The research of social processes at the university using big data

Abdullayev Vugar Hacimahmud[1], *Ragimova Nazila Ali*[1] *Khalilov Matlab Etibar*[1]

[1]Azerbaijan State Oil and Industry University, Baku, Azerbaijan

Abstract. The volume of information in the 21st century is growing at a rapid pace. Big data technologies are used to process modern information. This article discusses the use of big data technologies to implement monitoring of social processes. Big data has its characteristics and principles, which reflect here. In addition, we also discussed big data applications in some areas. Particular attention in this article pays to the interactions of big data and sociology. For this, there consider digital sociology and computational social sciences. One of the main objects of study in sociology is social processes. The article shows the types of social processes and their monitoring. As an example, there is implemented monitoring of social processes at the university. There are used following technologies for the realization of social processes monitoring: products 1010data (1010edge, 1010connect, 1010reveal, 1010equities), products of Apache Software Foundation (Apache Hive, Apache Chukwa, Apache Hadoop, Apache Pig), MapReduce framework, language R, library Pandas, NoSQL, etc. Despite this, this article examines the use of the MapReduce model for social processes monitoring at the university.
Keywords. Big Data, Sociology, Social Process, Monitoring of Social Process, Hadoop, MapReduce, University.

1 Introduction

We live in an information age where the main products become information and knowledge in the economy. The beginning of the information age can consider the appearance of microprocessors and personal computers. Those who possess them dominate the economy. According to the report of the analytical company IDC "Digital Universe Study", the volume of digital data was 0.18 zettabytes in 2006, and a volume was 1.8 zettabytes in 2011. In 2020, the volume of stored data was about 55-60 zettabytes. By 2025, it's up to about 400 zettabytes. Accordingly, managing structured and unstructured data with modern technologies is an area that is becoming increasingly significant [1, 2]

Promptly growing data volume provokes the search for new means for their storage and processing. In this regard, the term "big data" enters.

Big data is structured and unstructured data that is incapable to process by traditional methods of data processing. Traditional processing methods include the following methods:

- The method of comparison is a scientific method of cognition, in the process of its unknown (studied) phenomenon or subjects are compared with previously known ones studied to determine common features or differences between them;
- The grouping method means dividing the entire population of data studied into qualitatively homogeneous groups according to the selected characteristics. This method is most often used to identify the influence of factors on a homogeneous group of different characteristics. Selection correct characteristics is the main task of grouping.
- Graphical method. Graphs are scale images of measures and their dependencies using geometric shapes. A graph is an image of a mathematical dependence in the form of a certain curve that characterizes the changes in a function when the argument (s) change;
- Balance method. The balance method of analysis is widespread in the practice of accounting and planning, analysis of the availability of the enterprise with labor, financial resources, raw materials, fuel, materials, basic means of production, etc., as well as in the analysis of the completeness of their use. It is used to identify and reflect the ratios, proportions of two groups of interrelated and balanced economic indicators, the results of which should be identical.

These methods can implement by tools of SQL in combination with one of the object-oriented programming languages. The problem with big data is that they present in weblogs, a video, GPS data, machine code, etc. that different from the structured DB format. [3]

The main characteristics of big data consider volume, velocity, and variety. It also accepts to refer veracity, value, viability, variability, visualization to them.

It is possible to formulate the following basic principles for working with big data from the above-mentioned characteristics:

- Horizontal scalability is the main principle of the analysis of big data. It means that with the increase in data volume, it's necessary to increase the number of computing nodes, without losing performance;

[1] Corresponding author: abdulvugar@mail.ru

- Fault tolerance. The number of computing nodes may grow, so the probability of their exit out of operation grows. Therefore, tools of big data should be ready for such situations and are capable to take corresponding measures;
- The locality of data. It is necessary to store and process data in the same physical server, otherwise the

expenses of transfer of data between servers can be enormous.

1.1 Big Data solutions

There is a set of various methods and tools for the processing of big data. The international consulting company McKinsey&Company selected several main methods of the analysis of big data.

Data mining methods (DM) are a set of methods for detection in data of earlier unknown, uncommon, practically useful knowledge that are necessary for decision-making. Treat the DM methods:

- Association rule learning. Set of methods for detection of associative rules in data bulks;
- Classification is the purpose of objects, observations, or events to one of the earlier announced classes;
- Cluster analysis is the statistical technique reference of objects to groups by the identification of the general sign;
- Regression is a set of statistical techniques for studying of influence of one or several independent variables on dependent, etc.

Crowdsourcing is a method that allows at the same time different users to make data collection from different sources which quantity is not limited.

Data fusion and data integration are a set of methods for the integration of heterogeneous data from different sources for carrying out over them the intellectual analysis. It is possible to give examples of methods of digital signal processing, natural language processing as an example, including the tone analysis.

Machine learning is a set of methods, the self-training pursuing the aim creation of algorithms based on empirical data. A distinctive feature of machine learning is not the direct solution of an objective, but training during the application of solutions of similar tasks.

Artificial neural networks are mathematical models, which construct based on biological neural networks.

Pattern recognition is a set of methods of classification and identification of images, characterized by a certain property set and signs.

Predictive analytics is a set of methods of data analysis, which concentrate on forecasting of possible behavior of objects for acceptance of optimal solutions.

Simulation modeling is a type of mathematical modeling for the creation of the model, which describes a real system virtually. It use generally for forecasting and carrying out experiments.

Spatial analysis is a set of methods of the analysis of spatial data.

Statistical analysis is a set of methods of collecting, organizing, and interpretation of big data.

Data visualization is a set of methods of data view in the form of diagrams, charts, the animated images, and animation.

The above-presented methods of processing big data are implemented in various tools of different vendors. These include products 1010data (1010edge, 1010connect, 1010reveal, 1010equities), products of Apache Software Foundation (Apache Hive, Apache Chukwa, Apache Hadoop, Apache Pig), MapReduce framework, language R, library Pandas, NoSQL, and others to tools of big data. [3]

Company 1010data offers services of data analysis for receiving more intelligent solutions and acceptance of more optimal solutions for business. For this purpose, it is necessary to place the data on the servers of the company. Products of this company are:

- 1010edge - the intuitive platform of corporate data analysis, the need for data supporting everything, and analytics. Its opportunities are the main providing collecting and management of data, the analysis, and modeling, creation of reports and visualization, application development, sharing and monetization of data;
- 1010connect – the detailed portal allowing to share and control the distribution of data outside the enterprise that allows creating a basis for unprecedented cooperation with business partners or to turn data into the differentiated highly profitable generator of regular income. Its main opportunities are providing own cloud, flexible management and management of permissions, multilevel access, and controlled distribution, safety, and high availability;
- 1010reveal – a set of intelligent solutions for consumers of high definition;
- 1010equities – this product provides alternative solutions for data transmission on the party of the buyer.

The most widespread tool for work with big data is Apache Hadoop that implement based on MapReduce. Apache Hadoop is a set of libraries and utilities for the development and execution of the distributed programs that work at the clusters capable consist of several thousand nodes. Hadoop consists of four modules:

- Hadoop Common is a set of libraries of management of file systems of Hadoop and scenarios of creation of the necessary infrastructure and management of the distributed processing;
- Hadoop Distributed File System – the file system for storage of files of large volumes;
- Yet Another Resource Negotiator – the module of resource management of clusters and planning of tasks;
- Hadoop MapReduce – a program framework for coding of distributed computing based on MapReduce.

Apache built the whole ecosystem around Hadoop that contains the whole set of powerful tools, the facilitating work with big data:

- Hive – the tool for the creation of HiveSQL of requests over big data. It is capable to turn normal SQL requests into a series of MapReduce-tasks;
- Pig – a programming language of requests to big semi-structured data which one line is capable to turn into the sequence of MapReduce-tasks;
- Hbase – a columnar DB which implements based on NoSQL;
- Apache Spark is the engine of the distributed data processing using the Hadoop components (HDFS and YARN);
- Apache Tez – a framework that works over Hadoop YARN for processing of group data, persons in need of integration with Hadoop YARN;

- Apache Solr – the instrument of faceted and full-text search, integration with a DB, dynamic clustering, and document handling with a difficult format;
- Apache Sqoop and Flume – instruments of data flow control;
- Zookeeper and Oozie – tools for coordinating and task scheduling;
- Apache Storm – the tool providing safety of stream data processing;
- Apache Kafka – the tool for fast processing of program messages between programs.

The Hadoop installation represented a quite difficult task earlier: it was necessary to configure each machine in a cluster. Due to the increase in popularity of an ecosystem of Hadoop, there were companies providing assemblies of an ecosystem of Hadoop and powerful tools for management of a Hadoop cluster. Select with Forrester Research a number of the Hadoop distribution kit: Cloudera, and MapR.

Cloudera is the first company and the leading supplier of Hadoop and possesses its Hadoop distribution kit called by Cloudera CDH. It provides software for gaining access, storage, analysis, protection, and management, and information search. Cloudera has its module of management of a cluster of Cloudera Manager and a quite high price of technical maintenance (≈ $4,000 a year for one node of a cluster), only big corporations can afford it.

The MapR distribution kit prefers the distributed file system to MapR-FS, own DB of MapR-DB and the unique distributed broker of program messages MapR Event Store instead of Apache Kafka uses. MapR provides a balance between stability and high-speed performance, saving at the same time the simplest uses.

In addition, there is the Russian ArenaData distribution kit, which is completely localized into Russian and based on open the Apache Software Foundation projects.

The Pandas library does Python as a powerful tool for analysis and data visualization. SQL, HTML, Excel files, text files can serve as a data source. The main structures of data storage in Pandas are:

- Series is the indexed one-dimensional array of values;
- DataFrame is the indexed multidimensional array of values in which column is the structure of Series.

R is a programming language for statistical data processing and work with graphics. A distinctive feature of this language is its big range of statistical and numerical methods and expansibility using packets, the libraries providing themselves for the work of special functions or scopes.

NoSQL has no accurate definition. In the general NoSQL – the term for designation of the approaches implementing DBMS different from relational DBMS. Unlike traditional DBMS, NoSQL has the following properties: basic availability, flexible status, coherence eventually. For work with big data of NoSQL uses the Family of Columns model. The systems using this model store data as disperse matrixes where lines and columns use as keys. Generally, this model is used in Apache HBase, Apache Cassandra, ScyllaDB, Apache Accumulo, Hypertable.

1.2 Big data application in some scientific fields

The application of big data has found the reflection in many aspects of our life, including in economy (Einav, Levin, 2013), in management and business (Frizzo-Barker et al., 2016), anthropology (Sivkov, 2017), history (Bearman, 2015) and in sociology. [4]

According to McKinsey&Company, there are five basic approaches to the use of big data in an economy:

1. Organization of "transparent" economy;
2. Acceptance of mathematically justified management solutions;
3. Narrow segmentation of clients taking into account personal provisions;
4. Increase in speed in decision-making thanks to difficult analytics;
5. Development of goods and services of the next generation. [5]

According to IDC, 90% of the data stored on servers of the companies is practically useful, but it is not suitable for use. Useful information for business in the company generally obtain from CRM and telephony (automatic telephone exchange). The CRM systems contain information on sales on territories, seasons, the sum, and the number of orders. The automatic telephone exchange contains data on waiting duration on the line, durations of a talk, and algorithms of recognition of the entering and outgoing calls, phone numbers. [6]

Big data is capable to automate and generalize functional approaches to the search and selection of employees, improving the quality of work of personnel and increase in labor productivity, a solution to tasks in the field of education of commands by a ratio of qualities of people in management. The main task of big data is the management of talents and improvement of trial and error methods of personnel in human resource management. [7]

In the first decade of the XXI century, one of the important anthropological problems was the problem of exploiting hidden and inaccessible communities or populations eroded by large populations. Thus, big data with little focus is another way to solve an important ethnographic problem, namely, the problem of identifying the boundaries of some unrepresented or poorly represented community that does not have spatial localization. The contextuality of big data analysis is one of the main requirements on the part of the social and humanities sciences.

The application of advanced information technologies in history forms two new areas of scientific research. These areas are historical informatics and digital history. Historical informatics is positioned as an interdisciplinary direction of historical research, having a balanced relationship of applied (resource) and analytical components (with an emphasis on the latter), and digital history refers to itself as a multidisciplinary field of digital humanities, which is dominated by philological sciences and is more associated with the resource component.

2 Sociology and Information Technology

Implementation of big data in sociology generates two types of sociology: computational sociology (computational social sciences) directed to collecting and data analysis; social information science (e-social sciences, digital social researches), intended for accumulation and information analysis.[8]

According to C. Cioffi-Revilla "computational social sciences" is the integrated cross-disciplinary search in social research through calculations at increasing the scale of information processes.

D. Watts considers the term "computational social sciences" as a label which the agent-based models describe simulations. These sciences, interactions that include the analysis of web and large-scale data of observation, virtual experiments in laboratory-style and computing modeling.

There also create the division of computing social sciences in Microsoft Corporation. When determining "computational social sciences" statistics is added, considering that it is the cross-disciplinary area attracting examination, large-scale statistical and techniques of machine learning, covering several independent social sciences including sociology, economy, psychology, political sciences, marketing when large-scale demographic, behavioral and network data for research of human activity and relationship prevail.

It is possible to draw the following conclusion following from above-mentioned definitions of this term:

Computational social sciences are cross-disciplinary mutual the intersections of computer and social sciences aimed at cross-disciplinary finding in social researches using examination, the methods of machine learning, economies, sociology, psychology, marketing, political sciences.

C. Cioffi-Revilla has proposed five research methods for "computational social sciences," though it should be noted that these methods are not limited to this field:

1. Automatic collecting of information. In "computing social sciences" the analysis of the text and the content analysis on monitoring of information on events and studying political the rhetorician is applied;
2. Analysis of social networks. It is used for the safety of social networks, analysis of people's opinions;
3. Geospatial analysis. Applying geographic information systems, researchers study space layers of distribution of the ideas;
4. Modeling of complexity. Applies mathematical means to the understanding of the interaction between elements in a system as well as definitions of the intensive conflicts;
5. The agent-based modeling. Using this type of modeling, researchers study changes in environments and emergence of the news organizations.

The reverse of "computational social sciences" is social information science. Social information science is the science applying the modern digital technologies directed to studying social problems and opportunities of application of social researches.

One of the main objects of studying sociology is the processes proceeding in society, called social processes. Disagreements between social groups of people, which some groups profit indifference with other social groups of people, are in fundamentals of social processes. Such a deal of things is the natural evolution of society. The formulating criterion of what is the temporary component. This component gives the character of the object, which gives the chance to trace all object properties depending on time. The temporary component is especially interesting in studying social-economic and political processes.

The social process is a change of social character in society, which is caused by the desire of certain groups to influence the current situations in society for the satisfaction of the interests. Sources of these processes are people.

Heterogeneity in positions of subjects of social society defines a vector of social processes that aim to reach a balance with each other. Such interaction of interests of communities has resulted from the actions of unknown forces caused directly by this interaction. The result of these actions sets the direction of social processes. [9] Subordination of social processes by subjects of society to the vector of behavior and option of probable actions is their priority task.

Despite a big variety of social processes, Robert Park and Ernest Burgess are sociologists of the Chicago school of sociology could classify them into six groups:

1. Cooperation is a process association of persons in groups, for the sake of the general interest with the purpose to receive the benefit. For this purpose, mutual respect and the establishment of rules of cooperation in-group are necessary. Sociology is having made observation they established that cooperation is the cornerstone the mercenary purposes;
2. The competition is a fight of subjects for receiving different resources (money, power, love, etc.). In essence, the competition is a fight for remunerations for what it is necessary to be ahead of the rival with the same purposes;
3. Adaptation is a process when there is an adoption of norms and values of the new environment by the individual at the dissatisfaction of norms and values of the current environment of needs of the individual. In this process select subordination, a compromise, and tolerance;
4. The conflict is the process demanding full subordination or open counteraction of change, occurring in society;
5. Assimilation is a process where a part of society loses a certain degree of the cultural lines and replaces them with loans from another part of society;
6. Amalgamation is a process that arises when mixing groups of subjects of society. The difference from assimilation is in what after completion of the process of amalgamation of an edge between groups erase.

There are also added two more processes to these processes: maintenance of borders and systematic communications. Borders between social groups are one of the main aspects of social life. For their maintenance,

establishment, and modification a lot of time and energy were selected. Ambits of social groups separate their members from all other societies. [10] Processes of assimilation and amalgamation follow by erasing of borders between social groups, destruction of the available division for creation of common features of the group.

Systematic communications are processes that link establishment between social groups that are in the set social borders. The absence of any group of communication with other groups will lead to its isolation.

Monitoring of social processes is a set of methods of collecting and processing social processes for the detection of patterns of development of society and each individual separately. Objects of social monitoring are all set of social phenomena and processes. [11]

At this moment histories monitoring of social processes carried out by the analysis of social networks and a wide area network of Internet, then direct observation of an object. Social processes are capable to influence economic development and the political situation of the region or even the countries. Having results of monitoring, experts are capable to influence these processes.

Monitoring of social processes is the difficult process defining not only specifically the purposes, but also it implements based on certain principles:

- Completeness of social information;
- The efficiency of data collection;
- Comparability of the obtained data;
- A combination of the generalized and differential estimates and outputs in the course of social monitoring.

Tools of big data can be useful to the execution of functions and problems of monitoring social processes. Applying these means, an opportunity to collect information on an object from social networks, unstructured and semi-structured databases appears. Then use of different tools of big data allows defining processes in society. Further to carry out monitoring of social processes. Afterward, there is an opportunity to define to positive or negative changes will lead to social processes and to apply the appropriate measures.

All this allows watching objects and the phenomena, to reveal trends of development, forecasting of possible effects, to run for the search of necessary measures for prevention of negative trends and maintain positive, leading society to further development.

3 METHODS OF RESEARCH

The data stored at the university in documents of various kinds are historical in nature. This means that these documents almost never change over time. Therefore, the use of MapReduce and Hadoop is suitable for this example.

MapReduce is a distributed computing framework developed by Google, used to reliably perform parallel big data computing on large clusters up to several terabytes in size. The main advantage of MapReduce technology is the ease of scalability of data processing on several clusters. Each of these clusters can consist of several computers, if the number of computers changes, you should simply change the configuration. Due to this feature, many experts prefer to use this framework. The MapReduce paradigm consists of "sending a computer to where the data is located," that is, processing big data is done on the same cluster where it is stored. The MapReduce algorithm consists of three stages: Map, Shuffle and Reduce (Figure 1. [12]).

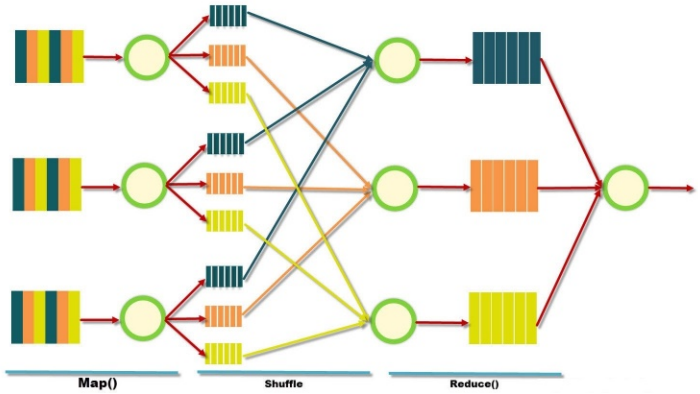

Figure 1. Algorithm MapReduce.

Map stage. At the input, the MapReduce receives data, which is usually stored in files or directories. Data is filtered and preprocessed at this stage. Here, the user-defined map () method processes the data. To implement this method, it is customary to use the Mapper class, which converts the input data into a pair < k1, v1 >, where k is the key, v is the value. This is because MapReduce works exclusively with this data format. At the output of this stage we get list < k2, v2 >.

Shuffle Stage. It goes unnoticed for users. Here, the map () output is sorted by one key obtained at the Map output. The output of this step can be represented as < k2, list (v2) >.

Step Reduce. The output of the second step is the inputs to the reduce () method, which is implemented by the Reducer class. This user-defined method calculates the end result for a single key. The set of values returned by this step is the output of MapReduce, presented as list < k3, v3 >.

The main advantages of MapReduce are support for parallel data processing and local processing of big data. This means that all functions of the map () and reduce () methods flow independently and in parallel. Note that the functions of the map () method will be executed first, and the function of the reduce () method will be at the top.

The disadvantage of this model is that it completely scans all data at the input and it does not apply indexes, which is not acceptable when making a decision in a short time. The MapReduce model is designed so that both intermediate and final results are stored on disk, which leads to an increase in the number of operations performed, thereby increasing the processing time of large data. [13, 14, 15]

4 Social processes monitoring at the university

The main task of the university is the socialization of students, i.e. training of students for social life and labor market. [16] It is possible to define as far as the student is ready to social life having analyzed data on students. Moreover, having carried out data analysis it is possible to learn about teachers how they promoted it. It is necessary to define and make observations of social processes at university for a solution to this task.

The main sources of social processes are students and teachers at the universities. The relationship of students with teachers, students, and teachers among themselves is the object of monitoring of social processes at universities.

During the semester and academic year, the university accumulates information and documents of various kinds about students and teachers. The following can serve as sources of information in universities:

- An electronic journal is a journal containing information about the attendance and achievement of students in the disciplines taught by them during the semester. The e-journal contains attendance data obtained by students for laboratory work, presentations, midterms, the score obtained by the student in the exam, as well as the total score for a certain discipline;
- The information system of the university is a single repository of information about the structures of the university, its students, and teachers. This information system can include an electronic journal, remote education support tools, corporate mail, and other components;
- The teacher's plan is the main document of the university teacher, which contains data on educational and methodological work, research work, scientific and methodological work, etc. This plan is for the next school year;
- The lecturer's report is the final document on the scientific staff of the department, containing information on the work done throughout the academic year;
- The final report of the department's employees is a final document containing qualitative and quantitative indicators that characterize the past educational process. Also, this document should reflect the dynamics of the department.

It is possible to define the proceeding social processes at university having analyzed data from these sources. Such data on students are results of intermediate examinations, laboratory works, independent works, and data on attendance, the point gained during a semester, results of exams, and the general point on a subject. Except for data on students, there is collected data about teachers at the universities: an experience, quantity released articles in foreign logs, qualification of teachers, the quantity the leading students, an academic load, a ratio of the number of students to one teacher, to the students teaching the number to groups, projects teaching quantity. It is possible to reveal social processes having browsed these data where the main source is the student and the teacher.

The social process of "**cooperation**" is presented in the form of combining students into groups, teachers into departments, in turn, groups and departments are combined into faculties.

Also, cooperation includes the unification of students and teachers into teams for start-ups or projects to win various kinds of grants in any competitions or participation in competitions (for example, hackathon). In addition, cooperation is manifested as an association of all university teachers to give appropriate knowledge to the student.

From the above, it can be concluded that in order to determine the social process of "cooperation," students need data on their participation in various events. For each student, the Map () function will receive information about the event and the number of times the student participated in it (for example, <start-up, quantity>, <project, quantity>, etc.). The Reduce () function will determine the total quantity of events in which the student participated. With a high value of this indicator, it can be concluded that the student is easily adapted to cooperation.

The social process of "**competition**" at the university is expressed as the struggle of students in one specialty for a limited number of scholarships. In this case, the main parameter is the average student score scored during the semester in all passing disciplines.

To identify competition among students in one specialty, it is necessary to have data on the specialty and students studying in this specialty, as well as to have information on the number of places on the scholarship. This will be obtained at the output of the Map () function (<specialty, student>). And in the function Reduce (), students will be sorted by the total score per semester. The first who fall into this number and do not have sections will receive a scholarship during the next semester. The fewer places on the scholarship, the higher the level of competition for the scholarship in the corresponding specialty.

In addition to the above, competition among the research staff of the department is observed as participation in competitions for a free position (for example, head of the department, dean) or in competitions for individual grants.

The social process of "**adaptation**" is observed in freshmen, where they adapt to the conditions of university life. This process is monitored on the basis of attendance data and performance statistics, and if it is improved, the adaptation process is successful. The end result of this process is graduation, otherwise, the student is expelled from the university.

This process will be monitored as follows. At the Map stage, information about the student will be obtained, or rather, his average points per semester (<student, average score>). In the next stage, the dynamics of this indicator will be studied. The successful end of this process is revealed upon graduation.

In the university, adaptation is also manifested during the communication of the teacher with students, where the student adapts to the conditions of the teacher, and the teacher looks for the desired approach of study. With the negative development of this process, it leads to administrative penalties, sometimes to conflicts.

Also, adaptation includes a student moving from the region to large cities for higher education. In the

university, adaptation is manifested during the communication of the teacher with students, where the student adapts to the conditions of the teacher, and the teacher looks for the desired approach of study. In addition, the educational burden of the teacher can also be attributed to adaptation.

The social process of "**conflict**," along with competition, is the most common social process in the university. The conflict in the university is expressed as a confrontation between a teacher and a student. Usually, a conflict between a teacher and a student arises due to a score, where students compete for a scholarship.

To determine the conflict between the group and the teacher, it is necessary to analyze the data on the number of sections and on the average score of the group in the corresponding subject. That is, at the Map stage, pairs <S, C> and <S, Sc> will be formed, where S is the subject that the teacher teaches to the corresponding group, C is the cut received by students in the corresponding subject, Sc is the student's score in the corresponding subject. The Reduce step calculates the total number of slices and the average score of the group for the corresponding item. With a large number of sections and a smaller average score of the group, we can talk about a conflict between the group and the teacher.

Also, in addition to the conflict between the group and the teacher, it is possible to identify a conflict between teachers and students. To determine a conflicting teacher, the following data are needed: the total number of students (S_{tn}) who did not pass the exam, the percentage of these students out of the total number of students (x), the number of students who were not admitted to the exam due to attendance, and the results of sessions on other exams for students. And to determine a conflicted student, you need - the average score per semester, the number of academic debts, attendance results. Usually, a conflict between a teacher and a student arises due to a score, where students compete for a scholarship.

The social process of "**assimilation**" at the university is observed mainly as obtaining students with new cultural and initial professional qualities. In the end, this should lead to the admission of students to work. In other words, the process of assimilation is the main task of the university.

In this case, at the Map stage, the specialty is the key, and the values will be the number of graduates working in the specialty, out of the total number. Look like <S, G> and <C, TNG>, where S is a specialty, G is a graduate in the corresponding specialty, TNG is the total number of graduates in the corresponding specialty. At the Reduce stage, the total number of graduates in the corresponding specialty will be calculated. The total number of graduates is necessary to find the percentage of working graduates to the total number of graduates.

An example of an assimilation process among teachers is the upgrading of his skills. This process is also observed in the activities of doctoral students who are just starting to engage in scientific research, and also sometimes works as an assistant.

The social process of "**amalgamation**" at the university occurs when freshmen are combined into groups, which is an integral part of the university is, where, there is a process of exchange of cultural qualities between students.

At the Map stage, a pair <G, Npe> <S, P> will be formed, G is the group in which amalgamation will be evaluated, Npe is the number of no passing exam per one subject, S is the student of this group, P is the student's point per subject. After sorting, the main processing takes place. Here, the average group score per item and the total number of slices per semester are calculated. If the average score per subject is slightly different from the scores of students of that group and the average number of sections does not significantly differ from sections per subject, then it can be concluded that the amalgamation process is successful.

Social borders are selected two groups of people at the universities: students and teachers. In turn, students divide into groups and teachers on departments.

The lesson act here as social communication where the contact between students and teachers is come acts here.

5 Conclusion

Monitoring of social processes is possible to carry out on the following algorithm at the university:

1. It is necessary to define information sources. In this case, sources of information can be the electronic education system, groups of students, and pages concerning the university service;
2. Follows will make the uniform database where would enter all information, received from sources. For drawing up base it is possible to use Hive, Hbase, or something from NoSQL;
3. Further, it is necessary to define data necessary for the definition of social process at the university. Afterward, it is necessary to carry out data analysis for receiving estimates of the social processes, which list above. For definition of estimates of social processes is possible to use Hadoop and its distributors, Spark, language R, and other means of the processing of big data;
4. After receiving estimates of monitoring, it is possible to build forecasts of these or those processes. It is necessary to take the corresponding actions having received the results of forecasts. For forecasting, it is possible to use simulation modeling for the creation of the virtual model of the proceeding processes.

References

[1] The volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2024, https://www.statista.com/statistics/871513/worldwide-data-created/

[2] Igor Ilin, Anastasii Kliminm, Anton Shaban, Features of Big Data approach and new opportunities of BI-systems in marketing activities Web of Conferences 110,0 (2019)

https://www.researchgate.net/publication/335067775_Features_of_Big_Data_approach_and_new_opportunities_of_BI-systems_in_marketing_activities
[3] Big data (Big_Data), http://www.tadviser.ru/index.php/Article:%20Big__data_ (Big_Data).
[4] K. Guba, "Big data in sociology: new data, new sociology?". The magazine "Social Review" volume 17 No. 1, p. 213-236, 2018.
[5] Big Data Technology in Economics, http://ru.datasides.com/big-data-in-economics/.
[6] V. Svilas, "Big Data will help increase your company's profit. How it works?". Edition "Rusbase", 2018.
[7] US S. Kesaev, V. V. Alekhno, "Prospects for the use of Big Data in personnel management". Electronic journal "Nauka-Rastudent.ru", 2017.
[8] E. Yu. Zhuravleva, "Sociology in the Network Environment: Toward Digital Social Research". The journal "Sociological research" No. 8, p. 25-33, 2015.
[9] The concept of the social process, https://psyera.ru/ponyatie-socialnogo-processa_8350.htm.
[10] Types of social processes, https://psyera.ru/vidy-socialnyh-processov_9845.htm.
[11] Monitoring of social processes, https://spravochnick.ru/sociologiya/suschnost_i_principy_socialnyh_processov/monitoring_socialnyh_processov/
[12] Big Data от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce. https://habr.com/ru/company/dca/blog/267361/
[13] Hadoop – MapReduce, https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm
[14] MapReduce Tutorial, https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
[15] MapReduce, https://ru.bmstu.wiki/MapReduce
[16] L. G. Vasilieva, "Socialization of students and the educational process management system in a branch of a university (for example, the Arsenyev city district). Magazine "Young Scientist" No. 10, p. 301-303, 2009.