ISSN: 1987-6521, E-ISSN: 2346-7541

JUNE 2022 VOLUME 59 ISSUE 02
REVIEW ARTICLE

© THE BALTIC SCIENTIFIC JOURNALS
BLACK SEA
SCIENTIFIC JOURNAL OF ACADEMIC RESEARCH

USE OF MODERN INFORMATION TECHNOLOGY FOR PREFORMING OF OBSERVATION OF SOCIAL PROCESS AT THE UNIVERSITY

Yusifov Samad, Ragimova Nazila, Abdullayev Hajimahmud, Khalilov Etibar

Department of Computer Engineering, Azerbaijan State University of Oil and Industry, Baku, Azerbaijan

Email: abdulvugar@mail.ru

Received: 13-Jun-2022, Manuscript No. BSSJAR-22-66539; **Editor assigned:** 15-Jun-2022, PreQC No. BSSJAR-22-66539 (PQ); **Reviewed:** 29-Jun-2022, QC No. BSSJAR-22-66539; **Revised:** 06-Jul-2022, Manuscript No. BSSJAR-22-66539(R); **Published:** 15-Jul-2022, DOI: 10.36962/GBSSJAR/59.2.007

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 interactions of large 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 monitor social processes at the university. For realization of monitoring of social processes different technologies are used: products 1010 data (1010 edge, 1010 connect, 1010 reveal, 1010 equities), products of Apache Software Foundation (Apache Hive, Apache Chukwa, Apache Hadoop, Apache Pig!), MapReduce framework, language R, library Pandas, NoSQL, etc.

Keywords: Big data, sociology, social process, monitoring of social process, hadoop.

INTRODUCTION

We live in an information age where the main products become information and knowledge in economy. The beginning of information age can consider appearance of microprocessors and personal computers. Those who possess them dominate in economy. According to the report of the analytical company IDC "Digital Universe Study", the volume of digital 1 data was 0.18 zettabyte in 2006, and a volume was 1.8 zettabyte in 2011, till 2020 all a volume of data of our planet will reach 40 zettabyte where statistically about 5200 gigabyte of data come to each inhabitant of Earth. Promptly growing data volume provokes to search of new means for their storage and processing. In this regard, the term "big data" enter.

LITERATURE REVIEW

Big data is structured and unstructured data that is incapable to process by traditional methods of data processing. It is possible to carry a method comparison, a method of application of absolute and relative values, a method of use of average values, a grouping method, a balance method, a graphic method, a graph-mathematical method to traditional methods of processing. These methods can implement by tools of SQL in combination with one of object-oriented programming languages. The problem of big data is that they present in web logs, a video, GPS data, machine code etc. that different from the structured DB format (McCall 2020).

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

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

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

- Fault tolerance. It is possible that the quantity of computing nodes will grow, so the probability of their exit out of operation grows. Therefore, tools of big data should be ready to such situations and are capable to take corresponding measures;
- Locality of data, It is necessary to store and process data in the same physical server, otherwise expenses of transfer of data between servers can be enormous.

Application of big data has found the reflection in many aspects of our life, including in economy, in management and business, anthropology, stories and in sociology (Javaid, et al. 2020).

According to McKinsey & Company, there are five basic approaches of use of big data in 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 (Kavakiotis, et al. 2017).

According to IDC, 90% of the data stored on servers of the companies are practically useful, but it is not suitable for use. Useful information for business in the company generally obtains 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 (Deibert and Rohozinski 2010).

Big data is capable to automate and generalize functional approaches to search and selection of employees, improvement of quality of work of personnel and increase in labor productivity, a solution of tasks in the field of education of commands by a ratio of qualities of people in management. The main task of big data is management of talents and improvement of trial and error methods of personnel in human resource management (Goodall, et al. 2009).

Sociology and information technology

Implementation of big data in sociology generates two types of sociology: The 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 (Kozlenkova, et al. 2014).

According Cioffi-Revilla "Computational social sciences" is the integrated cross-disciplinary search in a social research by means of calculations at increasing the scale of information processes. Watts consider the term "Computational social sciences" as a lable 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 a 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 by means of examination, the methods of machine learning, economies, sociology, psychology, marketing, political sciences. Cioffi-Revilla has offered 5 methods of researches for "computational social sciences, it should be noted that these methods are not exclusive for "computational social sciences":

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 safety of social networks, studying of the terrorist networks;
3. Geospatial analysis-Applying geographic information systems, researchers study space layers of distribution of the ideas;
4. Modeling of complexity-Applies mathematical means to understanding of 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 of environments and emergence of the new organizations.

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

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

Social process is a change of social character in society, which cause by desire of certain groups to influence the current situations in society for 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 balance with each other. Such interaction of interests of communities is resulted by the actions of unknown forces caused directly by this interaction. Result of these actions sets the direction of social processes. Subordination of social processes by subjects of society to the vector of behavior and option of probable actions is their priority task.

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

1. Cooperation is a process association of persons in groups, for the sake of the general interest with the purpose to receive benefit. For this purpose, the mutual respect and establishment of rules of cooperation in-group is 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, the power, love, etc.). In essence the competition is 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 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 other part of society;
6. Amalgamation is a process that arises when mixing groups of subjects of society. Difference from assimilation is in what after completion of process of amalgamation of an edge between groups erase.

There also add 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 select. Ambits of social groups separate their members from all other society. Processes of assimilation and amalgamation follow by erasing of borders between social groups, destruction of the available division for creation of common features of group.

Systematic communications are processes link establishment between social groups that are in the set social borders. At absence at any group of communication with other groups will lead to its isolation. Monitoring of social processes is set of methods of collecting and processing of social processes for detection of patterns of development of society and each individual separately. Objects of social monitoring are all set of the social phenomena and processes.

At this moment histories monitoring of social processes carry 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 a 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 implement based on certain principles:

- Completeness of social information;
- 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 execution of functions and problems of monitoring of social processes. Applying these means, an opportunity collecting 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. Afterwards there is an opportunity to define to positive or negative changes will lead 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 search of necessary measures for prevention of negative trends and maintenance positive, leading society to further development.

Today's situation

The main task of the university is socialization of students, i.e. training of students for social life and for labor market. 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 of this task.

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

Each semester at the university collects a huge number of information on students and teachers. It is possible to define the proceeding social processes at university having analyzed these data. 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 data on students, there is collected data about teachers at the universities: an experience, quantity released articles in foreign logs, qualification of teachers, quantity the leading students, an academic load, a ratio of 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. Example of process of cooperation is association of students in groups; association of teachers in departments of any area in science for its development and teach to students of training programs; writing of degree and dissertation works where the teacher and the student combine the efforts; collaboration of several teachers over one project or writing of article.

The competition is shown when students of one specialty fight for a grant. Here the only parameter is the GPA for a semester in all objects. The competition to a certain position (the manager of department, the dean) can be an example of the competition among teachers.

The most widespread social process at university is the conflict between the teacher and the student. The following data are necessary for definition of the conflict teacher: the general the number of the cutting-off students, percent of the cutting-off students from the total number of students, the number of the cutting-off students because of attendance, results of sessions on others of examinations of students. Moreover, are necessary for definition of the conflict student-GPA for a semester, quantity of cuts, results of attendance. Usually the conflict between the teacher and the student arises because of point where students conduct the competition for a grant.

Process of adaptation observes at first-year students where they adapt to the new environment. Monitoring of this process carry out based on data on attendance and statistics of progress if it is found, then process of adaptation is successful. The result of this process is the termination of university; otherwise, the student expel from university. Also moving by the student from the region to big cities for receiving the higher education belongs to adaptation. Adaptation show during communication of the teacher with students at the university where the student adapts to the teacher's conditions, and the teacher looks for the necessary approach of training. In addition, the academic load of the teacher belongs to adaptation.

Example of process of assimilation for teachers is professional development. Assimilation show at students when taking for work.

Associations of first-year students in groups are integral part of the university where there is an amalgamation process, i.e. there is a process of exchange of cultural qualities between students. Students integrate in groups on the selected specialties. The same occurs when teachers integrate in departments.

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.

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

Data mining methods: Intelligent Data Analysis (IAD) is a set of methods for detection in data of earlier unknown, uncommon, practically useful knowledge that are necessary for decision-making.

Treat the IAD methods:

- Association rule learning. Set of methods for detection of associative rules in data bulks;
- Classification is purpose of objects, observations or events to one of earlier announced classes;
- Cluster analysis is the statistical technique reference of objects to groups thanks to identification of the general sign;
- Regression is 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: Is set of methods for 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 set of methods, the self-training pursuing the aim creation of algorithms based on empirical data. Distinctive feature of machine learning is not the direct solution of an objective, but training during application of solutions of similar tasks.

Artificial neural networks: Are a mathematical model, which construct based on biological neural networks.

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

Predictive analytics: Is 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 creation of model, which describes a real system virtually. It use generally for forecasting and carrying out experiments.

Spatial analysis: Is set of methods of the analysis of space data.

Statistical analysis: Is set of methods of collecting, the organization and interpretation of big data.

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

It is possible to carry products 1010 data (1010 edge, 1010 connect, 1010 reveal, 1010 equities), 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.

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

- 1010 edge-the intuitive platform of corporate data analysis, the needs 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;
- 1010 connect-the detailed portal allowing to share and control distribution of data outside the enterprise that allows to create 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;
- 1010 reveal-a set of intelligent solutions for consumers of high definition;
- 1010 equities-this product provides alternative solutions for data transmission on the party of the buyer.

Map Reduce is a framework of distributed computing, the developed Google applied to parallel computing over data sets with a size up to several petabyte. It consists of three stages:

1. Stage Map. Here data are processed the map methods, determined by the user. At this stage, there is filtering and predate processing. The map method gives a set of couples a key value.
2. Stage Shuffle. Proceeds imperceptibly for users. Here the map function output is distributed on one key received on a stage map output.
3. Stage Reduce. The output of the second stage is inputs for the reduce method. This method set by the user calculates the result for a separate key. The set of values returned by this stage is MapReduce output.

The most widespread tool for work with big data is Apache Hadoop that implements based on MapReduce. Apache Hadoop is a set of libraries and utilities for 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 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 on the basis of NoSQL;
- Mahout-library of machine learning;
- Apache Spark is the engine of the distributed data processing using the Hadoop components (HDFS and YARN);
- Apache Tez-a framework which 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, a 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.

DISCUSSION

The Hadoop installation represented 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, Hortonworks and MapR.

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

The Hortonworks distribution kit is Hortonworks Data Platform including DataPlane services for integration with solutions third-party software vendors and own module of management of a cluster Hortonworks Management Center on the basis of Apache Ambari which is a part of each distribution kit absolutely free of charge.

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 balance between stability and high-speed performance, saving at the same time the simplest uses.

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

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

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

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

CONCLUSIONS

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 the electronic education system, groups of students and pages concerning the university serves;
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, which are necessary for definition of social process. For definition of social processes at the university, data describe above. Afterwards 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 processing of big data;
4. After receiving estimates of monitoring, it is possible builds forecasts of these or those processes. It is necessary to take the corresponding actions having received results of forecasts. For forecasting, it is possible to use simulation modeling for creation of virtual model of the proceeding processes.

REFERENCES

1. Deibert, R., & Rohozinski, R. (2010). Liberation vs. control: The future of cyberspace. Journal of Democracy, 21(4), 43-57.
2. Goodall, J. R., Lutters, W. G., & Komlodi, A. (2009). Developing expertise for network intrusion detection. Information Technology and People.
3. Javaid, M., Haleem, A., Vaishya, R., Bahl, S., Suman, R., & Vaish, A. (2020). Industry 4.0 technologies and their applications in fighting COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(4), 419-422.
4. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.
5. Kozlenkova, I. V., Samaha, S. A., & Palmatier, R. W. (2014). Resource-based theory in marketing. Journal of the Academy of Marketing Science, 42(1), 1-21.
6. McCall, B. (2020). COVID-19 and artificial intelligence: Protecting health-care workers and curbing the spread. The Lancet Digital Health 2(4), e166-e167.

Citation: Samad Y., Ragimova N., Abdullayev H & Khalilov E. (2022). Use of modern information technology for preforming of observation of social process at the university. GBSSJAR. 59(2), 1-8. DOI: 10.36962/GBSSJAR/59.2.007