia Svetsynska[1], Andriy Luntovskyy[2]

[1]Institute of Telecommunication Systems, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
[2]BA Dresden University of Cooperative Education, Saxon Academy of Studies, Germany

lgloba@its.kpi.ua, Andriy.Luntovskyy@ba-dresden.de

Abstract: *The main idea of the Data Mining (DM) is nowadays as follows: overcoming of the Big Data problematics is possible under use of "data compression" via their transformation into (fuzzy) knowledge. The "heavy-weighting approaches" involving precise analytical techniques and expensive specialized software are used for this aim. On the other hand, there is the opportunity to solve the Big Data problem under use of some "light-weighting approaches" based on agility: freeware, multipurpose techniques, minimal challenges on the personnel training and competencies! The paper examines the techniques and case studies on the both topics. The "heavy-weighting approaches" (ontologies, knowledge bases, fuzzy logic and fuzzy knowledge bases) are compared to light-weighting one. The existing reference solutions are discussed.*

Keywords: *Big Data, Ontologies, Fuzzy Knowledge Base, Heavy-Weighting and Light-Weighting Approaches*

1. Motivation

"Big Data" is a term for the some current and medium-term data sets. Huge as well as heterogeneous data volumes (approx. 100PB to 100EByte) are nowadays so immobile that their access and management becomes only very efficient last time without use of special techniques [1].

Big Data are nowadays too complex for their efficient processing with classical manual methods for data structuring. Herewith a recognition of the problem is given by one of the world's leading research and consulting company as very serious: the company Gartner Inc. notes "the problem of Big Data as one of the most important trends of IT-infrastructure development along with virtualization and energy efficiency of IT" [2].

Under use of conventional storage, retrieval and analysis methods (just like DB, data warehouses, tables and sheets, formatted texts, csv-data, web hypertexts, XML documents, graphics, audio and video sources etc.) "Big Data" become faster suitable for nothing and inadequate.

To solve this problem the computational techniques of so called "Data Mining" i.a. with ontologies, data pattern recognition, fuzzy logic etc. are used. These techniques can disassemble the complexity and heterogeneity of Big Data: via compression via fuzzy logic and knowledge bases is required [3, 4].

1.1. The sources of Big Data

Based on the book [4]: "The most outstanding Big Data sources are such modern technologies like GIS, parallel clusters and grids, semantic and social networks Web2.0, cloud computing, as well as intelligent Internet of Things.

The accumulation of Big Data is now typical for trading and marketing, electronic payments, process automation, for international justice and criminology, pharmaceutical and advertising industry etc. A large number of scientific and research institutes, organizations, universities accumulate, store and compute large volumes of technical and scientific information: industrial, home and automotive automation, patient data in healthcare domain, M2M, Business Analytics, research and pharma experimental data etc.

Often such large information amount is not structured, that is characterized with extra-proportional complexity of information management, a significant increase in network traffic and via heterogeneity of geographically distributed data replicates within multiple computing nodes.

5G will be surely actively involved to Big Data acquisition and processing" [4].

More than 50 billion sensors will be used in 5G by year 2020 to upload information on: therefore how we interact with the things surrounding us or in us?!

One of most exciting further topics on Big Data discussion is the exponentially growing use of up-to-date crypto currencies (Bitcoin, Ethereum, Ripple, Litecoin, Peercoin, NXT, Namecoin) as well as of world-wide Block Chain database.

1.2. Data Mining

The mostly common Data Mining tasks are separated into descriptive and predictive as follows [1, 3, 5]:

- Classification [Predictive],
- Clustering [Descriptive],
- Association Rule Discovery [Descriptive],
- Sequential Pattern Discovery [Descriptive], Regression [Predictive],
- Deviation Detection [Predictive].

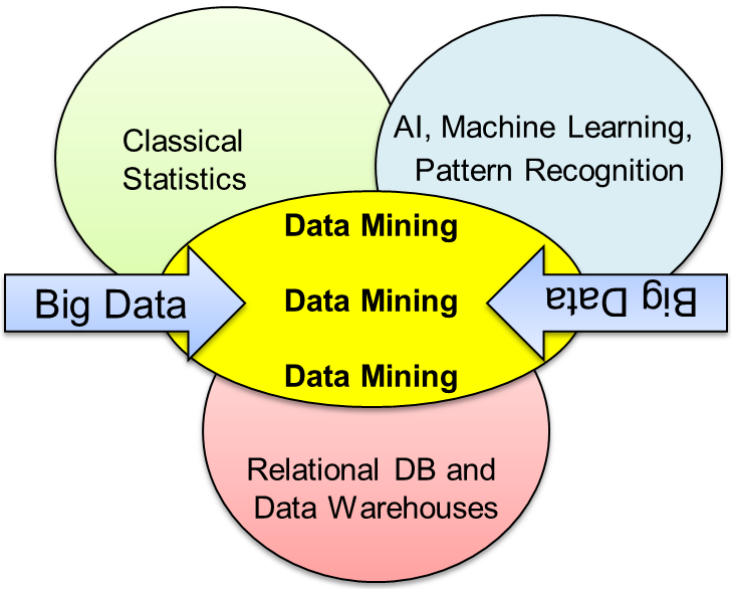

Figure 1. DM computational techniques [1, 3, 5]

Common data mining techniques are: artificial neural networks, decision trees, genetic algorithms, nearest neighbor method, and rule induction (Fig. 1). Additionally DM uses computational techniques based on advanced statistics and pattern recognition methods. Non-trivial extraction of hidden information and knowledge from large databases or data warehouses is an important challenge.

1.3. Own „6V-Paradigm" for Big Data

The most recent properties of Big Data can be formulated as a 6V-Paradigm (cp. 5V concept of Mehmet Ulima) [1]. The mentioned V-factors are not undependable however are interlocked with each other (Fig. 2). These V-factors are as follows:

1. Volume: the amount of generated and stored data is moving from 100PB up to 100EByte nowadays.
2. Velocity: it means the frequency and speed of data aquisition and processing (from slow batch techniques to real time with strong limitation on reaction time or latencies).
3. Variety: multiple codec types of multimedia data. This helps people who analyze it to effectively use the resulting insight.
4. Veracity: The quality of captured data can vary greatly, affecting accurate analysis.
5. Violation: The discussed data are often scattered, without any clustering or any structuring.
6. Value: Inconsistency of the data sets can negatively affect the processes of data handling and management.

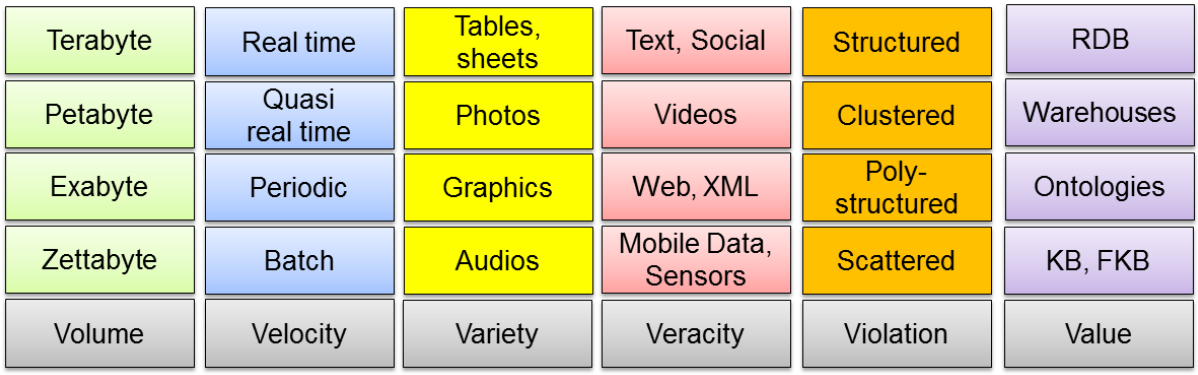

Figure 2. Own 6V-Paradigm

Unfortunately the depicted V-factors for Big Data are growing faster than the performance of their analysis (in Mbyte/s or in GFLOPS) via classical computational techniques [1 - 6].

2. Heavy-Weighting Approaches

The following own classification of the methods and approaches as well as a technology stack aimed to overcoming of the Big Data problematics is depicted in Fig. 3.

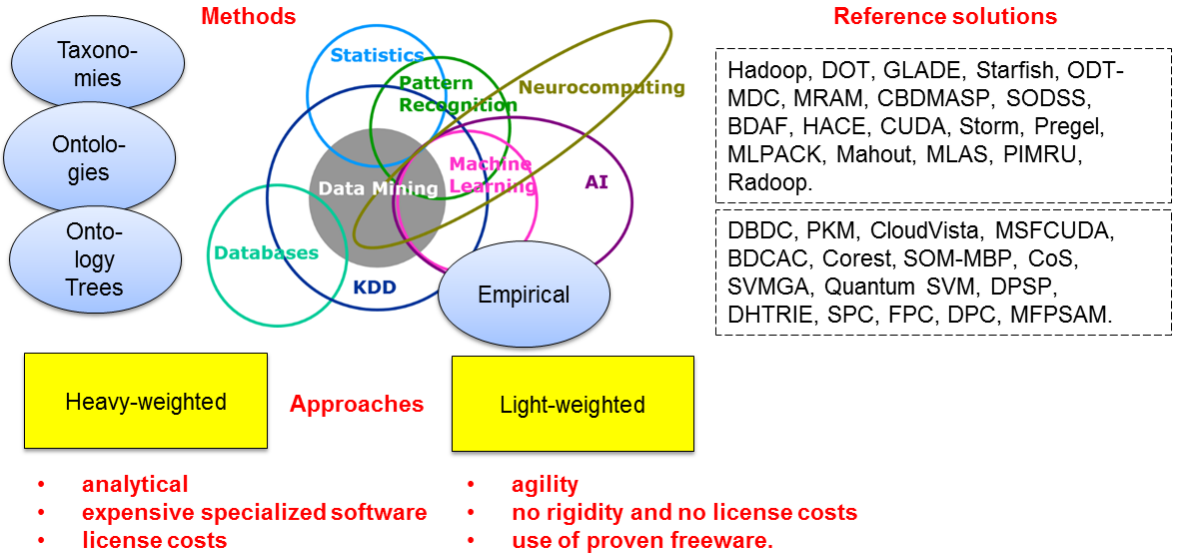

Figure 3. Methods and approaches aimed to overcoming of Big Data: based on [5]

In contrast to "data" and "information" under knowledge is usually understood a set of information (facts, theories and rules) available to persons or groups, which are characterized by the greatest possible degree of certainty, so that on their validity or truth can be trusted (refer Fig. 4).

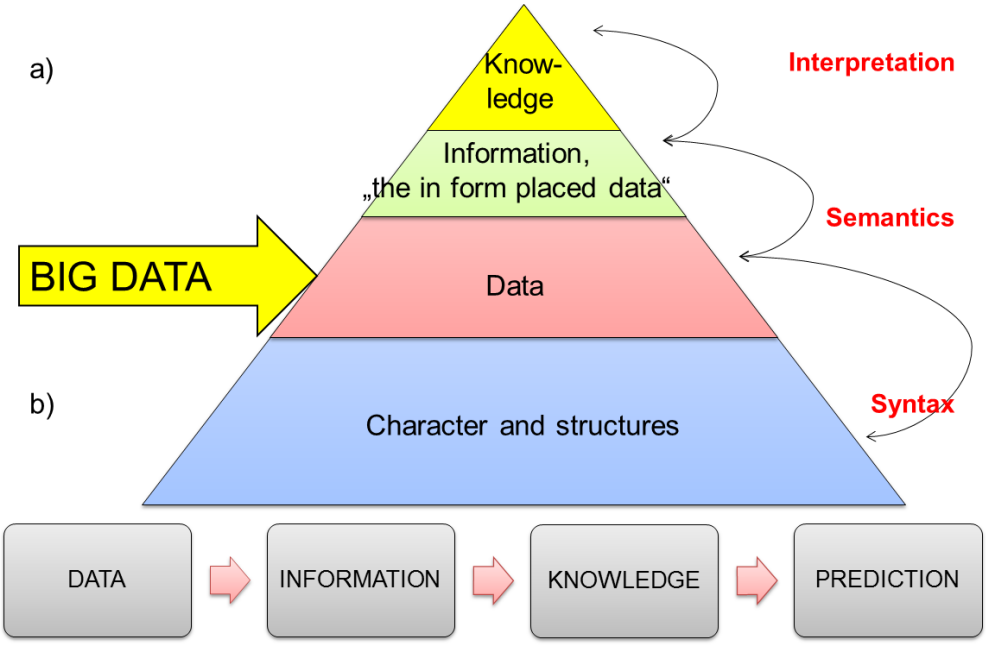

Figure 4. (a) Knowledge Pyramid; (b) Data Mining Workflow

Big Data can however be processed via so called Knowledge Discovery in Databases (KDD). The KDD includes also certain preliminary steps as follows: 1 - acquisition, collection, 2 - pre-processing, clustering and pattern evaluation, 3 - compression based on fuzzy logic and knowledge bases, as well as the real analysis (data mining) with processing and prediction [6 - 9].

The Knowledge Bases (KB) include the following kinds of entries: general knowledge as well as knowledge about specific subject areas and processes.

Can the ontologies be useful for Big Data too? Surely [5, 6, 10, 11]. The deployment of ontologies and trees of ontologies [11] can solve some of such V-problems which were identified for Big Data. Ontologies are usually linguistically constituted and formally ordered representations of a set of entities (conceptualities) and the relationships existing between them in a particular subject area, for example by OWL (based on XML). The ontologies are used to exchange knowledge in a digitalized and formal way (Fig. 5) between applications and services (e.g. via such editors and tools like Protégé, SMORE, Jena, FOAF) [Feldmann].

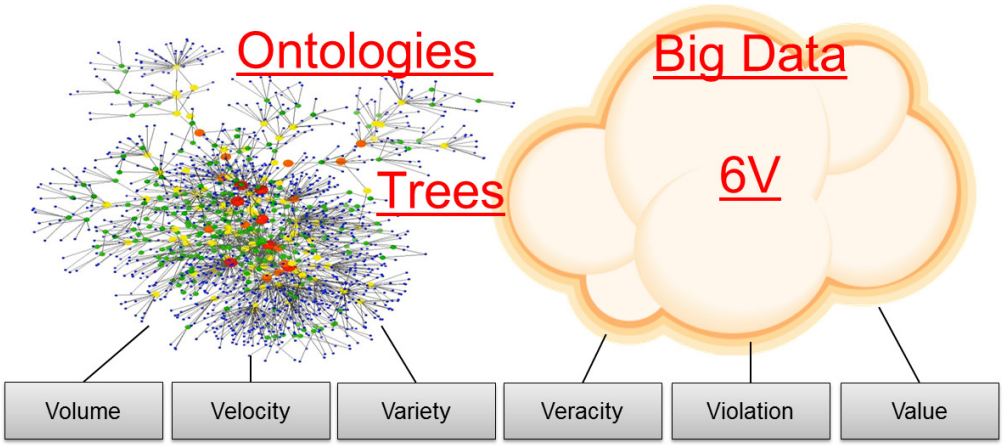

Figure 5. Can the ontologies be useful for Big Data?

Nevertheless such kind of techniques is heavy-weighting. Furthermore the heavy-weighting approaches (ontologies, knowledge bases, and fuzzy logic and fuzzy knowledge bases) are compared to light-weighting one. The existing reference solutions are discussed. Our paper examines the case studies on the both topics.

In contrast to the above described methods the light-weighting approach use empirical methods, freeware and common tools to data analysis.

3. Case Studies

3.1. Ontology Based Big Data Compression

This framework [6] is dedicated to health, a very demanding (challenging) domain. i.a. via the problems with large amounts of patient data, stored in different formats: e.g. handwritten texts, imaging results, lab results, genomic data etc. The unstructured data have to be transformed to structured formats via their taxonomation and forming of ontologies (cp. Fig. 6) in few steps [5, 6].

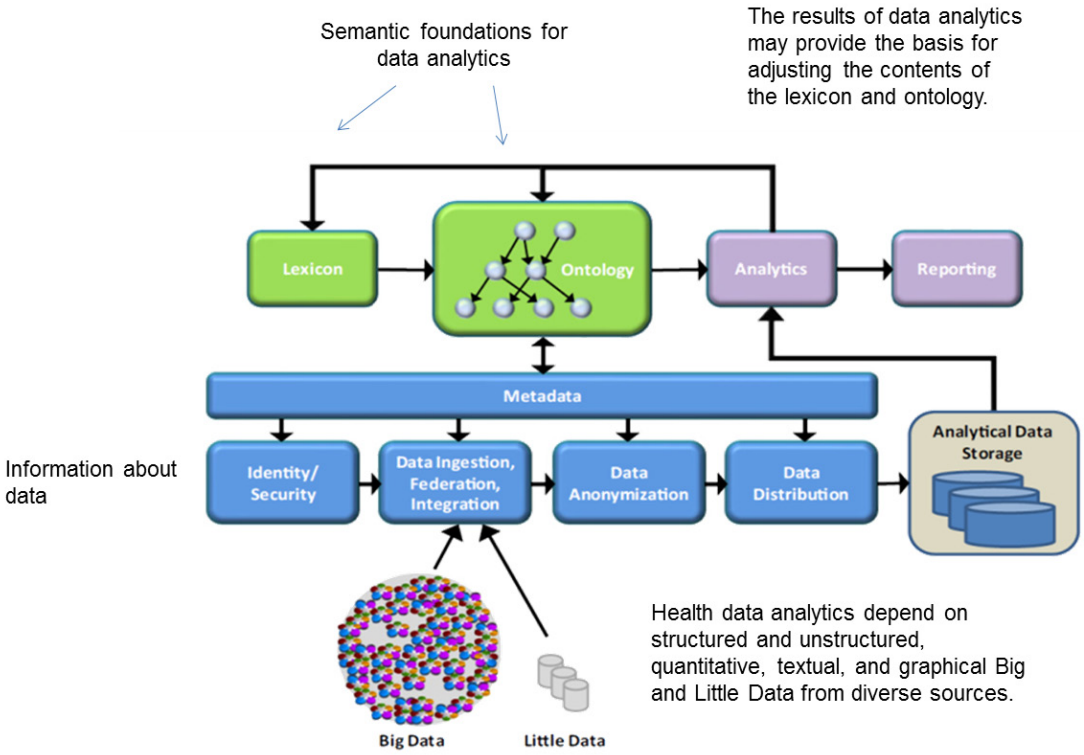

Figure 6. An ontology-based framework for data analysis [6]

3.2. Knowledge Based Big Data Compression

Telecommunication companies and mobile providers invest a lot of costs in analytical tools for service development and QoS increasing. The ongoing data analysis is aimed to [7 - 9]:

1. increase of sales;
2. assure revenue (detect and prevent revenue leakage);
3. reduce churn and fraud;
4. improve risk management;
5. decrease operational cost;
6. improve visibility into core operations, internal processes and market conditions;
7. discern trends and establish forecasts;
8. cross-sell/up-sell products and service plans etc.

The main idea of the method is as follows: overcoming of the Big Data problematics is possible under use of "data compression" via their transformation into knowledge [12 - 15]. The method is oriented primarily to the analysis of the data for telecommunication companies and mobile providers.

In this paper, we propose to consider a fuzzy logic approach that helps to reduce computational complexity in the process of classifying large amounts of data, the efficiency is considered on the example of obtaining a general estimation of quality index of providing services.

To determine the quality index of providing services, it is assumed to obtain it based on the statistical information of the operator's monitoring system to minimize "human" factor influence. Based on the obtained value of the integral quality index, following actions to improve the quality of service are determined.

To solve this problem, it is proposed to use decision-making methods based on fuzzy logic. Fuzzy expert rules are formed, which are the basis of the expert system. Fuzzy rules in such a system can be periodically adjusted to the current status of the technical infrastructure by means of their reformation (refinement) in the process of providing services by telecom operator.

The approach current status estimation of providing services by telecom operator consists of such stages:

1. Quality index of providing services obtaining (Y).
2. Summarized index characterizes the system status cumulatively on the basis of measured private indexes.
3. Fuzzy knowledge base (FKB) formation using the current values of integral quality index.
4. To compress large amounts of data, it is possible to form FKB in the form of fuzzy logic rules. FKB uses fuzzy logic methods to obtain conclusions.
5. Estimation service quality by telecom operator using an integral quality index of providing services.

FKB formation is considered on the example of service quality estimation. Formation algorithm in the generalized form can be presented as follows:

Initial data: measurement tables provided by telecom operators. The measurement table is a set of parameters, which we denote by $X_1 \ldots X_n$.

To build FKB it is necessary to split measurement table into 3 samples:

1. Training sample with M_1 rows, where $M_1 = \{1, k\}$, which is needed to form fuzzy logical rules of knowledge base;
2. Test sample with M_2 rows, where $M_2 = \{k + 1, n\}$, which is needed to check fuzzy logical rules quality of knowledge base;
3. Examination sample with M_3 rows, where $M_3 = \{n + 1, m\}$, which is required for the final verification of the correct operation of the obtained FKB.

Generic algorithm for determining an unknown value of Y:

1. Y_D calculation using the desirability function based on M_1 data.
2. FKB formation using M_2, and obtained values of Y_{FKB}. FKB is formed by a set of rules:

$$\text{«IF } X_1, X_2, X_3, \ldots \text{, THEN } Y \text{»}.$$

Therefore it is necessary to perform such steps:

2.1.Clusterization.

2.2.Membership function selection.

2.3.Y_{FKB} calculation based on the selected membership function.

2.4.If $|Y_D - Y_{FKB}| > \varepsilon$, then transition to 2.2, where ε is acceptable deviation.

2.5.If $|Y_D - Y_{FKB}| \leq \varepsilon$, then the membership function was chosen correctly.

The main points of the proposed approach are described in more detail below.

Such parameters of the monitoring system are used:

- Connection Success Rate, %
- Connection Block Rate, %
- Connection Drop Rate, %
- PS Attach Success Rate, %
- PDP Context Activation Success Rate, %
- Speed DL, kbitsps
- Iub Congestion, %
- Backhaul Accessibility, %

- DNS Success Rate, %
- DNS Response Latency, ms
- $W_{1...10}$ – linguistic variables efficiency.

It is possible to form FKB for estimating the integral quality index and predicting trends of its change in a short time interval based on integral quality index parameters.

Input information is the observation table $\boldsymbol{T} = \{\boldsymbol{t}_{M1}\}$, where i-th element $\boldsymbol{t}_i = (\boldsymbol{x}_i, y_i)$, $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ik}]$ — vector of input values, y_i — output value, M_1— number of observations, k — number of input variables, $\boldsymbol{x}_i \in DX$, $y_i \in DY$.

They are needed to be converted to such values, when system can estimate service quality level cumulatively, because the data values in the table do not carry information. It is proposed to use Harrington's desirability function.

The mathematical apparatus of converting specific parameters into abstract numerical values uses one of the logistic functions, which is called the "desirability curve":

$$d = \exp[-\exp(-Y)],$$

where coordinate axis Y is private indicators scale; axis d is desirability scale.

For evaluation integral quality index desirability scale is divided into five subranges in the range from 0 to 1: [0; 0.2] – «very bad», [0.2; 0.4] – «bad», [0.4; 0.6] – «satisfactorily», [0.6; 0.8] – «good», [0.8; 1] – «very good».

The obtained value d(i) for the i-th parameter is recalculated together with others into a generic desirability coefficient – D.

$$D = \sqrt{n \cdot d(1) \cdot d(2) \cdot \ldots \cdot d(n)},$$

where n is the number of used indexes.

The number of such indexes may not be the same for different systems. This allows comparing generic coefficients even when there are not some comparison parameters for different systems or data on them.

Consider the formation of fuzzy knowledge base.

Initial data: fuzzy model is defined as a system with input variables $\mathbf{X} = \{X_{k+1}, X_{k+2}, \ldots, X_n\}$, which are defined on the input area of discourses $DX = DX_{k+1} \cdot DX_{k+2} \cdot \ldots \cdot DX_n$, and one output variable Y, which is defined on the output area of discourses DY. Explicit value is denoted as x_i for input variable X_i and as y for output variable Y.

There are many methods of clusterization. In this paper, a fuzzy c-averages algorithm is considered. Fuzzy c-averages algorithm is based on distance minimizing from the observed data to cluster centers.

Next step is to consider Y_{FKB} getting based on the selected membership function. Now the rules of fuzzy base are formed.

The main stages of obtaining fuzzy logical inference:

1. Fuzzification (fuzziness introduction).
2. Aggregation (the truth degree of conditions for each rule of fuzzy logic inference algorithm determination).

If the rule condition has a simple form, its truth is equal to the corresponding value of the membership function. If the condition has the following form:

RULE<#>: IF «b_1 is a_1» AND «b_2 is a_2», THEN «b_3 is V»,

then truth degree is determined on the basis of known truth values of the subconditions and expressions for performing operations of fuzzy conjunction and fuzzy disjunction are applied:

Fuzzy logic conjunction operation (AND): $\mu_c(x) = \min\{\mu_A(x), \mu_B(x)\}$

Fuzzy logic c disjunction operation (OR): $\mu_c(x) = \max\{\mu_A(x), \mu_B(x)\}$

1. Accumulation (finding the membership function for each of the output linguistic variables).
   If conclusions relating to the same output linguistic variable belong to different rules of the fuzzy logic inference system, they are combined into fuzzy sets using the expression:

$$y = \min\{x_m\},$$

   where x_m is modal value (moda) of Fuzzy set, which is corresponded to the output variable after the accumulation step, obtained in accordance with expression:

$$x_m = \max\{\mu(x)\},\ x \in [a, b]$$

2. Defuzzification (leading to explicit).

Fig. 7. shows, that the value of membership function can be corresponded to several arguments (if Y = middle from P.6, then at this stage only the membership function, which is responsible for the value of "MIDDLE" is used). In this paper right modal value method is used:

$$y = \max\{x_i\}.$$

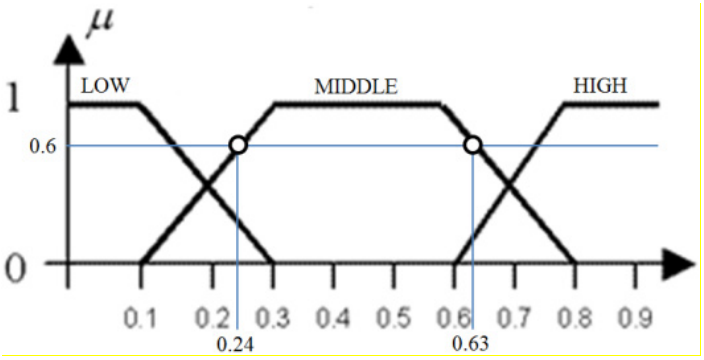

Fig. 7. An example of membership function [7, 8]

To verify the proposed approach, data from one of the Ukrainian telecom operators was used. Natural values were normalized in the range from 0 to 1. The fragment of the data is presented in Fig. 8.

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
0,01	0,99	0,98	0,95	0,95	0,98	0,99	0,99	0,0075	0,99
0,001	0,9948	0,9989	0,9893	0,9958	0,9967	0,9967	0,9981	0,0033	0,9959
0,001	0,994	0,9989	0,9892	0,996	0,9972	0,9972	0,9968	0,0031	0,996
0,0006	0,9983	0,9993	0,9814	0,9643	1	0,9991	0,9992	0,0025	0,9984
...	...	...	...	...	...	...	...	...	...

Figure 8. An example of initial data of Ukrainian telecom operator

According to the algorithm for determining unknown *Y*, the data were divided into three samples: teaching, test, examination.

From the first sample, *Y* was obtained using desirability function.

From the second sample fuzzy knowledge base was formed, which was consisted of such rules:

IF X1 = middle AND X2 = not low AND X3 = not low AND X4 = not low AND X5 = not low AND X6 = not low AND X7 = not low AND X8 = not low AND X9 = not very high AND X10 = not low,
THEN Y = not low

IF X1 = middle AND X2 = not low AND X3 = not low AND X4 = not low AND X5 = not low AND X6 = not low AND X7 = not low AND X8 = not low AND X9 = not very high AND X10 = not low,
THEN Y = not low

IF X1 = very low AND X2 = high AND X3 = high AND X4 = high AND X5 = high AND X6 = high AND X7 = high AND X8 = high AND X9 = very low AND X10 = high,
THEN Y = high

IF X1 = very low AND X2 = high AND X3 = high AND X4 = high AND X5 = high AND X6 = high AND X7 = high AND X8 = high AND X9 = very low AND X10 = high,
THEN Y = high

During the experiment, it was investigated that fuzzy knowledge base was formed at 4^{th} iteration.

Correctness of the algorithm was tested by the third sample. And the results of the algorithm fully corresponded to the expert estimates provided by one of the Ukrainian telecom operators [7, 8].

3.3. A Light-Weighting Approach

Let us to examine a system example for overcoming of Big Data complexity which was developed via TIQ Solutions in Leipzig [16, 17]. The mentioned reference solution of TIQ Leipzig is represented below. The system is based on the light-weighting approach and is oriented to processing of Big Data in the domain of Business Analytics.

The system architecture is depicted via Fig. 9. One of the important criteria of the quality of the referred solution is scalability.

This can be considered as a freeware reference solution interfaced to the use on a Linux-cluster. The system allows the connectors to binding of the conventional data systems (DB, Data Mining etc.).

The complex poly-structured redundant retrieved data can be processed with higher performance within an enterprise or institution computing center or cluster. Some java implemented modules allow real time processing control. The CAPEX and OPEX are reduced based on the commodity hardware as well as freeware components.

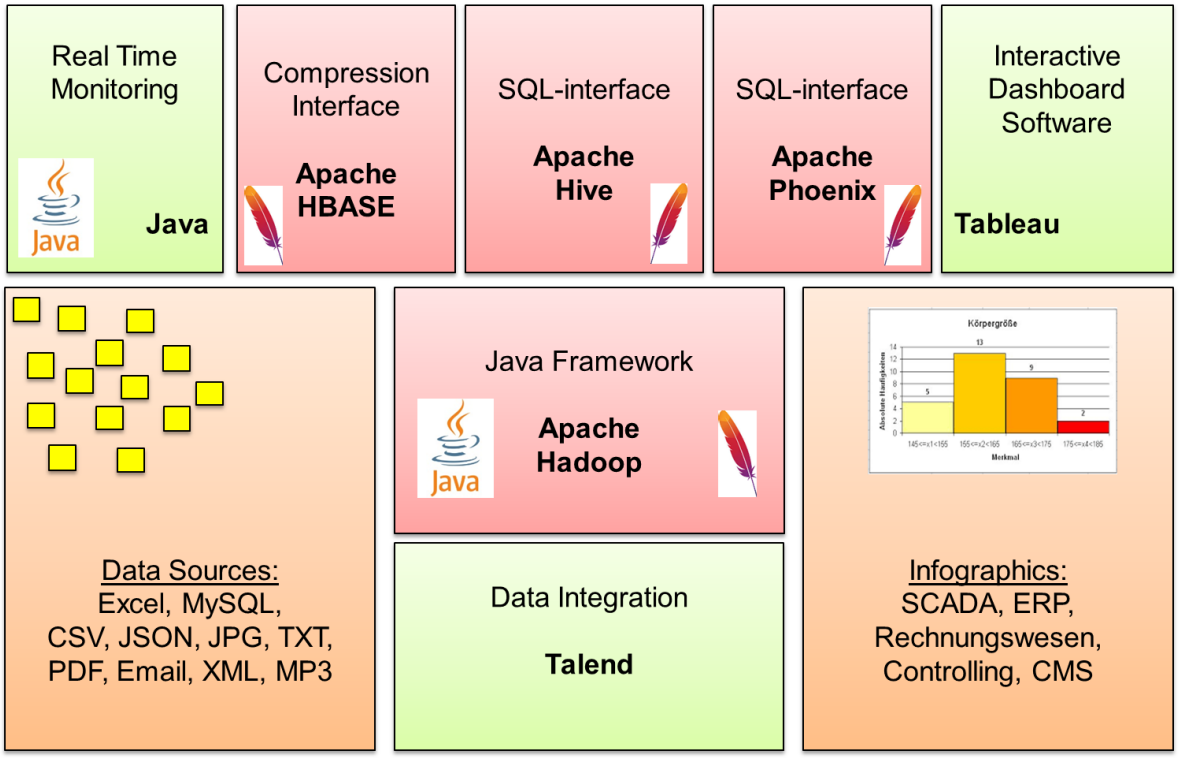

Figure 9. The architecture of a light-weighting solution for Big Data [16, 17]

The discussed architecture is characterized via:

- agility
- no rigidity and any large license costs
- use of proven freeware.

The components of the architecture are as follows [17]:

1) Apache Hadoop (since 2008):
   - in Java implemented freeware framework (FW) for scalable distributed applications.
   - Use of Google-similar algorithms as well as GFS (Google File System)
   - Support of intensive computing with significant data (PByte-area) on the computer clusters.
   - Famous users are such as: FB, AOL, Baidu, IBM, Yahoo.
2) Apache HBASE:
   - scalable DB for management of Big Data within the Hadoop clusters.
3) Apache Phoenix:
   - Massively paralleled RDB-Engine with OLTP concept for Hadoop with Apache HBase as Backstore.
   - SQL queries are processed, then compiled into the series of HBase scans and are orchestrated to produce the regular JDBC results.
4) Apache Hive:
   - DBMS for Hadoop aimed to data aggregation, queries and analysis
   - SQL-similar interface aimed to generating of queries to the data retrieved in diverse DB and file systems under use of Hadoop.
5) Tableau (Seattle, 2003):
   - Software for data visualization and reporting.

6) Talend:
   - Cloud software for Big Data Integration.
7) SCADA (Supervisory Control and Data Acquisition):
   - SCADA is the computer-assisted monitoring and control of technical processes.

4. Conclusions and Outlook

1) A so called "6V paradigm" on Big Data is formulated.
2) The mostly common techniques classification and overview is given.
3) So called "heavy-weighting" and "light-weighting" approaches are discussed.
4) Case studies on use of ontologies, FKB as well as empirical concepts are examined.

This work can be qualified as a work-in-progress. The authors try to find new efficient methods to overcome the above mentioned 6V via combination of the examined methods: FKB-based data reducing, ontology-based as well as agile, rigidity-free and freeware based solutions.

References

[1] Ulema M.: Big Data and Telecommunications Telecom Analytics – Tutorial II, BlackSeaCom'2016.
[2] Gartner Inc. IT Consulting and Reports (online 2017): http://www.gartner.com/.
[3] Keberle N.: Modeling of dynamic domains under use of the ontologies, Bulletin of Kharkiv Air Force University, vol. 3, 2009, pp. 121-127.
[4] Luntovskyy A., Spillner J.: Architectural Transformations in Network Services and Distributed Systems: Service Vision. Case Studies, XXIV, 344p., 238 pict. , Springer Nature Verlag, April 2017 (ISBN: 9-783-6581-484-09).
[5] Konys A., Rogoza W.: Big Data and Ontologies. Talk at ACS Int. Conf. 2016 in Międzyzdroje, Oct. 2016, 3 p.
[6] Kuiler E.: From Big Data to Knowledge: An Ontological Approach to Big Data Analytics, Review of Policy Research, Volume 31, Number 4 (2014)
[7] Globa L., Svetsynska I., Volvach I.: Integral Quality Index of Providing Services Calculation, Conference 2017, 6 p.
[8] Globa L., Zacharchuk A., Ischenko I.: Expert system for decision-making on the basis of fuzzy logic algorithm, ACS 2016 Conference 2016, 12 p.
[9] Globa L., Novogrudska R.: Workflow of Ukrainian National Antarctic Scientific Centre Portal Modeling, Ukrainian Antarctic Journal, pp. 229-237, Antarctic, 2016.
[10] Ivanytska N., Stryzhak O.: Role of Ontology in the System of Formation of Educational and Cognitive Competences on Physics of Secondary School Pupils, in: Information Technologies and Learning Tools, vol. 39 (1), 2016, pp. 160-169.
[11] Stryzhak O., Globa L., Kovalskyi M.: Increasing web services discovery relevancy in the multi-ontological environment, in: Advances in Intelligent and Soft Computing Serices (AISC), Springer, 2014, 5 p.
[12] Russell S., Norvig P.: Artificial Intelligence: A modern approach, New Jersey, Upper Saddle River, 2010.
[13] Jones M.: Artificial Intelligence: A Systems Approach, Hingham, Massachusetts, New Delhi, Infinity Sci. press LLC, 2008.
[14] Kriesel D.: A Brief Introduction to Neural Networks (online 2017), http://www.dkriesel.com/en/science/neural_networks.
[15] Marr B., Wiley J.: Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance, Sons Ltd, 2015.
[16] Vortraege der Hausmesse des IBH 2017 0n 23.3.2017 (in German).
[17] TIQ Solutions Leipzig (online 2017): https://www.tiq-solutions.de.