IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

A Method to Solve Uncertainty Problem for Big Data Sources

Andrii Berko
Information Systems and Networks Department
Lviv Polytechnic National University
Lviv, Ukraine
andrii.y.berko@lpnu.ua

Vladyslav Alieksieiev
Applied Mathematics Department
Lviv Polytechnic National University
Lviv, Ukraine
vladyslav.i.alieksieiev@lpnu.ua

***Abstract*— Big Data analysis and processing is a popular tool for Artificial Intelligence and Data Science based solutions in various directions of human activity. It is of a great importance to ensure a reliability and a value of data source. One of the key problems is the inevitable existence of uncertainty in stored or missing values. Any uncertainty in a source causes its disadvantageous, complexity or inapplicability to use. That is why it is crucial to eliminate uncertainty or to lower uncertainty influence. Here in this research, we offer ontology-based method to solve an uncertainty problem for big data sources.**

Keywords*— *big data; data sources; data uncertainty; ontology; uncertainty elimination

I. INTRODUCTION

Nowadays there are many areas requiring to solve problems with artificial intelligence solutions and tools supplemented with the necessity to use big data sources. It concerns many tasks in business, finance, medicine, politics, ecology and ecological surveillance, and many other all requiring artificial intelligence.

These tasks need to take into consideration such features of big data like volume, velocity, and variety. Meanwhile, significance and reliability of the data should be kept. Special preparations should be made with big data sources before use. Those can be ETL (extract, transform, load) processes, normalization, aggregation etc. This is the step of data source processing when the problem of uncertainty appears.

An anomaly appears in some missing values, incomplete data, inaccuracy, inconsistency, unreliability etc. Generally, this lowers the value of the big data source, reliability of final results, makes it difficult or impossible to work with the resource. These are the reasons to consider the importance of the problem of elimination or lowering the influence of uncertainty level in big data sources.

The problem of data uncertainty has been discussed for a long time. Solutions for the problem and its different aspects were offered in [1–5]. Also, there were researches on some particular cases of custom IoT based monitoring system like a problem of data losses [6] and a problem of aggregation of obsolete data [7].

Difficulties of solving the uncertainty problem for big data sources are explained with its features: huge volume, high level of velocity and variety. Due to these features, all standard tools are not applicable. And this is the motivation to develop some new approaches oriented to interact with sources of big data.

II. PREREQUISITES OF BIG DATA UNCERTAINTY PROBLEM

A. Why Big Data?

The primary question is to understand the peculiarity of uncertainty problem in big data. First, the answer comes from its key features. Basic characteristics differing big data from other sources types are so named "triple V" – Volume, Variety, and Velocity. These are the specific features responsible for the appearance of effect and problem of uncertainty in big data sources. Now, let's discuss the influence of these features.

1st. Big and huge volumes of incoming source require its distribution. In the meantime, different parts can be managed with different tools. This kind of architecture does not allow to maintain the global integrity of the data. Inability to ensure source integrity causes inevitably to the appearance of inconsistency in data, data losses and data distortion. As a result, one gains the uncertainty of some part of a big data source.

2nd. The variety of big data requires using different schemas, descriptors, and another mechanism to describe data within the global resource. Thereafter, this causes to appear an inconsistency, duplication, incompleteness, ambiguity, and different interpretation of data units. In an example, in one source the data may exist, and in another source, similar by meaning, it may be absent. Another inconsistency is to have the same presentation for different data or different presentation for the same data in different parts of the global source. Again there is a reason for an effect of uncertainty within a global big data source.

3rd. The velocity of big data combined with no control of integrity makes any changes asynchronous and inconsistent. There may also happen, that some values of data endure some changes but a corresponding or similar values remain unchanged. Besides, fast and unsynchronized changes make it appear incorrect and inconsistent values, or unpredictable loss of relevance of some data, etc. This is one more reason for uncertainty to appear in big data sources.

Thus, the conclusion can be made, that the effect of uncertainty is natural for big data. Uncertainty seems to exist almost always in big data sources, due to the basic features – Volume, Variety, and Velocity.

There are some more factors causing the uncertainty in big data sources. These are the requirements for the resource known as "*another double V*", what means Value and Veracity [5]. Value means the cost and applicability of data

to solve some definite set of problems. Veracity is the relevance, accuracy, and reliability of the data. It is obvious, that the high-level uncertainty makes it is impossible to fulfill the requirements. Some traditional data sources (Databases, Data Warehouses, formatted documents, XML etc.) to fulfill the requirements due to a low level of uncertainty, and some other data sources (private sources, texts, pictures, social networks) has no such requirements to be followed strictly. At the same time, those big data sources the problem of uncertainty needs a special way for solution.

Usage peculiarities of big data can be defined as a third aspect, establishing the specifics of uncertainty problem. While the traditional data sources can be considered available and ready to use, the big data sources have a need to implement some preparation procedures. This preparation procedure is typically made as ETL (extract–transform–loading), normalization, aggregation etc. Nevertheless, in all cases, a data cleaning should be made to prepare big data to use. And one of the most important steps of data cleaning is to eliminate the uncertainty or to lower the uncertainty level to reach some appropriate level.

B. Types of data uncertainties

C. J. Date [1] asserts not all the uncertainties to be the same. It can be divided by the origin factors, origin nature, and abilities of interpretation. Really, there is a difference between those cases, when the value does not exist due to impossibility, when the value exists but remains unknown, and when the value exists and known, but it is inconsistent or ambiguous. According to C. J. Date concept [1], there are the following types of uncertainties:

- **inexistent** value for the data element,
- value **not formed** yet at the moment,
- value exists, but **unknown**,
- value exists, but not received (**obtained**),
- value is invalid (**unacceptable**),
- value not **determined**,
- value is inconsistent (**corrupted**),
- value is **ambiguous**,
- value is not accurate enough,
- value is an empty, etc.

The list is incomplete and there more items could be added. Depending on contents and peculiarities of the source of big data, there can appear some different causes for data uncertainty. Why should we categorize data uncertainty in some resource? First, the way of uncertainty elimination depends on the nature and factors of origin of the uncertainty. For example, if the value exists, but remains unknown, then it can be queried again; if the value is inaccurate, then it can be refined; if the value is not formed, then it can be received later; etc. Second, when uncertainty elimination is impossible, then the process of the source processing could be planned accordingly; if the value is inappropriate, or invalid, or empty, then it can be excluded from processing; if the value is inaccurate or inconsistent, then the level of accuracy or confidence can be changed, etc. That is why the categorization of uncertainties is recognized as a key element to solve the uncertainty problem for big data source.

C. Approaches to data uncertainty problem solution

There are a number of approaches to solve the problem of uncertainty in big data sources. The primary target is to eliminate uncertainty or to lower uncertainty level to support effective resource usage. The most popular approaches to uncertainties elimination are:

- **repeat a request** to receive a value,
- refine inconsistent or inaccurate values (**rectification**),
- eliminate the origin of inconsistency, repeat or ambiguity of values,
- **replace** uncertainty with some aggregate value (average value, probable value, standard or default value, initial value, some calculated value, estimated value, expert value, etc.),
- use of **fictitious** value as artificial surrogate marks instead of uncertainty,
- **remove** of uncertain value or data element from the resource,
- **ignore** the data uncertainties while processing the resource,
- create **special tools** to process uncertainties.

Besides those mentioned, there are some other approaches to eliminate uncertainties that can be used according to peculiarities of a big data source. The choice of appropriate approach is rather a difficult task. The appropriate approach should match the following conditions:

1st. Match the type of uncertainty.

2nd. Match the contents and peculiarities of big data source.

3rd. Ensure a correct result.

4th. Determine the best approach to reach the aim.

The solution of a problem of choosing the appropriate approach to eliminate uncertainties is one of the important steps to prepare a big data source for the use.

III. ONTOLOGY-BASED SOLUTION OF DATA UNCERTAINTY PROBLEM

A. Data uncertainty and ontology

As it was shown above, there is a direct tie between the approach to eliminate data uncertainty and the nature of the uncertainty. Meanwhile, the question comes to define the best approach to match the uncertainty type. The answer is ambiguous. Using an expert approach there were defined some cases of conformity. The cases are presented in Table I.

Obviously, to use an appropriate approach to eliminate data uncertainty it can be not enough only to define the nature of uncertainty.

Generally, the process of elimination of data uncertainty can be described as a formation of a new value for data element to replace uncertain value. Thus, the new value becomes explicit, definite, exact, unambiguous (consistent) and acceptable.

TABLE I. THE RELATION BETWEEN UNCERTAINTY TYPES AND METHODS OF ITS ELIMINATION

Uncertainty Type / Method of Elimination	Not exists	Not formed	Unknown	Not obtained	Unacceptable	Not determined	Corrupted	Ambiguous
Repeated request	+	+	+	+	-	+	+	+
Rectification	-	-	-	-	-	-	+	+
Replacement	-	-	+	-	-	+	+	+
Factitious value	-	-	+	-	+	+	-	-
Remove	-	-	-	-	+	-	-	+
Ignore	+	-	-	-	-	+	-	-
Special tools	+	+	+	-	+	-	-	-
No Action	+	-	+	-	-	+	+	-

The value v_{ij} of some data unit V_i, which was formed to eliminate uncertainty, depends on uncertainty category U_k and elimination approach S_l. The model of new value formation to be described as consequent transitions: "data unit – uncertainty – elimination – new value", or as

$$V_i \rightarrow U_k \rightarrow S_l \rightarrow v_{ij} \quad (1)$$

or as a function

$$v_{ij} = \Phi(V_i, U_k, S_l) \quad (2)$$

Basic assignment of the model is to answer the question: which method of uncertainty elimination, for which data unit, of what uncertainty type, when and how should be implemented. One solution of the problem is to implement a special knowledge base within cleaning tools for big data. The knowledge base should include some expert and synthetic knowledges like:

- structure and contents of a big data source;
- types of uncertainties in a big data source;
- approaches to eliminate uncertainties in big data source;
- correspondence between data units, uncertainties types, and elimination approaches.

The key tasks for that knowledge base are 1st – to accumulate expert knowledges about approaches for elimination of different types of uncertainties for particular data units of the big data source; 2nd – to develop and to supplement knowledge base with new knowledges; 3rd – to use knowledges for the purpose of data uncertainty elimination. The use of knowledge base allows to exclude or to reduce the influence of human factor and to increase the quality of preparation results for big data source.

The basis of knowledge base is the special kind of ontology. Generally, the ontology can be defined as

$$O^O = < C^O, R^O, F^O >, \quad (3)$$

where $C^O = \{C^V, C^U, C^S\}$ is the set of concepts (classes), which are

C^V – data units of the resource,

C^U – types of uncertainties of the resource,

C^S – approaches to eliminate data uncertainties.

$R^O = \{R^{VU}, R^{US}, R^{VS}\}$ is the set of relations between ontology classes, respectively,

R^{VU} – relation between data units and uncertainties types,

R^{US} – relation between uncertainties types and approaches for uncertainties elimination,

R^{VS} – relation between data units and approaches for uncertainties elimination.

F^O – the set of rules (axioms) of data uncertainties elimination. Each rule defines the approach to eliminate particular uncertainty type within a particular data unit. Unlike to those conformities from Table I, the result of such rule implementation should be definitely unambiguous.

The common structure of the ontology for elimination of uncertainty in big data sources is described at Figure 1.

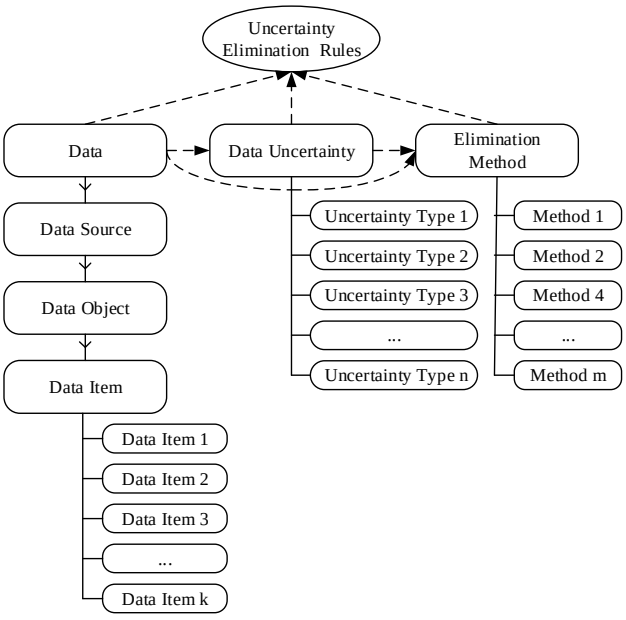

Fig. 1. Common ontology structure for Data uncertainty elimination in Big data Sources

The given ontology has the following concepts (classes) defined:

- The class "Data" describes the common content of big data source. Class is hierarchical and includes subclasses. 1) "Data Source" – local resource, like database, document repository, web-resource, content storage etc. 2) "Data Object" – sub source of a local resource, like a table, file, document,

container. 3) "Data Item" – class object, the elementary data unit having its own interpretation, like a column, field, XML-element, JSON-element and so on.

- Class "Data Uncertainty" describes data uncertainties of big data source. These uncertainties to be defined as it was shown above. The content of these uncertainties to be specific for each resource.
- Class "Elimination Method" consists of set of methods to be implemented for elimination of data uncertainty at a particular resource.

Relations defined in the ontology describe: 1) between data items and data uncertainty types; 2) between data uncertainty and related methods of its elimination; 3) between data items and possible data uncertainty elimination methods.

The defined ontology classes are connected to a set of rules to use methods of uncertainty elimination for particular units of big data source. Each rule can be a production rule of the form "if–then–else". It defines the approach of elimination of uncertainty in big data source following the principle

FOR <*Data Item i*> **IF** <*Data Uncertainty*>

IS <*Uncertainty Type j*> **THEN**

< *Elimination Method k*> **WOULD BE USED**.

This is the way the ontology allows to answer definitely to an earlier posed question: which method of uncertainty elimination to use with what kind of data while preparing some resource of big data.

B. The algorithm of ontology-based data uncertainty problem solution for big data sources

With the ontology for big data source to eliminate uncertainties a specialized algorithm is offered. The algorithm consists of three stages. The stages are to describe the solutions for the types of tasks respectively: 1) elimination (lower the level) of uncertainty of big data source; 2) analysis and identification of problem situations that appeared during the data uncertainties elimination process; 3) learning and improvement of ontology. For the purpose of best effectiveness stages 1 and 2 to be done first. Stage 3 can be executed after the elimination of uncertainties in big data source and after fixing problem situations. Now there is a description of the algorithm.

/* Part 1 – Data Uncertainty Elimination */

Step 1. *Research on resource (sub-resource, unit) of data*. The step is to reveal data uncertainties. In case of success (uncertainty found) Step 2 should be made, otherwise, (no uncertainties found) current step to be repeated for the next data unit of the resource (sub-resource). When research is finished for all the data units, then move to Step 8.

Step 2. *Uncertainty categorization*. The uncertainty to be apprised to find out its type according to ontology defined types. If the type is found, then Step 3 should be made, otherwise – Step 6.

Step 3. *Search for an approach to solve uncertainty*. This step is to find out whether the ontology has the definition of the required approach to eliminate the particular uncertainty type. If there is an appropriate approach defined, then Step 4 should be made, otherwise move to Step 6.

Step 4. *Making a decision to eliminate uncertainty*. Among the set of rules, there should be found a rule to conform data unit and uncertainty type and its elimination approach. If the rule was found, then Step 5 should be made, otherwise (no rule found in the ontology) move to Step 6.

Step 5. *Execution of uncertainty elimination according to the rule*. This step is processed with the call to some specific procedure, which was previously defined. If the uncertainty is eliminated, then the changes to the data are made. After that move to Step 1 for the next data unit. If the uncertainty was not eliminated, then move to Step 6.

/* Part 2 – Problem Analysis */

Step 6. *Processing the problem situation*. This step is to recognize and fix the problem situation, appeared during the attempt of data uncertainty processing. These problem situations can be categorized according to its origin:

- no description of data unit (sub-resource) in the ontology;
- no description of uncertainty type for the data unit in the ontology;
- no description of approach to eliminate some type of uncertainty or in some type of data unit;
- no description of rule to eliminate uncertainty for particular data unit;
- implementation of rule to eliminate uncertainty did not effect.

Step 7. *Fix the problem situation*. If the situation is recognized and categorized it can be fixed in a special log (or register) using some particular format. After fixing the problem situation move back to Step 1.

/* Part 3 – Ontology Learning and Evolution */

Step 8. If the result of Steps 1–7 there is no fixed problems, or there are no unprocessed records in the log (register), then the ontology-based algorithm is finished. If there are records in the log, then move to Step 9.

Step 9. *Unsupported Learning*. A specially prepared procedure for autonomous improvement of the ontology should be made. It is supposed to make supplements with new descriptions of the data unit, uncertainty types, uncertainty elimination approaches, relations and rules to eliminate uncertainties. Some service resources (catalogs, vocabularies, thesauruses, etc.) should be used for that purpose. If the supplement is successful, then the Step 8 is to be repeated. Otherwise, if unsupported learning failed, then Step 10 should be made.

Step 10. *Supported Learning*. The problem situation to be analyzed by an expert to make a decision on how to eliminate uncertainty. According to that decision, there should be made some changes to the ontology of big data source. If the expert is unable to solve the problem situation, then the situation is marked like unsolvable. The description

of the problem situation should be made in the ontology. And then move back to Step 8.

This is the end of the algorithm. If the result of execution of all algorithm steps is that the data is cleared to the required level, then the algorithm considered to succeed. If the level of uncertainty for the data is higher than an accepted level, then an algorithm can be executed again.

Finally, the algorithm allows to achieve two aims: 1) the big data source can be cleared from incomplete, ambiguous, or other non-quality data; 2) the ontology, as a core of knowledge base, can be learned and improved. The development of the ontology is recognized to be the key to improve the results of solving the uncertainty problem of big data sources.

IV. AN EXAMPLE OF ONTOLOGY-BASED PROCESSING OF DATA UNCERTAINTY

As an example of ontology-based method elimination of data uncertainty using, input data stream of news portal is considered. Such data stream is the time-serialized sequence of news data block. News data block is formed by a robot, news aggregator etc. Usually, news data block may contain uncertain data such as absent, incorrect, invalid, or unreliable values because data were obtained from various sources. So, procedures of data uncertainty processing are needed to be performed before download news data block into repository of news portal.

In considered case, each news data block contains its ID and set of records. Each record is structured according to the document-oriented model and is called "document". Document contain description of one of news message about any new event, presented in JSON format. Proposed structure of document is the next:

News ID – unique identifier of document using for its identifying and search in the news repository,

News Category – classify message by predefined category (so as Politics, Society, Culture, Sport etc.),

News Priority – describe event value in the grneral news context,

News Date&Time – when presented event occurred,

News Place – where (country, region city) presented event occurred,

News Object – persons, organizations, institutionsor etc., to which the event relates,

News Subject – describe what the event means,

News Source – define where did the news come from,

News Text – contains text of message about presented event and its details.

Each block of news may contain some uncertainties in the data, because it formed by special program tools using information sources of various format. Such data uncertainty types as value absence, invalid values, out of range values, and unreliable values has been defined as possible uncertainties for news data stream, Each type of data uncertainty is associated with a certain condition, which lets detect this uncertainty in data set (Table II).

Some method of data uncertainty elimination in news data stream has been defined as well in considered example. These are such as (1) repeated request for a message; (2) set default value; (3) reject message; (4) no any action.

TABLE II. THE ASSOCIATION BETWEEN UNCERTAINTY TYPES AND CONDITIONS OF ITS DETECTION IN NEWS DATA STREAM

Uncertainty type	Uncertainty condition
absent value	Is Null
invalid value	not in *ValueDomain*
value out of range	not between *MaxValue* and *MinValue*
unreliable value	not in *ReliableValuesSet*

Problem of defining what method should be applied for elimination of uncertainty of any type for certain data item processing. For solution this problem using developed method special ontology has been developed. Such tools as Protégé ontology editor and OWL ontology model were applied for ontology development.

Ontology for data uncertainty elimination in news data stream include main class **NewsDataStream** divided into three classes:

(1) class **NewsBlock** which include entries **NewsBlocdID** and **NewsMessageDoc**;

(2) class **UncertaintyType** includes entries **InvalidValue**, **NoValue**, **OutOfRange**, **UnReliable** which corresponds to types of uncertainties in the input news stream,

(3) class **UncertaintyProcessing** contains entries **Request**, **SetDefault**, **Reject**, **NoAction** which corresponds to uncertainty processing methods developed for input news stream data.

Entry **NewsMessageDoc** also is class includes entries corresponding to each data item of news message according to the structure of the document (see Fig.3).

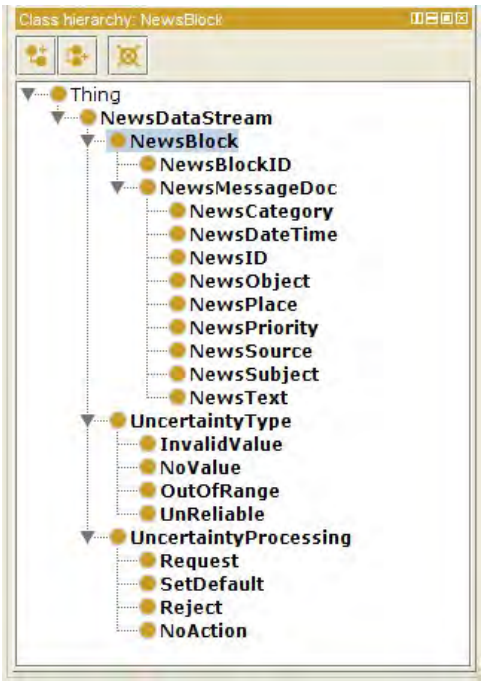

Fig. 2. Ontology of news portal input data stream structure created by Protégé

Relations between ontology classes of classes has been defined by using object properties tools. Two class of properties defined in developed ontology:

(1) object property class **IsUncertaintyOfType** is defined for establishing correspondence between the data items (class **NewsMessageDoc**) and data uncertainty types (class **UncertaintyType**);

(2) object property class **ProvessingMethod** is defined for establishing correspondence between the data uncertainty types (class **UncertaintyType**) and data uncertainty elimination method (class **UncertaintyProcessung**).

Each property defined to describe directed functional dependency between items of certain classes. Using of such dependencies allows define a rule for choose of uncertainty eliminated method for each given data item.

TABLE III. THE RESULTS OF EXPERT EVALUATION OF ONTOLOGY-BASED METHOD OF DATA UNCERTAINTY ELIMINATION

Evaluation criteria	**Manual Uncertainty Elimination**	**Ontology-Based Uncertainty Elimination**
Uncertainty Detection Level	low	high
Correctness of the method selection	middle	high
Performance	low	high
Absent value	high	high
Invalid value	middle	middle
Value out of range	middle	high
Unreliable value	high	middle

Experimental application of ontology has been evaluated by expert way in quality score: "low", "middle', "high". The results obtained after expert evaluation of processing about 500 news documents are presented in table III.

V. UNSOLVED PROBLEMS AND DISCUSSIONS

There are some problems appeared as a result of this research which require further discussion and investigation.

1st problem is to build an initial ontology of big data source. The problem requires some expert knowledge of data processing and detailed description of the resource. Typically, there is no such description for big data. That is why the building of the initial ontology is not enough defined problem.

2nd problem is to qualify the type of uncertainty. The problem has no formal solution often. That is the motivation for further researches in areas of machine learning and artificial intelligence to be able to solve the problem effectively.

3rd problem is the learning and development of the ontology. Unsupported learning requires using some additional methods and algorithms based on experience and analysis of numerous precedents of data uncertainty problem solving.

The solution of these problems can be a separate task for scientific researches. Making these researches can strongly improve the method presented in the paper.

VI. CONCLUSIONS

An approach to solve an uncertainty problem for big data sources was offered in the paper. The peculiarity of the solution is to use an ontology as a core of knowledge base. The ontology can be considered as a special type of metadata. The developed approach allows, first, to make better data clearing in a big data source, and, second, to accumulate for further use a knowledge and an experience to solve data uncertainty problem.

REFERENCES

[1] C. J. Date, Database in Depth: Relational Theory for Practitioners. O'Reilly, CA, 2005.

[2] K. Alieksieieva, and A. Peleshchyshyn, "Application of incomplete and inexact data for commercial web-project management," Scientific announcements of Lviv Polytechnic National University, Lviv, Ukrane, no. 805, pp.345-353, 2014.

[3] N. Marz , and J. Warren, Big Data: Principles and best practices of scalable realtime data systems. Manning Publications, 2015.

[4] J. Chen, D. Dosyn, V. Lytvyn, and A. Sachenko, "Smart Data Integration by Goal Driven Ontology Learning. Advances in Big Data," Advances in Intelligent Systems and Computing, Springer International Publishing AG, pp. 283-292, 2016.

[5] D. Losin, Big data analytics. Elsevier Inc., Waltham, MA, USA, 2014.

[6] V. Alieksieiev, and O. Gaiduchok "About the problem of data losses in real-time IoT based monitoring systems," Mathematical Modeling, STUME, Sofia, BULGARIA, Year I, issue 3, pp. 121–122, 2017.

[7] V. Alieksieiev, G. Ivasyk, V. Pabyrivskyi, and N. Pabyrivska, "Big data aggregation algorithm for storing obsolete data," Industry 4.0 – STUME, Sofia, BULGARIA, Year III, issue 1, pp.20–22, 2018.