The Concept and Evaluating of Big Data Quality in the Semantic Environment

Oleksandr Novytskyi[1]

[1], *Institute of Software Systems of the National Academy of Sciences of Ukraine, Academician Glushkov Avenue, 40, Kyiv, 03187, Ukraine*

Abstract

Big data refers to large volumes, complex data sets with various autonomous sources, characterized by continuous growth. Data storage and data collection capabilities are now rapidly expanding in all fields of science and technology due to the rapid development of networks. Evaluating the quality of data is a difficult task in the context of big data, because the speed of semantic data reasoning directly depends on its quality. The appropriate strategies are necessary to evaluate and assess data quality according to the huge amount of data and its rapid generation. Managing a large volume of heterogeneous and distributed data requires defining and continuously updating metadata describing various aspects of data semantics and its quality, such as conformance to metadata schema, provenance, reliability, accuracy and other properties. The article examines the problem of evaluating the quality of big data in the semantic environment. The definition of big data and its semantics is given below and there is a short excursion on quality assessment. The model and its components which allow to form and specify metrics for quality have been developed. This model includes such components as: quality characteristics; quality metric; quality system; quality policy. A quality model for big data that defines the main components and requirements for data evaluation has already been proposed. In particular, such evaluation components as: accessibility, relevance, popularity, compliance with the standard, consistency, etc. are highlighted. The problem of inference complexity is demonstrated in the article. Approaches to improving fast semantic inference through materialization and division of the knowledge base into two components, which are expressed by different dialects of descriptive logic, are also considered below. The materialization of big data makes it possible to significantly speed up the processing of requests for information extraction. It is demonstrated how the quality of metadata affects materialization. The proposed model of the knowledge base allows increasing the qualitative indicators of the reasoning speed.

Keywords

Big data quality, semantic big data, reasoning optimization in the semantic big data

1. Introduction

The concept of Big Data in the broad sense of this word is used to define data processing, spread, and analytics [1]. The main special feature of this data is increased exponentially. Many efforts are aimed at solving the problem of big data, this is due to the need to develop new methods and algorithms for BD processing.

Defining big data is primarily related to the difficulty of defining a quantitative definition of a set of information objects. The most accepted definition is indicated in the report [2], where the problem of managing large data sets is based on the three Vs: Volume, Velocity, and Variety. They are expressed due to the growth of data volumes, the heterogeneity of data formats and metadata which make the rapid management of data more complicated. Later, such a criterion as Veracity [3] was

13th International Scientific and Practical Conference from Programming UkrPROGP'2022, October 11-12, 2022, Kyiv, Ukraine
EMAIL: alex.googl@gmail.com (A. 1)
ORCID: 0000-0002-9955-7882 (A. 1)

added to the definition of big data. This term was clarified and supplemented with criteria that affected the complexity and unstructuredness of the data [4], [5]. A number of big data definitions came from real business problems. However, we assume that the semantics and structure are given through external ontologies and fixed through metadata for semantic big data. We do not consider the problem of normalization and data extraction but evaluate the quality of such data. But this does not solve the problems of operating with such data and creates additional problems related to the reasoning of information from such a BD set. Our semantic data model must satisfy such requirements as Findable, Accessible, Interoperable and Reusable data or metadata [6].

2. Big Data Semantics

The issue of semantics was studied in works [7], where big data was considered on the basis that data semantics refers to the meaningful and effective use of a data object to represent a concept or object in the real world. Such a general concept unites a wide variety of applications [8]. Big Data semantic knowledge refers to numerous aspects of rules, expert knowledge and domain information [9]. One of the specific properties of big data in the semantic environment is the increasing complexity of reasoning even though this data not to big for the first view. Online web-application is very sensitive for delay for response and union approach reasoning and web technology provide high requirement to velocity big data. Our article surveys the problem of big data quality for web application and means for increasing velocity.

3. Model quality of Dig Data

The practical suitability of BD is determined primarily by its quality. The urgency of solving the BD quality problem is determined by the scale of its creation and distribution.

Let us consider the main concepts related to the quality of BD [10] some concepts was taken from the digital library domain and adapting to big data. Quality is a set of properties of objects that give them the ability to satisfy the stipulated or anticipated needs of the consumer following the purpose.

The quality characteristic is a property or a set of object properties, with the help of which quality can be described and evaluated. Each object has its nomenclature characteristic. A characteristic can be a composition of other characteristics, forming a hierarchical structure.

Metric is a formula or rule for determining the degree to which an object possesses a characteristic.

A quality indicator is a quantitative or qualitative value, obtained as a result of the procedure for evaluating the quality of a characteristic according to the evaluation methodology. Quantitative indicators have a numerical expression within a certain scale. Qualitative indicators have a verbal expression within a certain verbal ordered scale.

Quality level is the degree of acceptability of the obtained quality indicator from the view of the expected (planned) quality.

The quality system is a set of organizational structures, methods, processes, procedures and resources necessary for the general direction and management of quality by established methods. It includes quality policy, quality model; quality achievement system; quality system documentation.

The quality policy is a document developed by the responsible management. It expresses the goals in the quality field, the acceptable level of quality, the duties of various persons and structures for quality assurance, a set of measures to achieve quality. The quality policy is defined based on tasks set in the quality field.

Quality model is a set of objects for which it is described, evaluated and supported. Also, it includes quality characteristics, methods and means of quality assessment, metrics and algorithms for determining quality indicators. A specific quality model is selected based on the developed quality policy and other factors.

Achieving quality is a set of organizational structure, responsibilities, procedures, processes and resources that implement general quality management [11].

The quality management system is an organizational structure that includes personnel who implement quality management functions using established methods.

Quality management is the general management of quality provided by resources, particularly human resources. It organizes quality assurance work, interacts with the external environment, defines policies, goals and plans in the quality field, and makes strategic and important operational decisions regarding quality.

Also an quality assurance is creating confidence that quality requirements will be met. It includes administrative and procedural measures carried out within the framework of the quality system to ensure the fulfillment of requirements and goals. This is a systematic measurement, comparison with a standard, process monitoring, making technological or any other process adjustments to achieve the required quality.

Quality control is a set of measures, procedures, methods and means that allow performing a systematic and independent analysis. It is possible to determine the compliance of activities and results in the quality field with the planned measures and the effectiveness of their implementation and compliance with the set goals. The quality assurance system is the subject of the system analysis.

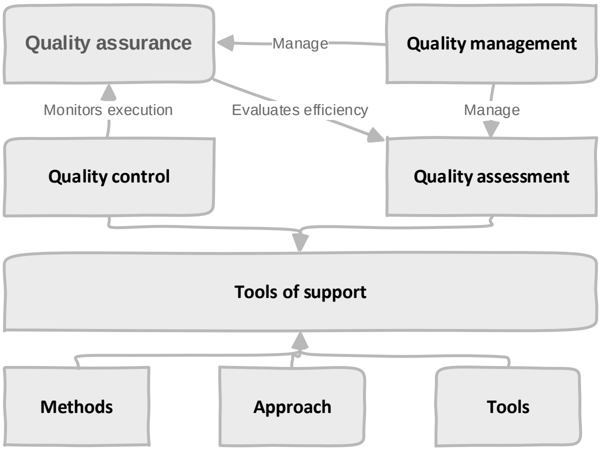

Figure 1: The system of quality achieving

Quality assessment measures the achieved or expected level of quality overall at every stage of the BD life cycle. There is a distinction between objective and subjective assessment. Objective assessment is a clearly defined assessment process, usually fixed by mathematical formulas, which does not depend on subjective perception. Subjective assessment is based on personal feelings, views and opinions.

We propose considering the main requirements for the quality model [12], which are also applied to BD.

A. The quality model should provide an opportunity to highlight the quality of the product itself and its interaction with the environment. The following components are distinguished in this context as:

- the quality of the product itself, without taking into account its behavior with the external environment (internal quality);
- product quality regarding its behavior in the external environment (external quality);
- the quality of technological processes of product development (process quality);
- the quality of the product to its use in different contexts (and the quality experienced by the user in specific scenarios of product use (quality during use)).

B. The quality model should include all stages of the BD development and use life cycle starting from requirements development and ending with the industrial operation.

C. The quality model is relevant to all structural elements of BD. It contains all types of support for the software system — functional, informational, mathematical, technical, etc.

D. An important component of the quality model is the structure of quality characteristics and metrics that assess elementary characteristics.

BD consist of two components are data and data base application, information is retrieved from a computerized BD by using a computer program.

The semantic information model for BD defines as a set of information objects in which each predicate define through top-level ontology.

Each information IO object in the BD environment is defined as a certain directed acyclic graph where the information object consists of a list of statements in the triplet «subject - predicate - object». The set of such triplets forms a directed graph, in which vertices are subjects and objects, and edges are predicates. Certain metadata describes each node of such a graph. That is, the model of the information object in the BD environment is defined as $IO = \left(s(m), p(m), o(m)\right)$.

Evaluating the quality of elementary characteristics involves determining their metrics represented by formulas or rules for determining the degree to which an object has an elementary characteristic [13]. The metric of an elementary characteristic reflects the degree to which an object or a set of objects possesses a certain property. Let a set of equivalent objects $M = \{M_i\}$ where ($i = 1,...,N$), be given, which may or may not have a certain property. We define the following characteristic function:

$$c(M_i, p) = \begin{cases} 1, & object\ M_i\ has\ property\ p; \\ 0, & another\ case. \end{cases} \quad . \tag{1}$$

Then the estimate of the degree to which the set of objects M has the property p is equal to:

$$M(p) = \frac{\sum_{j=1}^{N} c\left(M_j, p\right)}{N} . \tag{2}$$

If the objects M_i ($i = 1,...,N$) are unequal and their weighting factor $K_i : 0 \le K_i \le 1$ ($i = 1,...,N$), is given for each of them, which determines the relative importance of the objects, then the above formula takes the following form:

$$M(p) = \frac{\sum_{J=1}^{N} K_j \cdot c\left(M_j, p\right)}{N} . \tag{3}$$

Similarly, a metric can be defined for a situation where one object can have multiple properties and it is necessary to determine to what extent they are inherent to the object.

Establishing acceptable values for certain characteristics and adding a qualitative measure to the appropriate range is important for metrics. This range can be determined experimentally or algorithmically. An expert establishes it in many cases. For example, let's imagine as j an expert with K_j competence specifying a range of values $\left[X_{ij}, Y_{ij}\right]$ for the i characteristics, where Y_{ij} - the optimal value of the characteristic is X_{ij} - its worst value.

M experts evaluated the characteristics. The final score for the range of values is calculated as follows:

$$X_i = \frac{\sum_{j=1}^{M} K_j \cdot X_{ij}}{\sum_{j=1}^{M} K_j} \quad Y_i = \frac{\sum_{j=1}^{M} K_j \cdot Y_{ij}}{\sum_{j=1}^{M} K_j} . \tag{4}$$

It should be noted that intervals $\left[X_{ij}, Y_{ij}\right]$ are set by experts or determined algorithmically only for elementary characteristics. At other levels, i.e. for integral characteristics, the minimum and maximum values are calculated according to the defined formulas based on the given or calculated values of the previous levels [13].

4. Quality properties of information objects in Big Data

Next, the issues of evaluating the quality of semantic information objects are considered. IO quality characteristics.

Accessibility is a complex function that depends on many factors, including:

- the IO is actually available in the BD (the information object may be in the BD, but for some reasons, it may be removed from public access or due to the amount of data, it may not be identified among a set of objects);
- there is a service that can find the IO (one of the ways to remove an information object from public access is to deactivate its searching characteristics);
- it is the network and data transmission system in the network operational;
- there are no restrictions on access to the IO or if there are such restrictions they do not apply to specific persons or groups of persons.

It should be noted that in the given context, they talk about the availability of the IO to perform a single operation as reading. Our review does not include other possible operations with IO (changes, deletion, administration).

For BD this is availability provided by a specific service that interacts with BD. As a rule, a distinction is made between availability for all and certain services. In this case, the restriction of access rights $Acc\left(S_i, IO_j\right)$ where the S_i service for IO_j, means a function that acquires the following values: 1 — the service does not have access restrictions or it belongs to the group to which access is open; 0 — otherwise.

Now, if we mark other availability indicators as P_i except for access rights restrictions which take the following values: 1 — the indicator is satisfied, 0 — the indicator is not satisfied, then the general availability formula is calculated as follows:

$$MIN\left(P_1,\ \ldots, P_n, Acc\left(S_i, IO_j\right)\right). \qquad (5)$$

Relevance is the measure to which the information content of the information object meets the information needs of the user. Both cannot be strictly formalized. This assessment largely depends on the depth of the user's knowledge about their information needs at the current time and the tasks facing them. The user's information needs at the current moment are expressed through his information search query as a result of knowledge reasoning. The query implicitly defines the context in which relevance is evaluated. The user carries out an evaluation of this compliance as a result of receiving a response to the request (the user can be a group of people).

The relevance evaluation function is as follows $Relevance\left(IO_i, S_j, Query_k\right)$:

$$Relevance\left(IO_i, S_j, Query_k\right) = \begin{cases} 1 & - \quad \text{Servise } \mathrm{S}_j \ \textit{appove that } \ \mathrm{IO}_i, \textit{is relevant for Query}_k \\ 0 & - \quad \text{another case} \end{cases} \qquad (6)$$

Accuracy of storage. In the process of existence, the object can go into different states caused by the transition to other software and technology platforms. Big data is characterized by constant changes, and errors in these data also tend to accumulate and scale, including changing the storage format, using newer versions of BD, etc. All this can lead to a loss of storage accuracy of the new version of the information object compared to the old one. This characteristic requires assessing the loss degree of storage accuracy based on comparing states in the dynamic environment that in general is a complex task and required additional research [13].

Credibility means that the IO has the ability to confirm that it is what it should be. The ability to verify and measure the extent to which an IO is what it is claimed to be is fundamentally important in its correct perception and use. Reliability determines the extent to which the IO can be relied upon. This is largely determined by the developer's credibility and origin source. The credibility of the IO can be measured by:

- the attitude of users towards the IO itself;
- the attitude of users towards the source of the IO;
- the availability of information on the chronology of IO changes;
- the attitude of users to the BD in which the IO is located.

Integrity determines to what extent the IO is complete and correct from the point of view of the software object it represents. Integrity contributes to increasing trust in the IO [14]. Accuracy of reproduction determines the degree of accuracy of the reproduction of the IO of its original. For example, a text document reproducing an ancient book can accurately reproduce the text and completely ignore its artistic design.

Timeliness indicates that the IO is introduced and updated on time, as this issue is specific to BD. This characteristic evaluates how quickly the set $s(m), p(m), o(m)$ in IO is updated compared to the real state of affairs.

The characteristic is measured by the ratio of the actual delay time compared to the permissible one:

$$Timeliness\big(IO(s,p,o)\big) = \frac{real\,time\,delay}{\exp ected\,time\,delay}. \qquad (7)$$

Origin is a characteristic of the quality of an IO. It indicates how well (correctly, completely, qualitatively) the entire prehistory of the origin and change of an IO is presented, and how accurately and during what period it is possible to trace the prehistory of the existence of an IO. This is an important characteristic since inference over semantic data depends on the data itself. Understanding the historical information about the data helps to determine the reasons for changing the system's behavior, which is not a trivial task in the BD environment.

Susceptibility indicates how easily a person can understand and accept IO. It can be used to analyze which set of IO is most easily perceived by a group of persons due to the solved tasks.

Practical aspects of assessment of the quality of BD. One of the most challenging tasks in achieving data quality metrics is the early detection of data-related problems. Typical problems include completeness, the integrity of data and lack of contradictions. The problem lies in that in the conditions of the BD, the time to detect such issues may exceed the time requirements for receiving a response to the information from the BD. That is why it is necessary to develop methods that will allow the detection of such problems at an early stage. There are various approaches to deal with the task, like the way to control all data entered into the system through the ontology. In practice, it is often not known what the data model should be since the requirements for the BD system can change as the data increases. These requirements can be constantly updated. This means that data previously entered into the BD management environment in the previously specified structure may not correspond to the quality model after some time. Identifying these problems due to the scale is a difficult problem.

One of the criteria of the quality model is the ability of BD to give a quick response to user requests. The most effective method of increasing such speed is materialization [15]. Materialization can be used to improve performance at query time by making the required information explicit in advance. Thus, recalculation of the necessary information for each separate request is avoided. However, this method can be ineffective if there is excessive materialization.

Consider a certain graph of semantic data G in which the connections between concepts are built on the basis of descriptive logic. We will briefly describe the DL, which is the basis for all DL of the family. ALC means «Attributive Language with Complements». It is defined in [16]. The language is based on the previously introduced language AL (Attributive Language), to which the addition constructor (negation) was added. Syntax describes a set of correctly constructed language expressions, and semantics indicates their formal meaning.

Let $CN = \{A_1, \ldots, A_m\}$ i $RN = \{R_1, \ldots, R_n\}$ be finite, non-empty sets of atomic concepts and atomic roles. The ALC syntax is defined as follows:

- M and L are concepts;
- an arbitrary atomic concept A is a concept;
- if C is an arbitrary concept, then ШC , ChD and CgD are concepts, corresponding constructors are called addition, intersection and union;
- if C is a concept, R is an atomic role, then j $R.C$ and i $R.C$ are arbitrary concepts.

ALC semantics is defined through the concept of interpretation. An interpretation is a pair of I = (D, g^I) where Δ – is a non-empty set, called the domain of interpretation, a^I is an interpreting function that assigns the relation A^I8 Δ to each atomic concept A and to each atomic role R as an binary relation R^{I}8 ΔЧΔ. Other formulas are interpreted as follows:

$$M^I = \Delta,\ L^I = ;\tag{8}$$

$$(\mathrm{y}A)^I = \mathrm{D} \setminus A^I,\ (C\mathrm{h}D)^I = C^I 1 D^I,\ (C\mathrm{g}D)^I = C^I 2 D^I\tag{9}$$

$$\mathrm{j}\, R.C = \{a 9 \mathrm{D}\ |\mathrm{j}\, b 9 \mathrm{D}\ ((a,b) 9 R^I\ \mathrm{o}\, b 9 C^I)\}\tag{10}$$

$$\mathrm{i}\ R.C = \{a 9 \mathrm{D}\ |\mathrm{i}\ b 9 \mathrm{D}\ ((a,b) 9 R^I\ ®\ b 9 C^I)\}\tag{11}$$

Next the essence of the ($TBox$) terminology is revealed for DL ALC. However, all introduced concepts are easily transferred to other DL. Terminologies describe general knowledge about concepts and roles. To describe knowledge about specific individuals (their belonging to concepts and roles), the DL offers a system of facts about individuals or $ABox$. For this, a set of names of individuals is entered into the DL. There are two types of facts: a statement about an individual's belonging to a concept (written as $C(a)$); the statement about the belonging of a pair of individuals a and b to role (written as $R(a,b)$).

A system of facts or $ABox$ is a finite set of statements of form $C(a)$ and $R(a,b)$, where a and b are individuals, C is an arbitrary concept and R is a role.

Here are some ALC extensions that were used to fulfill the tasks of the dissertation work.

R-follower is an individual who is the right part of the role R. We denote the set of R-followers for e that can be written as $R^I(e)$, where $e\ \mathrm{O}\,\mathrm{D} : R^I(e) = \{d\ \mathrm{O}\,\mathrm{D}\ |\ (e,d)\,\mathrm{O}\, R^I\}$. We denote the power of such a set by $|\mathrm{R}^I(e)|$. The following constructors are called numerical role constraints. If R is a concept, n and 0 is a natural number, then:

- $\mathrm{J}\ 1R$ is a concept for limitation of functionality;
- $\mathrm{J}\ nR$ and $\mathrm{i}\ nR$ is a concept for quantitative limitation;
- $\mathrm{J}\ nR.C$ and $\mathrm{i}\ nR.C$ is a concept for qualitative limitation.

The following constructors are interpreted as follows:

$$(\mathrm{J}\ 1R)^I = \{e\ \mathrm{O}\,\mathrm{D}\ |\ |R^I(e)|\ \mathrm{J}\ 1\},\tag{12}$$

$$(\mathrm{J}\ nR)^I = \{e\ \mathrm{O}\,\mathrm{D}\ |\ |R^I(e)|\ \mathrm{J}\ n\},\tag{13}$$

$$(\mathrm{i}\ nR)^I = \{e\ \mathrm{O}\,\mathrm{D}\ |\ |R^I(e)|\ \mathrm{i}\ n\},\tag{14}$$

$$(\mathrm{J}\ nR.C)^I = \{e\ \mathrm{O}\,\mathrm{D}\ |\ |R^I(e)\, 3\, C^I|\ \mathrm{J}\ n\},\tag{15}$$

$$(\mathrm{i}\ nR.C)^I = \{e\ \mathrm{O}\,\mathrm{D}\ |\ |R^I(e) 1 C^I|\ \mathrm{i}\ n\}.\tag{16}$$

There are cases when it is necessary to describe specific characteristics of an object In order to describe the real world, for example, the number of pages in an information resource. To solve this problem, a specific area with a fixed set of predicates is created [17]. A concrete domain is a pair $\mathrm{D} = (D, \Phi)$, where D is a non-empty set and Φ is a set of predicates in D. It can be assumed that given a set of predicate symbols PN where each predicate symbol $P\ \mathrm{O}\, PN$ is associated with an n-arity and Φ maps an n- relation to it as $P^{\mathrm{D}} 8\, D^n$. It should be noted that Φ always contains a single predicate D, that is PN always includes M symbol and is interpreted as $M^{\mathrm{D}} = D$. Also is always closed with respect to the complement, that is for every n-predicate symbol P in PN there is an n-predicate symbol in P, which is interpreted as $\mathrm{D}^n \setminus P^{\mathrm{D}}$.

Let be a given concrete area D with a set of predicate symbols PN. Also let a finite set of symbols be given: CN are atomic concepts, RN are atomic roles, $AF\text{8}RN$ are atomic abstract attributes, CF are atomic concrete attributes. A sequence of $f_1\text{j}\ f_k h$ з k і 1 with atomic abstract attributes f_i ОAF and one concrete attribute h ОCF will be called a complex, concrete attribute.

Concepts of $\mathrm{ALC}(\mathrm{D})$ logic are defined by grammar [17]:

$$\text{M|L|}\, A\ \text{|y}C\text{h}D\ \text{|}C\text{g}D\ \text{|j}\, R.C\ \text{|i}\ R.C\ \text{|j}\ \overset{\text{Й}}{\text{Л}}u_1\text{j}\ \ u_n\overset{\text{Ш}}{\text{Ы}}P \tag{17}$$

where A ОCN, R ОRN, $u_1,...,u_n$ are arbitrary attributes, P ОPN is the n-concrete predicate. The semantics of $\mathrm{ALC}(\mathrm{D})$ logic is considered as $I = (\mathrm{D},\text{g}^I)$ interpretation with the following additions:

- sets Δ and D must not intersect;
- each atomic abstract attribute f ОAF is assigned a partial function f^I : D ® D ;
- each atomic abstract attribute h ОCF is assigned a partial function f^I : D ® D.

A composite concrete attribute $u = f_1\text{j}\ f_k h$ is interpreted as a composition of partial functions $u^I(x) = h^I(f_k^I(\text{j}\ \ f_1^I(x)\text{j}\ \ )$. As a result, a partial function u^I : D ® D is formed.

The only new (compared to) type of concept is interpreted as follows:

$$\begin{gathered}(\text{j}\ \overset{\text{Й}}{\text{Л}}u_1\text{j}\ \ u_n\overset{\text{Ш}}{\text{Ы}}P)^I = \{e\ \text{О}\,\mathrm{D}\ \text{|j}\, x_1\text{j}\ \ x_n\ \text{О}D : u^I{}_1(e) = \\ x_1\text{oj}\ \ \text{o}u^I{}_n(e) = x_n\text{o}\{x_1,\text{j}\ \ ,x_n\}\text{О}P^D\}\end{gathered} \quad . \tag{18}$$

The set of points on which the attribute u is defined is expressed by the concept φu.M, where M is a specific predicate that is always present in the PN signature. The following equivalence is valid:

$$\text{y}\$\overset{\text{Й}}{\text{Л}}u_1,\text{j}\ \ ,u_n\overset{\text{Ш}}{\text{Ы}}P\ \text{є}\ \text{y}\$u_1.\text{Mgj}\ \ \text{h}\,\text{y}\$u_1.\text{Mg}\$\overset{\text{Й}}{\text{Л}}u_1,\text{j}\ \ ,u_n\overset{\text{Ш}}{\text{Ы}}\text{y}P. \tag{19}$$

Indeed, the condition e О$(\text{yj}\ \overset{\text{Й}}{\text{Л}}u_1,\text{j}\ \ ,u_n\overset{\text{Ш}}{\text{Ы}}P)^I$ means that either one of functions u_i^I is undefined at point e or the tuple $\langle u_1{}^I(e),\text{j}\ \ ,u_n{}^I(e)\rangle$ does not belong to the predicate P^{D}, P, but belongs to its complement $(\text{y}P)^{\mathrm{D}}$. So, the G graph we have is given by BD

$$\begin{gathered}\text{M|L|}\, A\ \text{|y}C\text{h}D\ \text{|}C\text{g}D\ \text{|j}\, R.C\ \text{|i}\ R.C\ \text{|J}\ \ 1R\ | \\ \text{J}\ \ nR\ \text{|i}\ \ nR\ \text{|J}\ \ nR.C\ \text{|i}\ \ nR.C\ \text{|j}\ \overset{\text{Й}}{\text{Л}}u_1\text{j}\ \ u_n\overset{\text{Ш}}{\text{Ы}}P\ .\end{gathered} \tag{20}$$

When building a materialization, rules are set according to which it should be built. Consider the problem of excessive materialization, which can be caused by the following way of constructing concepts. For example, let's take the computer components motherboard and RAM. The concept that will determine the compatibility of these two components will be defined as follows:

$$\begin{gathered}RamDDR4\_2\_MaimboardDDR4\ \text{є} \\ (\text{"}hasSlotType.DDR4\text{g}Memory)\text{h}(\text{J}\ \ 1hasSlotType.DDR4\text{g}Mainboard)\end{gathered} \tag{21}$$

Table 1
Example of configuration

RAM	Main Board	Slots
Ram Model 1 DDR4	MainBoard Model 1 DDR 4	2
Ram Model 2 DDR4	MainBoard Model 2 DDR 4	4

As a result of the materialization, we will get the next G graph that will be set $C_6^2 = 15$ possible combinations that will determine the concept $RamDDR4\_2\_MaimboardDDR4$. If we take into account that the motherboard also has limitations in terms of supporting the maximum size of RAM and the real situation will become even more complicated.

Table 2

Example of configuration with combination

RAM	RAM Size	Main Board	RAM Slots	Max Memory support
Ram Model 1 DDR4	32	MainBoard Model 1 DDR 4	2	32
Ram Model 2 DDR4	12	MainBoard Model 2 DDR 4	4	128

Such dependence means that even with a small number of components, the knowledge base representation system will have to store a huge number of relationships that will determine the materialization. Accordingly, the inference on such a graph will work very slowly due to the huge number of combinations that form nodes of the graph available for search, as stated in [17], such an inference problem belongs to the P Space class. This means that the complexity depends on the size of the input data and to solve the problems of inference and feasibility of concepts, it is necessary to reduce the set of input data. To avoid such a problem, it is proposed to divide the knowledge base, which traditionally consists of TBox and ABox into two components, so that the subject area is described DL $SHIFT$ and then $ALC(D)$ (Fig. 2)

•

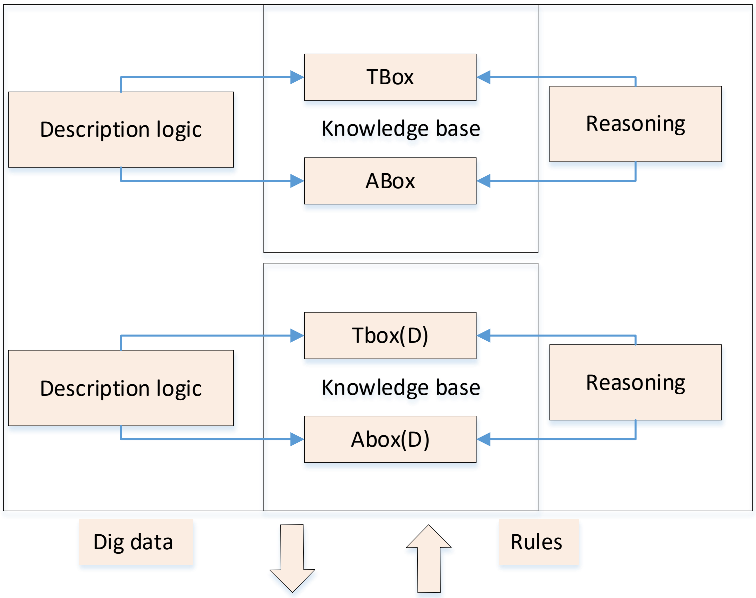

Figure 2: Knowledge base with separation

Thus, the knowledge base consists of two *TBox* T, T^D and two *ABox* A, A^D $K^D = (T, A, T^D, A^D)$. An I interpretation satisfies in K^D if I QT, T^D and I QA, A^D, in this case is K^D is called executable and the I interpretation is called a K^D model and written as I QK^D.

5. Estimate quality of metadata and an information object family in Big Data

Metadata quality assessment is intended to find out to what extent certain metadata or metadata schemas present in a BD meet the tasks that were set before the BD when it was designed. They contribute to the quality functioning of the semantics in the BD. The quality of metadata affects many processes related to the use of inference, building connections between the IO description, their input, storage, identification, search and access.

There are two aspects of quality related to metadata. The first of them refers to IO metadata (what IO metadata is, how fully it describes IO, whether it meets a certain metadata schema standard).

The second aspect is related to the schema of metadata (is the schema of metadata standard, to what extent the chosen schema meets the needs of the description of IS in a specific subject area). The quality of both aspects is described below.

Compliance with the standard. This characteristic indicates whether a standard IO metadata description scheme is used. The use of a standard metadata scheme is a fundamental issue in the consideration of the problem of the organization of search and retrieval of knowledge. The existence of IOs in the DB, the metadata of which do not meet or do not fully meet the standard, significantly reduces the resolution of fundamentally important issues facing the DB and reduces its quality. The measure of compliance with the standard can be the ratio of the number of non-standard metadata to the total number of metadata used in the description of the IO [13]:

$$Standard\left(IO\right) = 1 - \frac{w(IO(md))}{n(IO(md))}\,, \qquad (22)$$

where $n(IO(md))$ is the total number of IO metadata, a is the number of metadata that does not meet the standard adopted for this BD model

The *completeness* of the description of the IO in relation to the metadata scheme. This characteristic indicates the extent to which the metadata schema is fully used to describe the IO. Please note that not all metadata of the selected scheme can be applied to some types of IOs. Several metadata schemes can be used simultaneously in the DB network, but the completeness is determined relative to only those metadata that participate in the construction of semantic links between IOs.

Therefore, the degree of completeness of the description of the IO, according to the selected MS metadata scheme, is determined as follows:

$$Completeness\left(IO, MS\right) = \frac{\mathrm{Present(IO(md))}}{\mathrm{Required(IO(md))}}\,, \qquad (23)$$

where: md is metadata, MS is metadata schema, $\mathrm{Present(md)}$ is the total number of metadata required to describe the IO, which is actually present in the IO description, $\mathrm{Required(IO(md))}$ is the total number of MS metadata required to describe the IO.

Compliance with metadata schema. A metadata schema can set certain properties to its metadata. The characteristic of matching the metadata scheme determines how well the properties of the metadata of the IO correspond to the properties of the corresponding metadata of the selected scheme. Such properties include the type of data or attributes of relations between IOs, which in general are also included in the quality model.

Let n is the number of metadata in the MS scheme, m_i – is the number of properties of metadata md_i, $Conformance\left(i,j\right)$ is compliance of property j of metadata md_i IO with the standard specification of the MS schema. $Conformance\left(i,j\right)$ is calculated by the formula:

$$Conformance\left(i,j\right) = \begin{cases} 1 - \; iff\ i -\ metadata\ property\ belong\ to\ j -\ property\ from\ MS \\ 0 - \; \mathrm{otherwise} \end{cases} \qquad (24)$$

The correspondence of the IO to the i MS schema metadata is calculated according to the formula:

$$Conformance\left(md_i\right) = \frac{\sum_{j=1}^{m_i} Conformance\left(i,j\right)}{m_i} \qquad (25)$$

Then the correspondence to the *Conformance(MS)* metadata schema is calculated using the formula:

$$Conformance\left(MS\right) = \frac{\sum_{i=1}^{n} Conformance\left(md_i\right)}{n} \qquad (26)$$

Metadata scheme quality characteristics. A set of specially selected metadata make up a metadata schema. In the general case, such a set can be arbitrary, but this significantly reduces the quality of the BD, because our BD environment becomes isolated from other data sets and will not be able to take (at least fully) in the process of integration and reasoning information, in a sense the system becomes isolated because even using mappings between data schemas will be inefficient due to the scale of the data. In this regard, efforts are being made to develop and use standard metadata schemas, which are usually aimed at describing IOs of a certain class. There are many metadata schemes. In this connection, the question of choosing the most suitable for a certain subject area arises. This task is facilitated by the evaluation of the quality of the metadata scheme.

Compliance with standard metadata schema. This characteristic evaluates the extent to which all DB information objects conform to the standard. For IO, the characteristic of compliance with the standard is also significant, but it is at the IO level. In general, compliance with the standard scheme is evaluated as the arithmetic mean of compliance with the IO standard

$$Standard\left(MS\right)=\frac{\sum_{i=1}^{n} Standard\left(IO_i\right)}{n}. \qquad (27)$$

The completeness (usage) of the metadata scheme. This characteristic provides an opportunity to assess how much a certain scheme is used to describe the entire population of BD IOs. It is based on the characteristic of the completeness of the description of the IO in relation to the metadata scheme and is its arithmetic average for all IOs of the BD:

$$Completeness\left(MS\right)=\frac{\sum_{i=1}^{n} Completeness\left(IO_i, MS\right)}{n}. \qquad (28)$$

This characteristic makes it possible to assess to what extent the decision to use a certain metadata scheme is justified, and, if necessary, to make a decision to replace it.

Let's introduce metrics for evaluating the IO family. A family is a systematized set of IOs that are united into a single whole based on some meaningful or formal criteria of belonging, for example, regarding the general content, sources, purpose, semantic independence, method of use, etc.

Completeness of the family of IO. This characteristic establishes to what degree of completeness the family contains those IOs that it should contain. Completeness can be measured only when it is known what exactly the collection should contain, that is, when the original family, which acts as a sample, is known [13]. As a rule, families are distinguished on the basis of IO attributes.

The formula for measuring family completeness is as follows:

$$Completeness\left(F\right)=\frac{\sum_{i=1}^{n} IO_i(F)}{\sum_{i=1}^{n} IO_i(F_{original})}. \qquad (29)$$

Conformity of the collection to the standard. Determines the extent to which collection IOs conform to the standard. Compliance with the standard of the family can be considered as the arithmetic average of compliance with the standard of its IO:

$$Standard\left(F\right)=\frac{\sum_{i=1}^{n} Standard\left(IO_i(F)\right)}{n}, \qquad (30)$$

where n is the number of IOs in the collection

A variety of standards. It is believed that the family should be based on one standard metadata scheme specified in the external ontology, as the use of many schemes deteriorates the operational characteristics. The quality of this feature can be measured as the inverse of the number of metadata schema standards used.

Consistency. There are many different situations where a collection can be considered inconsistent (conflicting). For non-limiting generalizations, we consider only one situation when there are two IOs with absolutely identical values of their metadata.

Let the function $IdentMd\left(IO_i, IO_j\right)$ acquire the following values:

$$IdentMd\left(IO_i, IO_j\right) = \begin{cases} 1 - & IO_i \text{ and } IO_j \text{ have the same set of metadata} \\ 0 - & \text{otherwise} \end{cases}. \qquad (31)$$

Then the family matching function is defined as follows:

$$Consistency\left(F\right) = 1 - \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \frac{IdentMd\left(IO_i, IO_j\right)}{n \cdot \left(n - 1\right)}. \qquad (32)$$

For modeling our approach, we are using Neo4j as a system for storing and managing big data [18], [19]. Neo4j is a database whose data model is a graph, specifically a property graph. We took a database for electronic components consisting of boxes, main boards, and memory modules. Our goal is to find all available interpretations which will be models for our knowledge base. It means the need to find all compatible components or find a list of components that are compatible with the selected. This problem more detail describe in [20], [21], [22]. As specified in these works the quality of the result depends on the quality of metadata. And another important characteristic for semantic networks is the speed of reasoning for checking interpretation. It is related to time which needs to get answers about the compatibility of electronic components.

The metrics of quality data are allowing us to reveal a problem with missing required metadata for interconnecting components. Due to this information and metrics like compliance with metadata schema as a result of cleaning data, we built graph storage which consists of 44195 relations, we don't have any nodes without missing important data. This graph has a relation between memories, main boards, and cases. At first look, this graph does not belong to big data but if we take only 54 different types of memories, 113 types of mainboards, and 119 types of cases the result of materialization gives 246912 available combinations for our system. This materialization is not included in the concrete domain. Materialization in the concrete domain will bring an enormous quantity of available nodes because if we have for example attribute which describes the count of ram slots on main board it allows putting on these slots a different combination of memory modules. Our optimization also includes checking only bi-directional dependencies between components.

Our idea to split the knowledge database into two-part brings the possibility of extracting information from a database with materialization without a concrete domain.

We are build relation in our graph that it responsibility to DL $SHIFT$ the main condition for building relation avoid concrete domain. On Fig. 3 demonstrate relation between our components.

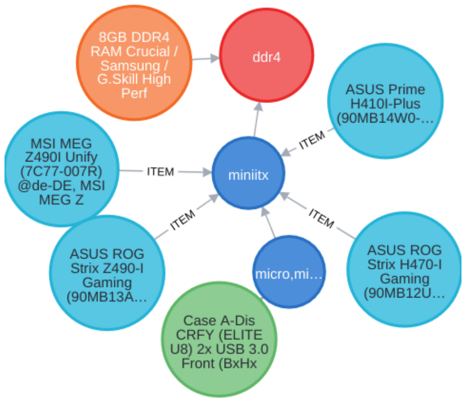
8GB DDR4 RAM Crucial / Samsung / G.Skill High Perf
ddr4
ASUS Prime H410I-Plus (90MB14W0-…
MSI MEG Z490I Unify (7C77-007R) @de-DE, MSI MEG Z
ITEM
miniitx
ITEM
ITEM
ITEM
ASUS ROG Strix Z490-I Gaming (90MB13A…
micro,mi…
ASUS ROG Strix H470-I Gaming (90MB12U…
Case A-Dis CRFY (ELITE U8) 2x USB 3.0 Front (BxHx

Figure 3: Knowledge base with separation

Three approaches were tested on the test data set. The first time q_1 when relation were built taking into account all possible variations, including the quantities of the selected components.

The second approach q_2 consisted in grouping components by common value of attributes in such a way as to avoid building additional connections. And the last optimization q_3 consisted in the fact that first all compatible components were searched, and only then the conditions of quantitative restrictions for a concrete domain were checked for satisfaction.

One problem is that the same component can be reinstalled twice or more depending on the number of previously selected components. That is, if the motherboard has 8 RAM sockets, then there may be a situation when 8 identical memory modules are selected, and there may be 8 different modules. Moreover, for the motherboard, we must check not only the quantitative limitation of the number of occupied sockets, but also the limitation regarding the maximum amount of memory supported by the motherboard

Table 3
Example of configuration with combination

Type optimization and query	Execution time	Count results
q_1 list all mainboards	5612 ms	6850
q_1 list all mainboards for the specific memory modules	4630 ms	6432
q_2 list all mainboards (specification was grouped)	3400 ms	6850
q_2 list all mainboards for the specific memory modules (specification was grouped)	2530 ms	6432
q_3 list all main boards (specification was grouped and quantity restriction included) time for two query	780 ms	6850
q_3 list all main boards for the specific memory modules (specification was grouped and quantity restriction included) time for two query	43 ms	6432

As we can see, the simplification of requests gives a significant increase in the speed of execution. But result BD systems depend on characteristics such as the completeness of the description, compliance with the metadata scheme. It should be noted that according to the expert evaluation of work with web resources, the response of the web service should be up to 600 ms.

6. Conclusions

The complexity of big data applications combined with the lack of standards for the representation of information objects, processing and storage requires significant resources. Data quality is one of the approaches that will allow achieving modeling of data that will require simpler algorithms for analysis. Analysis of data quality allows increasing their accuracy in various aspects. Enrich data semantics is a complex process of describing big data by ontological means. However, there is a problem with the speed of inference, the article proposes a method of knowledge base materialization in the environment of big data to optimize inference. The quality of the data plays a key role in this, allowing to build of appropriate graphs of schematic data on the basis of metadata.

Higher data quality levels can help produce better reasoning results but also help improve data maintainability and reusability and integration.

7. References

[1] J. Stuart Ward and A. Barker, "Undefined By Data: A Survey of Big Data Definitions" 2013.

[2] D. Laney, «3D data management: Controlling data volume, velocity and variety» META group, 2001.

[3] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales and P. Tufano, "Analytics: The Real-World Use of Big Data" IBM, 2012.

[4] I. C. Intel IT Center, "Centre. Big Data Analytics: Intel’s IT Manager Survey on How Organizations Are Using Big Data" Santa Clara, 2012.

[5] S. Suthaharan, «Big data classification: Problems and challenges in network intrusion prediction with machine learning» ACM SIGMETRICS Performance Evaluation Review, т. 41, № 4, pp. 70-73, 2014.

[6] M. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, L. da Silva Santos, P. Bourne and J. Bouwman, "The FAIR Guiding Principles for scientific data management and stewardship" Scientific data, pp. 1-9, 2016.

[7] P. Ceravolo, A. Azzini, M. Angelini, T. Catarci, P. Cudré-Mauroux, E. Damiani, A. Mazak, M. Van Keulen, M. Jarrar, G. Santucci and K. Sattler, "Big data semantics" Journal on Data Semantics, vol. 7, no. 2, pp. 65-85, 2018.

[8] R. Amsler, "Application of Citation-based Automatic Classification» Austin, 1972.

[9] W. A. Woods, «What's in a link: Foundations for semantic networks.» Representation and understanding, pp. 35-82, 1975.

[10] O. Novytskyi, G. Y. Proskudina, V. Reznichenko та O. Ovdiy, «Evaluation of the quality of electronic libraries in the web environment» Software engineering, т. 20, № 4, 2014.

[11] O. Novytskyi, G. Proskudina та O. Ovdiy, «Development of an digital library quality model» в Інформація, комунікація, суспільство 2014 : матеріали 3-ої Міжнародної наукової конференції ІКС-2014, 21–24 травня 2014 року, Україна, Львів, Славське, 2014.

[12] O. M. Spirin, S. M. Ivanova, O. V. Novytskyi, Z. V. Savchenko, V. A. Reznichenko, A. V. Yatsyshyn, N. M. Andriychuk, V. A. Tkachenko, M. A. Shinenko and Y. A. Labzhynskyi, Collective monograph. Electronic library information systems of scientific and educational institutions, В. Ю. Биков and О. М. Спірін, Eds., Kyiv: Pedagogical press, 2012, p. 176.

[13] A. Novitsky, V. Reznychenko та E. Romanov, «Characteristics and quality metrics of electronic libraries in the semantic web» Software engineering, т. 1, № 25, pp. 17-36, 2016.

[14] Kudym, K.A., Novitsky, O.V., Proskudyna, G.Y. та Reznychenko, V.A., «Statistics on the use of the scientific electronic library of periodical publications of NAS of Ukraine» Science of Ukraine in the global information space, т. 10, pp. 60-67, 2014.

[15] G. Antoniou, S. Batsakis, R. Mutharaju, J. Z. Pan, G. Qi, I. Tachmazidis, J. Urbani та Z. Zhou, «A survey of large-scale reasoning on the Web of data» The Knowledge Engineering Review, т. 33, 2018.

[16] F. Baader, D. Calvanese, D. McGuinness, P. Patel-Schneider та D. Nardi, The description logic handbook: Theory, implementation and applications, Cambridge university press, 2003.

[17] C. Lutz, «The Complexity of Description Logics with Concrete Domains» Hamburg, 2002.

[18] J. J. Miller, «Graph database applications and concepts with Neo4j» In Proceedings of the southern association for information systems conference, т. 2324, № 36, 2013.

[19] P. Shi, G. Fan, S. Li та D. Kou, «Big Data Storage Technology for Smart Distribution Grid Based on Neo4j Graph Database» IEEE 4th International Conference on Electronics Technology (ICET), pp. 441-445, 2021.

[20] A. Trentin, E. Perin та C. Forza, «Product configurator impact on product quality» International Journal of Production Economics, т. 135, № 2, pp. 850-859, 2012.

[21] [21] B. Thorsten, A. Nizar, G. Kreutler та F. Gerhard, «Product Configuration Systems: State of the Art, Conceptualization and Extensions» в Génie logiciel & Intelligence artificielle. Eight Maghrebian Conference on Software Engineering and Artificial Intelligence (MCSEAI 2004), Munich, 2004.

[22] Y. Wang, Z. Wenlong та X. W. Wayne, «Needs-based product configurator design for mass customization using hierarchical attention network» IEEE Transactions on Automation Science and Engineering, т. 18, № 1, pp. 195-204, 2020.