A generalized methodology for the construction of predictive analysis systems as exemplified by the mining equipment in the BIG DATA environment

Andrey Kupin
Department of Computer Systems and Networks
Kryvyi Rih National University
Vitaliya Matusevycha str. 11, Kryvyi Rih, Ukraine
kupin.andrew@gmail.com

Rodion Ivchenko
Department of Automation Computer Science and Technologies
Kryvyi Rih National University
Vitaliya Matusevycha str. 11, Kryvyi Rih, Ukraine
ivchenko.ra@gmail.com

***Abstract* – It is necessary to determine the optimal methodology for the system of predictive analysis of equipment to prevent emergency situations. The system may include, in particular: data input/reading from sensors, processing/storage of information in a database using algorithms for processing Big Data and decision trees [1]. Identifying possible types of problems and making decisions on how to respond to them; training the system for more accurate response and decision-making.**

Keywords—Industry 4.0; Big Data; Predictive analytics.

I. Introduction

The issue of introducing smart management of technical maintenance services at large industrial enterprises on the basis of new approaches within the framework of the modern concept of Industry 4.0 is being considered.

The current trends in the scientific and technological progress of the global industry are quite often described in such terms as “Smart Plant”, “Smart Production”, “Smart Factory” and “Factory of the Future”. At present, the development of those research areas is quite well illustrated by the concept of the 4th industrial revolution (Industry 4.0). The implementation of that concept involves the use of some key technological trends, such as Big Data processing, cyber-physical systems, autonomous robots with various smart sensors, simulators for 2D and/or 3D modeling, 3D printers, Internet things, augmented reality, etc. Consequently, according to the estimates of the world's leading experts, those trends will determine the main vector of the modern competitive industries [1].

The predictive analytics is a class of data analysis methods that focuses on predicting the future behavior of objects and subjects in order to make optimal decisions.

The predictive analytics uses statistical methods, smart data analysis methods, analyzes current and historical facts to compile predictions about future events. In business, predictive models use patterns found in historical data and the data being acted upon in order to identify risks and opportunities. Models capture relationships among many factors to make possible assessing risks or the potential associated with a particular set of conditions thereby guiding decision making.

II. Problem Statement

Another task for the predictive algorithms is equipment maintenance and repair. Companies mainly use basic control mechanisms provided by equipment manufacturers. But the potential of those means is limited, because they do not make it possible to analyze additional factors influencing the state of the equipment, and to predict a critical situation in advance. So, the maintenance department employees receive a lot of data, but they do not know how the items of that data are related to each other. As a result, the response from the repair services follows only after the actual equipment failure, which leads to downtime and, consequently, additional costs. The predictive analytics conducts, by means of machine learning and artificial intelligence, continuous analysis of Big Data, performs data visualization on the current state of the equipment, and predicts scenarios for the occurrence of equipment failures. As a result, unplanned downtime is reduced, equipment maintenance and repair work are optimized, maintenance time is decreased, and the management personnel receive an in-depth analysis of the causes of equipment failures.

Modern IoT and Big Data capabilities, along with the advanced predictive analytics methods, are becoming an efficient tool to reduce costs, improve product quality and increase company productivity. The predictive analytics has become a new trend of modern times that opens up broad prospects for the further development of companies. In addition to manufacturing industry, forecast platforms are successfully used in banking, insurance, retail, logistics, marketing and many other areas. Thus, banks and other financial institutions use the predictive analytics tools to forecast their customers' behavior, e. g. to find out whether they will acquire a new banking product or whether they will continue to cooperate with the bank. Exploring their historical behavior in conjunction with external factors, mathematical algorithms can help make profitable personal offers, anticipate sales increase, customer churn, etc. In marketing, mathematical modeling makes it possible to segment customers by certain characteristics, predict their future behavior (indicate customer value, calculate the likelihood of another purchase, etc.) and determine services and products that may be of interest to the customers. Predictive models accurately determine the efficient channels of interaction with the target groups. Analyzing social media data helps marketing researchers better understand customers and potential buyers. In retail business and logistics, predictive models help forecast demand, optimize inventories and analyze the risks of delays in deliveries [2].

When embarking on the introduction of a predictive analytics system, it is necessary to keep in mind that such a system cannot work without a large amount of the historical data and continuous collection of the current data. The less data is involved, the less accurate the predictions will be. When promoting predictive analysis tools, companies often face the problem of data shortage, which becomes an insurmountable barrier to forecasting.

It is therefore necessary to start collecting data immediately. Otherwise the company risks not sustaining rivalry with more technologically equipped competitors. Organizing stable high-quality data collection and storage is the primary task that must be accomplished in the near future if the company intends to remain on the market.

III. DESCRIPTION OF CONTROL OBJECTS AND THEIR MATHEMATICAL MODELS

Usually, concentrating plants are classified according to the nature of the concentrating processes used (flotation, gravitational, magnetic ones) or according to the raw material being processed (coal, ore, etc.). For any type of concentrating plants, the concentrating process technology is determined by a large and diverse number of factors influencing the said process.

The complexity of analyzing the mineral dressing technology as an object of control makes it natural and necessary to observe for its study a certain hierarchical structure. The distinctive feature of such a structure is the sequential division of the technical system into subsystems, between which the relationships of subordination are established. The hierarchical structure of ore concentrating plants from the standpoint of control tasks can be represented in the form of three stages [3]:

The lower stage is the so-called typical processes of the mineral dressing technology (crushing, grinding, separation, etc.);

The middle stage is a group of processes or pieces of equipment that perform an independent technological process task for the production of a given product, e. g. the flotation area of an ore concentrating plant; grinding, classification and magnetic separation stages of magnetic concentration plants; ore concentrate drying area, etc.;

The upper stage is a technological line for dressing a particular mineral raw material considered as a whole.

Each hierarchy of mineral dressing technology is characterized by its own control tasks that involve their respective functions, which actually determines the type of a mathematical model describing the process of functioning of a given object.

In general, the mathematical model of each stage of the hierarchy can be viewed as a complex object of the mineral dressing technology and presented as a function of variables. Here, three types of actions serve as the input of the object:

1) uncontrolled (but monitored) input variables Y = $\{y_1,...y_r\}$ constitute a disturbance vector and, as a rule, characterize, as far as concentrating production is concerned, quality indicators of the source material to be processed and those of its intermediate products obtained during the concentrating process;

2) controlled input variables U = $\{u_1,...,u_n\}$ constitute a control vector and characterize, as a rule, quantitative indicators (expenditure) of material and energy flows;

3) the uncontrolled factors Z = $\{z_1,...,z_k\}$ constitute an interference vector. Basically, this is a disturbance vector, about which the developer of the control system knows very little or nothing at all. Most often that vector is not taken into account at all.

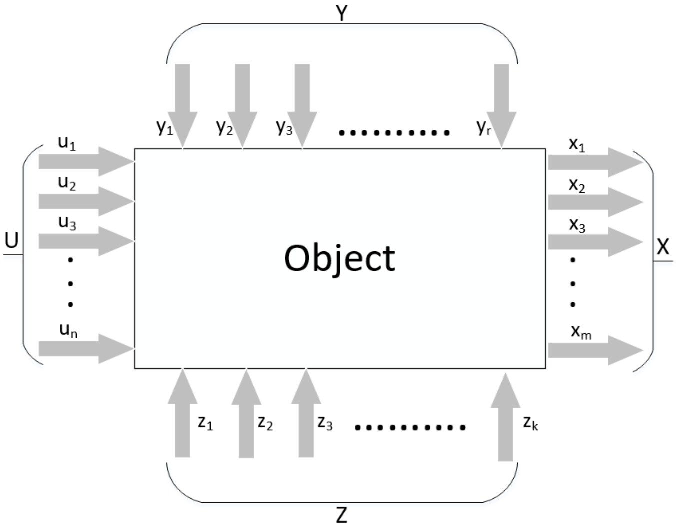

Fig. 1 – The structure of a complex control object

The essence of any technological process, including the process of concentrating mineral raw materials, is in converting the У, U, Z input actions into the X = $\{x_1,...,x_m\}$ output ones. The X vector is called the state vector, and characterizes, as far as concentrating production is concerned, qualitative and quantitative features of the output products (concentrate, tailings).

The objects of the concentrating process can be conveniently represented in the form of models describing typical processing operations, as far as the distribution of flows of mineral raw materials or intermediate products is concerned, resulting in a quantitative and qualitative change in the parameters of those flows. In total, one can identify four typical operations that characterize the mineral dressing technology: simple, mixing, separating, and combination ones.

The mixing operation is characterized by the presence of the "n" process flows Y = {y_1,...y_n} at the input of the object and by one "x" output flow. The operation equation looks like this:

$$x = f(Y, U) \quad (1)$$

A mixing operation can be exemplified by a ball mill used before the flotation operation. At its entrance, the mill is fed by the flows of ore Q_r (t/h), water W_v (m^3/h), and reactants W_r, i. e. Y = {Q_r, W_v, W_r}, while at its exit, a flow "x" of the slurry from finely ground particles of ore is formed. The control actions can represented by the consumption Q_b (t/h) of balls for the mill, ore consumption (Q_r (t/h), and water consumption W_v (m^3/h), i. e. U = {Q_r, W_v, W_r}.

IV. SENSOR TYPES

A sensor is an item of a measuring, signaling, regulating or controlling device that converts a monitored parameter (temperature, pressure, frequency, luminous intensity, voltage, current, etc.) into a signal convenient for measuring, transmitting, storing, processing, recording, and sometimes for influencing processes being controlled. Or more simply put, a sensor is a device that converts the input action of any physical parameter into a signal convenient for further use.

The sensors used are very diverse, and can be classified according to various criteria [4]:

Depending on the type of the input (measured) parameter, the following sensors are identified: those of mechanical displacements (linear and angular); pneumatic and electric ones; flow meters; speed sensors; those of acceleration, force, temperature, pressure, etc.

Currently, there is approximately the following distribution of the shares of measurements of various physical parameters in the industry: temperature – 50%, consumption (by mass and by volume) – 15%, pressure – 10%, level – 5%, quantity (mass, volume) – 5%, time – 4%, electrical and magnetic parameters – less than 4%.

According to the type of the output, into which the input is converted, non-electrical and electrical sensors are distinguished: direct current sensors (respective EMF or voltage sensors), alternating current amplitude sensors (respective EMF or voltage sensors), alternating current frequency sensors (respective EMF or voltage sensors), resistance sensors (effective resistance sensors, inductance sensors or capacitance sensors), etc.

Most sensors are electric. This is due to the following advantages of electrical measurements:

- electrical parameters are convenient to transmit over a distance, with the transmission being carried out at a high speed;
- electrical parameters are universal in the sense that any other parameters can be converted into electrical ones and vice versa;
- electrical parameters can be accurately converted into a digital code, and make it possible to attain high accuracy, sensitivity and operation speed of measuring instruments

By the operating principle, the sensors can be divided into two classes: generator and parametric ones (modulator sensors). The generator sensors (transducers) directly convert the input parameter into an electrical signal.

Parametric sensors convert an input parameter into the change in some electrical parameter (R, L, or C) of the sensor.

According to the operating principle, the sensors can also be divided into ohmic, rheostat, photoelectric (optical-electronic), inductive, capacitive ones, etc.

Three classes of sensors are identified:

- analog sensors, i. e. sensors that produce an analog signal that is proportional to the change in the input parameter;
- digital sensors generating a sequence of pulses or a binary word;
- binary sensors that produce a signal of only two levels: "on/off" (in other words, "0" or "1"); they are widespread due to their simplicity

Requirements for sensors:

- non-ambiguous one-to-one relationship between the output parameter and the input one;
- performance stability over time;
- high sensitivity;
- small size and weight;
- no counter-effects on the process and the parameter being controlled;
- operation under various conditions;
- various mounting options

Parametric sensors (modulator sensors) convert the X input parameter into a change in some electrical parameter (R, L or C) of the sensor. It is impossible to transmit over a distance a change in the above sensor parameters without an electric signal (voltage or current). It is possible to identify the change in the corresponding sensor parameter only by the sensor's response to current or voltage, since it is the above parameters that characterize that response. Therefore, parametric sensors require the use of special measuring circuits supplied by direct or alternating current.

Ohmic (resistance) sensors – their operating principle is based on the change in their effective resistance when the length L, the sectional area S or the resistivity *p* change:

$$R= pl/S \quad (2)$$

In addition, the relationship between the effective resistance value and the contact pressure and the illumination of the photocells is used. In accordance with that, ohmic sensors are divided into contact, potentiometric (rheostat), strain gauge, thermistor, and photo-resistor ones.

Contact sensors are the simplest type of resistor sensors that convert the movement of a primary element into an abrupt change in the resistance of an electrical circuit. With the help of contact sensors, forces, displacements, temperature, dimensions and shapes of objects, etc. are measured and controlled. The contact sensors include position and limit switches, contact thermometers and so-called electrode sensors that are mainly used to measure the critical levels of electrically conducting fluids.

Contact sensors can operate using both direct and alternating current. Depending on the measurement range, contact sensors can be single-range and multi-range ones. The latter are used to measure values that vary significantly, with the parts of resistor R incorporated into the electrical circuit being successively short-circuited.

The disadvantage of contact sensors is the complexity of continuous control and a limited service life of the contact system. But owing to the extreme simplicity of those sensors, they are widely used in automation systems.

Rheostat sensors are actually resistors with varying effective resistance. The input parameter of the sensor is the displacement of the contact, while the output one is the change in its resistance value. The moving contact is mechanically connected with the object, whose displacement (angular or linear) needs to be converted.

The most widespread is the potentiometric circuit for operating a rheostat sensor, into which the rheostat is incorporated according to the voltage divider circuit. It shall be reminded that the voltage divider is an electrical device for dividing DC or AC voltage into parts; the voltage divider makes it possible to take up (to use) only a portion of the available voltage through the items of an electrical circuit consisting of resistors, capacitors or inductors. A variable resistor included in the voltage divider circuit is called a potentiometer.

Usually, rheostat sensors are used in mechanical measuring devices to convert their readings into electrical parameters (current or voltage), e. g. in float level meters for liquids, in various pressure gauges, etc.

A sensor in the form of a simple rheostat is almost not used due to the considerable nonlinearity of its static characteristic line $I_n = f(x)$, where I_n is the current in the load.

The output parameter of such a sensor is the voltage drop U_{out} between the movable contact and one of the fixed ones. The relationship between the output voltage $U_{out} = f(x)$ and the *x* displacement of the contact corresponds to the regularity of the electrical resistance change along the potentiometer. The regularity of the resistance distribution over the length of the potentiometer that is determined by the design thereof can be either linear or nonlinear.

Potentiometric sensors, that are constructively variable resistors, are made of various materials, viz. winding wire, metal films, semiconductors, etc.

Strain gauge resistors (strain gauge transducers) are used to measure mechanical stresses, small deformations, and vibration. The operating principle of strain gauges is based on the strain-resistive effect, which is actually the change in the effective resistance of conductor and semiconductor materials under the influence of forces applied to them.

Thermometric sensors (thermistors) – their resistance depends on temperature. Thermistors, as sensors, are used in two ways:

1) The temperature of the thermal resistor is determined by the environment; the current passing through the thermal resistor is so small that it does not cause heating-up thereof. In view of that, the thermal resistor is used as a temperature sensor and is often called a "resistance thermometer".

2) The temperature of the thermal resistor is determined by the degree of heating by constant-value current and cooling conditions. In this case, the steady-state temperature is determined by the heat transfer conditions of the thermal resistor surface (the speed of the medium – gas or liquid – relative to the thermal resistor, the density of the medium, its viscosity and temperature); the thermal resistor can therefore be used as a flow rate sensor, thermal conductivity sensor of the medium, gas density sensor, etc. In sensors of such a type, something like a two-step conversion occurs: the parameter being measured is first converted into a change in the temperature of the thermal resistor, which is then converted into a change in the electrical resistance.

Thermal resistors are made both from pure metals and semiconductors. The material, from which such sensors are made, should have a high temperature resistance coefficient, a linear, if possible, relationship between the electrical resistance and the temperature, good reproducibility of properties and inertness to environmental influences. Platinum satisfies all the above properties to the highest degree, copper and nickel do that to a slightly lesser degree.

Semiconductor thermal resistors (thermistors) have a higher sensitivity as compared to metal ones.

Inductive sensors are used to receive information about the movements of the working parts of machines, mechanisms, robots, etc. in a contactless manner and to convert that information into an electrical signal.

The operating principle of an inductive sensor is based on a change in the magnetic core winding inductance relative to the position of the individual elements of the magnetic circuit (armature, core, etc.). In such sensors, linear or angular displacement X (input parameter) is converted into a change in inductance (L) of the sensor. Inductive sensors are used to measure angular and linear displacements, deformations, for dimensional control, etc.

In the simplest case, the inductive sensor is an inductance coil with a magnetic core, the movable element of which (armature) is displaced under the action of the parameter being measured.

An inductive sensor recognizes all electrically conductive objects and responds accordingly. An inductive sensor is contactless, does not require mechanical action, and operates in a contactless manner by utilizing changes in the electromagnetic field.

Advantages:

- no mechanical wear and tear, no failures related to the state of the mechanical contacts
- no contact rattling sound and false response
- high switching frequency of up to 3000 Hz
- resistant to mechanical effects

The disadvantages of inductive sensors are the relatively low sensitivity, the relationship between the induction and the frequency of the supply voltage, a significant reverse effect of the sensor on the parameter being measured (due to the attraction of the armature to the core).

Capacitive sensors – their operating is based on the relationship between the capacitor capacitance and the size and the relative position of its plates, as well as the permittivity of the medium between the plates.

For a double-plate flat capacitor, the capacitance is determined by the following formula:

$$C = e_0 eS/h \qquad (3)$$

where e_0 – the dielectric constant; e – the relative permittivity of the medium between the plates; S – the effective area of the plates; h – the distance between the plates of the capacitor.

The C(S) and C (h) relationships are used to convert mechanical displacements into changes in capacitance.

Capacitive sensors, as well as inductive ones, are supplied with alternating voltage (usually of heightened frequency – up to dozens of megahertz). As measuring circuits, bridge circuits and resonant ones are usually used. In the latter case, the relationship between the generator oscillation frequency and the capacitance of the resonant circuit is as a rule used, i. e. the sensor has a frequency output.

The advantages of capacitive sensors are their simplicity, high sensitivity and low response time. Their disadvantages are their propensity to the influence of external electric fields and the relative complexity of measuring devices, into which they are incorporated.

Capacitive sensors are used to measure angular displacements, very small linear ones, vibrations, speeds of movement, etc., as well as to reproduce given functions (harmonic, saw-tooth, rectangular ones, etc.).

Capacitive transducers, whose permittivity *e* changes relative to a displacement, deformation or changes in the composition of the dielectric medium, are used as level sensors of non-conducting fluids, bulk and powdery materials, thickness sensors of the layer of non-conducting materials (thickness gauges), as well as control devices for the humidity and composition of a substance.

V. SUBSEQUENT INFORMATION PROCESSING

Interfaces control the data flow. A clear definition of the required data and its connection to the network data world are the basis for reliable data exchange over the network. That being the case, an important role is played by the choice of the correct data transfer protocols for a given section of the path. The Ethernet-based solutions are at the forefront. But IO-Link also makes connection to the network possible, especially for devices that need only a reduced data transfer capability.

In smart manufacturing, many sensors collect a large amount of data in many places. As a result, the importance of decentralized data processing is rising. Additional interfaces in the data or software systems make possible new analyses and functions, help improve the flexibility, quality, efficiency and transparency of the production.

Through the application field-oriented connection technologies and bypassing the control system, the sensor data can in the future be directed immediately into the cloud.

Thanks to the successful and comprehensive integration of all sensors into a network with centralized or decentralized data processing systems, a previously unknown number of solutions appears, while the whole process becomes transparent due to data transfer and communication protocols.

BIG DATA methods make it possible to handle structured and unstructured data of very large volumes to obtain results that are efficient under the conditions of continuous growth and distributed over the nodes of the computing network. The need for such methods is caused by the ever-increasing development of technological processes and the equipment itself at the companies. Those methods can be used to collect information from sensors for the predictive analyses and data processing. They are also used to improve security and modularity (e. g. for the purpose of preventing equipment breakdowns).

The known technologies include [5]:

1) Recognition of graphic elements as a new part of speech recognition implementation.
2) Adaptation operations and related automated vehicles.
3) Semi-automatic flexible machines for additional services.
4) Fully automated quality assurance for adapting to rapid changes in demand.
5) Smart automatic control of objects for greater efficiency.
6) Enhancing security and modularity.

Enhancing the safety of production is of prime importance. The safety improvement is therefore possible mainly due to the predictive analysis to prevent emergency situations.

In recent years, SAP Predictive Analytics [6] has focused on the development of machine learning, Big Data processing and the IoT development. Those are the three most important technological areas, where the company is developing its solutions. SAP works not only on the development of a tool, but also on the application of the respective technologies in practice. If you have a large number of customers, automating your business processes with the use of SAP products allows you to analyze customer needs in a comprehensive way and offers you new approaches in utilizing client data to increase the efficiency of those processes.

You can download a temporary series into a data analysis tool, but without data pre-processing, the resulting model will be of poor quality. When preparing data, it is necessary to fulfill two stages of processing. The first stage is Data Engineering, i. e. collecting, understanding, clearing and primary data processing. The second stage is Feature Engineering: the formation of descriptive data features that contain information on various aspects of the behavior of the object, whose model is under construction. In terms of the CRISP-DM [2] methodology, those stages are similar to Data Understanding and Data Preparation (cf. Fig. 2).

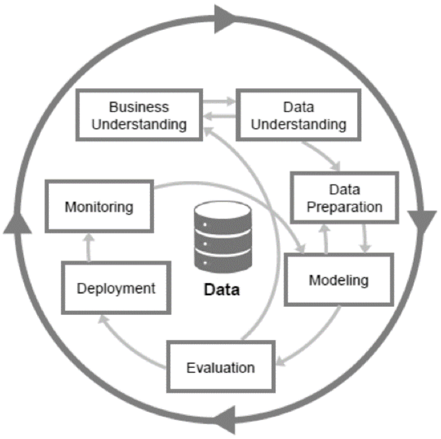

Fig. 2 – Stages in the CRISP-DM methodology with the possible transition paths between the stages [7]

An integral part of the initial stages of the machine learning process is the feature selection, i. e. the selection of variables on the basis of which the model undergoes the process of learning. The selection can be done using various tools, and also be dependent on many factors, e. g. the correlation of features with the target variable or the data quality. The next (and more important) step may be the creation of new features based on the already existing ones, the so-called feature engineering. That operation can allow engineers to improve the quality of the model, while at the same time making it possible to obtain a more complete explanation of the data, if the model is interpreted. In our case, the first step in constructing a SAP Predictive Analytics model was to create new features using the built-in Data Manager solution.

In the data set prepared, there are indicators that affect the target variable at the given moment of time. However, you can get additional information if you determine the impact of those indicators during a certain period up to the given moment. In our case, the following time intervals were chosen: during 1 hour and 1 day until the given moment of time. Even more informative may be the degree of the indicator change from the moment in the past until the given moment. As part of the method, the natural logarithm of the quotient of the current indicators and those with an interval of 1 hour and 2 days was selected. Consequently, we managed to obtain the degree of the indicator change from the moment in the past (whether it had increased or decreased and if so, then to what extent).

All modern techniques of working with Big Data invariably follow the three principles given below. In order to comply with them, it is necessary to use methods, techniques and paradigms of data processing. One of the most established methods is called MapReduce.

MapReduce is a distributed data processing model offered by Google for processing large volumes of data in computer clusters. MapReduce:

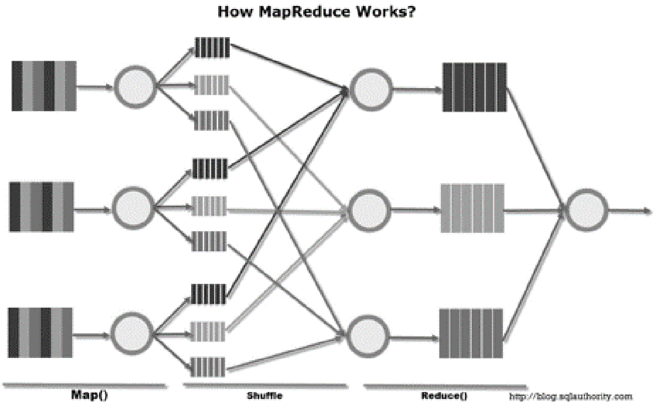

Fig. 3 The MapReduce model [8]

MapReduce assumes that the data is organized in the form of some entries. Data processing takes place in 3 stages:

1. The "Map" stage. At this stage, the data is converted using the "map ()" function, which is defined by the user. The role of this stage lies in the preliminary processing and filtration of the data. This operation is very similar to the "map" operation in functional programming languages. This user-intended function is applied to each input section.

   The "map ()" function is applied to one input entry pair and returns a set of key-value pairs. The word "set" means that it may return only one entry, may not return anything, and may return several key-value pairs. What is in the key and in the value is defined by the user, but the key is a very important thing, since the data with one key will later on find its way into one instance of the "reduce" function.

2. The "Shuffle" stage. It goes unnoticed by the user. At this stage, the output of the "map" function is "distributed into baskets" – each basket corresponds to one output key of the "map" stage. Later on, those baskets will serve as inputs for the "reduce" function.

3. The "Reduce" stage. Each "basket" with the values formed at the "shuffle" stage finds its way into the "reduce ()" function.

The "reduce" function is determined by the user, and calculates the final result for a separate "basket". The set of all values returned by the "reduce ()" function is the final result of the MapReduce task.

VI. DEVELOPMENT OF KEY EFFICIENCY INDICATORS

The methodology and tools for working with structured data have long been created. This is a relational data model and database management systems. But under the present-day conditions, companies need to process large amounts of unstructured data of various types (Table 1), so the previous methods are not quite suitable for that task. New methods of handling data are needed. At present, the model of work with

Big Data implemented in the Apache Hadoop project is becoming increasingly popular [9].

TABLE I. PREVAILING TYPES OF INFORMATION FOR VARIOUS FIELDS OF ACTIVITY WITH THE DEGREE OF USE

Field of activity\Type of information	Video	Images	Audio	Text/Numbers
Banking sector	Medium	Medium	Medium	Medium
Insurance	Low	Low	Low	High
Securities and investments	Low	Low	Low	High
Manufacture	Medium	Medium	Low	High
Retail trade	Medium	Low	Low	High
Wholesale trade	Low	Low	Low	High
Professional services	Medium	Medium	Medium	High
Entertainment	Medium	Low	Medium	Medium
Healthcare	Low	High	Low	High
Transportation	Medium	Medium	Low	High
Mass media	High	Medium	High	High
Utilities	Medium	Medium	Low	High
Civil engineering	Low	High	Low	Medium
Resources	Medium	Medium	Low	High
Government	High	Medium	High	High
Education	High	Medium	High	Low

Most products for working with Big Data have a highly efficient system for processing huge amounts of information and analyzing it in real time. The expected effect of the Big Data implementation may vary depending on the type of activity and the actual policy of a particular company (Fig. 4). When working with Big Data, methods of knowledge manipulation are used: various methods of the theory of recognition and classification, methods of intelligence analysis and data generalization, smart approaches in the form of genetic algorithms, neural networks and other branches of artificial intelligence. The sources indicate the relationship between the expected effect of the Big Data implementation and the field of activity and policy direction of a particular company [9].

The main tasks of the Hadoop platform are data storage, processing and management. The main components of the Hadoop platform are:

- Failure-resistant Hadoop Distributed File System (HDFS), which is used for storage purposes;
- The Map Reduce software interface, which is the basis for writing applications that process in parallel large volumes of structured and unstructured data on a cluster of thousands of machines;
- Apache Hadoop YARN performing data management function.

This reflects the Google citation index (Fig. 4). The Hadoop platform makes it possible to reduce the time for data processing and preparing, expands the possibilities for analysis, enables one to handle new information and unstructured data.

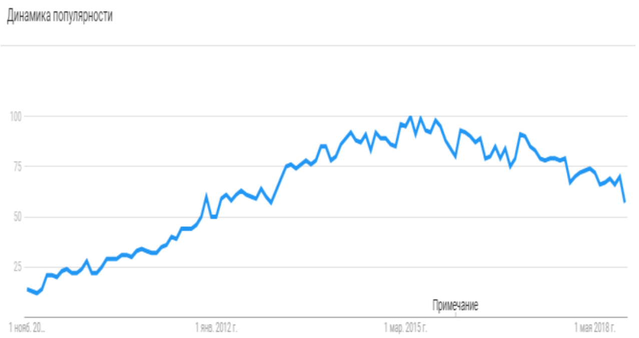

Fig. 4 Google citation index

The results of the project on the implementation of the Hadoop technology confirm the feasibility of its use (Table 2).

TABLE II. PROJECT RESULTS [9]

Platform	Description of the equipment	Approximate cost of the equipment, $	Average time for one report, min.
Oracle database	Hi End Class Server	300 thousand	59
Hadoop cluster	10 workstations	7 thousand	66
Optimized Hadoop cluster	10 workstations	8 thousand	40

Solutions based on the Hadoop technology have a number of significant advantages. The main ones are given in Table 3.

TABLE III. ADVANTAGES OF A SOLUTION BASED ON HADOOP [9]

Advantage	Short description
Reduced data processing time	When processing data using a cluster, it is possible to significantly reduce the time for data processing.
Reduced equipment cost	The use of the Hadoop technology makes it possible to reduce the cost of equipment required for data storage and processing, by dozens of times.
Increased failure resistance; the technology makes it possible to come up with a failure-resistant solution.	Failure of one or several cluster nodes affects only the system productivity, while the system as such continues to operate properly providing service to the end users.
Linear scalability	The solution makes it possible to increase productivity simply by adding new cluster nodes. With that said, cluster productivity increases linearly.
Unstructured data operation capability	The technology makes it possible to conduct complex processing of any files including unstructured ones, so that such data can be efficiently processed and used.

The Big Data storage solution is based on business requirements, workload management and smart storage ideas. As shown here, end users have many reporting possibilities and indicators that help them understand the use of tables, table spaces and workloads, since they relate to the frequency of access to data and its changes.

VII. Conclusion

Scientific methods aimed at introducing IT for processing large amounts of data with a distributed infrastructure based on smart agents and parallel algorithms have been examined. The emphasis is laid upon innovative methods based on smart agents and the principles of Industry 4.0. The implementation and simulation of parallel processing algorithms for Big Data and decision trees are being done.

The methods of operating BIG DATA have been formulated. The SAP Predictive Analytics modeling will be important in the process of collecting information from equipment sensors and in terms of the possibility of the follow-up analysis of the equipment, for example, for its proper functionality.

A review of technologies and business process control models has been carried out, during analysis of which recommendations were given that would be used for the predictive analysis in the BIG DATA environment.

References

[1] A. Kupin, I. Muzyka, R. Ivchenko: “Information Technologies of Processing Big Industrial Data and Decision-Making Methods”,Problems of Infocommunications. Science and Technology, 2018.

[2] “Predictive analytics features: case from Beltel Datanomics” [Electronic resource] – Available: https://iot.ru/promyshlennost/vozmozhnosti-prognoznoy-analitiki-keys-ot-beltel-datanomics

[3] A.M. Mariuta, Yu. G. Kachan, V. A. Bunko: “Automatic Control of Technological Processes of Concentrating Plants”. Moskow «Nedra». p 1983.

[4] “Electrotechnical Encyclopedia. Sensors” [Electronic resource] – Available: http://www.electrolibrary.info/subscribe/sub_16_datchiki.htm

[5] R. Ivchenko: “Technology predictive analysis based on IoT TA BIGDATA”. III International Scientific and Practical Conference "Information Security and Computer Technologies". Central Ukrainian National Technical University, Kropivnitsky, April 19-20, 2018.

[6] Magazine "Habrahabr". “How to predict the exchange rate of the ruble to the dollar using SAP Predictive Analytics” [Electronic resource] – Available: https://habrahabr.ru/company/sap/blog/345108/

[7] “Comparison of metadata editors”. [Electronic resource] – Available: https://en.wikipedia.org/wiki/Comparison_of_metadata_editors

[8] Magazine "Habrahabr". “Big Data from A to Z. Part 1: Principles of working with big data, the MapReduce paradigm”. [Electronic resource] – Available:: https://habrahabr.ru/company/dca/blog/267361/

[9] Ivanov P.D., Vampilovv V.Zh. “Big Data technologies and their application in a modern industrial enterprise.” Engineering Journal: Science and Innovation, 2014. [Electronic resource] – Available: http://engjournal.ru/catalog/it/asu/1228.html

[10] “Using big data in marketing research.” [Electronic resource] – Available: http://www.ovtr.ru/stati/bolshie-dannye-big-data-v-marketingovyh-issledovaniyah.