IIoT Competencies Support in the Master's Program “Automation of technological Processes and Production”

Vasilij A. Usachev[1], Vjacheslav I. Voronov [2]
Department of Intelligent Systems of Control and Automation
Moscow Technical University of Communications and Informatics
Moscow, Russia

[1]usvas.01@gmail.com, [2]vorvi@mail.ru

Abstract. **The department of ISCA MTUCI is training specialists to manage industrial Internet of things systems, based on a number of disciplines that form students’ competences for working with IIoT systems. The article discusses the features of teaching one of these disciplines “Intellectual databases and data warehouses” using the freely distributed software DBMS Hadoop.**

Key words: **Industrial Internet of Things, Big Data, Hadoop.**

I. INTRODUCTION

The professors of the ISCA department developed a number of disciplines on the direction “Automation of technological processes and production” (“Intelligent automated information management systems” profile) that form students’ competences for working with IIoT systems, such as “Intelligent databases and data warehouses”, “Intellectual methods data processing”, “machine learning and big data”, etc. [1].

The purpose of mastering the discipline “Intellectual databases and data warehouses” by students is to study the methods of building knowledge bases and data warehouses and their application possibilities for intellectualizing automated processes for storing and processing information.

Tasks of mastering the discipline are: the study of the principles of building knowledge bases and data warehouses; mastering the methods of constructing queries in the NoSQL language; mastering the methods of data mining for automating the storage and processing of information [2].

The total complexity of the discipline, studied in the second semester, is 4 credit units. The discipline provides lectures, practical exercises, laboratory work, independent work, the exam.

The main sections of the discipline are: introduction to the industrial Internet of things technology, Big Data, Data Mining, methods for creating, processing and storing data in NoSQL databases and data warehouses, intellectual analysis tools and their application in databases and data warehouses, developing applications based on Apache Hadoop, ensuring information security of Big Data class systems.

II. ABOUT THE CONTENT OF LECTURE COURSE

As part of the lecture course, attention is focused on the fact that the industrial Internet of things consists of three main components (Fig. 1): a set of intelligent, network-enabled products, product systems and other “things” connected through a communication infrastructure similar to the Internet infrastructure.

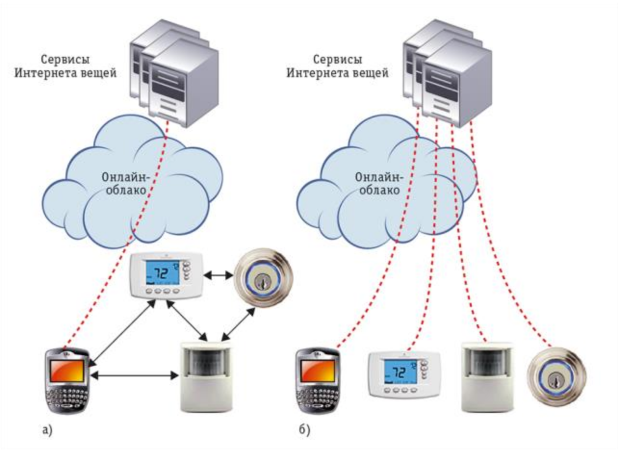

Fig. 1. Structure of the Internet of Things [3]

When developing and implementing the Internet of Things at the enterprise, special attention is paid to the critical infrastructure and the need for real-time analysis of all information coming from numerous sensors, controllers and devices [4]. Data on the state of the product, its operation and the environment are transmitted in real time to the monitoring systems in order to ensure the management, maintenance and updating of the product, as well as the efficiency of the entire system.

When analyzing solutions within the Industrial Internet of Things (IIoT), examples of implementation in the housing and utilities sector and in the mining sector are of interest.

Perhaps the most striking example of IIoT is the experience of the Moscow United Energy Company (MUEC). MUEC, together with the cellular operator MTS, is implementing a project for monitoring energy consumption [5], within which a unified automated system for controlling and recording the transfer of thermal energy and hot water has been created. For this purpose, MUEC installed 47 thousand consumption meters in municipal houses and social facilities, which continuously transmit data to the server of the MUEC Central Energy Accounting System.

The company also launched a system that allows remotely, via the Internet, to take readings of heat consumption in residential buildings in Moscow. Such meters are installed on 23,000 MUEC facilities. The system allows power engineers to quickly find out about failures in heat supply and predict possible breakdowns.

The introduction of intelligent technologies, taking into account the length of linear objects, leads to an increase in reliability and lower operating costs. This allows you to go to the management of the network “on state”, and not to carry out repairs in accordance with strict regulatory deadlines.

Another example of the use of IIoT in real production is the experience of Alrosa, which manages several mining and processing plants. Alrosa and the Sum of Technologies system integrator implemented a project to create an automated information management system for the production of the Aikhalsky Minerals processing plants (Yakutia Republic) [6]. With the introduction of the system, JSC Alrosa received a single source of data on the operation of process equipment and a tool for analyzing the efficiency of production activities.

In the course of the project, automatic data collection was organized from process equipment control systems, energy accounting systems and information systems used to manage the mining and transport complex, drilling operations, underground mining, equipment maintenance planning and repair. Production information is consolidated in a unified repository, acting as a single source of technological data for production services and information systems of JSC Alrosa. Detailed data on key performance indicators are displayed on video walls located in the control rooms of the Aikhalsky Minerals processing plants and in the Administration of JSC Alrosa and transferred to the production services for analysis.

Based on the information collected, the system generates analytical reports for the operational control of the implementation of planned production indicators by the mining and processing divisions, as well as for the analysis of technical and operational indicators of mining and transport, mining and process equipment.

Regardless of the type of enterprise activity, the multitude of connected sensors and devices generates a large flow of the most diverse information that needs to be collected, processed and provided an analysis of the state of production as a whole and its components [4].

III. DBMS NoSQl (Hadoop)

To handle the flow of information resulting from the advent of the Internet of Things, it becomes more convenient to use non-familiar relational DBMS such as Oracle, Microsoft SQL Server. Now these products are being supplanted by free tools for working with unstructured data. One of them is Hadoop. The university does not always have the opportunity to work with paid software, so Hadoop is great for use in research tasks. Hadoop contains tools such as:

- Hadoop Distributed File System (HDFS) is a distributed file system that allows you to store information of almost unlimited size;
- Hadoop YARN is a framework for managing cluster resources and managing tasks, including the MapReduce framework;
- Hive is a tool for SQL-like queries on large data (turns SQL queries into a series of MapReduce tasks).

At the heart of Hadoop is the MapReduce paradigm.

The work of MapReduce is based on splitting data processing into two phases: the phase of the map and the phase of convolution reduce. Each phase uses as input and output data a key-value pair, the types of which are selected by the programmer. The map function is simple. We extract the fields we need. The map function is also well suited for eliminating unwanted entries: missing, questionable or erroneous values are filtered out here. The mapping function can run on every machine in the Hadoop cluster. Thus, the possibility of parallel processing of the initial information on several machines appears [7]. The output of the mapping function is processed by the MapReduce infrastructure before it is passed to the reduce function. During this process, key-value pairs are sorted and grouped by key. In the phase of reduce, the processing of values takes place: the calculations need to be made to solve the problem.

In the course of mastering the discipline “Intellectual databases and data warehouses” undergraduates get practical skills in solving problems in the field of data analysis, write the program code of MapReduce tasks, and gain experience in working with the job tracker and task tracker. A MapReduce job is a unit of work that a client wants to do: it consists of input data, a MapReduce program, and configuration information [8]. To perform a job, Hadoop splits it into tasks, which are divided into two types: map tasks and reduce tasks.

Practical acquaintance with big data begins with a course on the Udacity portal and the virtual machine from Cloudera [9]. In this course, students are provided with various data sets: information about purchases, information about records on the forum, logs of the web server. Undergraduates must calculate the average cost of goods sold or determine the city in which they paid the most for Visa cards, find the most active forum participant for a certain period of time or the most visited page on the site.

The system requirements of the Udacity virtual machine are not large, so its launch is possible on all modern computers. Also, when teaching students the discipline used more powerful tools. The department uses a virtual machine image taken from the Cloudera website [10]. This image differs from Udacity in a richer toolbox. The virtual machine has all the necessary components of the Hadoop ecosystem. In this machine, a full-fledged Hadoop cluster is already deployed and configured. Users do not need to install and configure anything themselves. Also, there are already data sets for learning, and therefore students can start writing their MapReduce tasks immediately after starting the virtual machine. It is possible to upload your data to the Hadoop file system (HDFS) and process it as if they were located in a real Hadoop cluster. The volume of the processed data is not limited. The system requirements of this virtual machine are significantly larger and not every computer is suitable for this purpose. Especially to ensure the educational process, the department has a server that runs virtual machines in an amount sufficient to perform laboratory and coursework by undergraduates. Access to machines is remote.

In the work program of the direction 15.04.04 “Automation of technological processes and production” there is a laboratory course. Some of the works from the course are described in [11].

IV. ABOUT THE LABORATORY PRACTICE

Laboratory work 1. Study of the construction of queries and ways to output a relational database

The goal of the work is to learn how to perform SQL queries in the Cloudera environment with Hadoop tools.

Progress. You need to download data to the HDFS file system using Apache Sqoop. To do this, you need to open a terminal in Cloudera and execute the appropriate command.

This action may take some time. The MapReduce task is launched to export data from the MySQL database and import it into HDFS. Also, tables are created to represent HDFS files in Impala with the appropriate schema. After the import is complete, you need to check whether it has occurred correctly.

Laboratory work 2. Correlation of structured data with unstructured data

The purpose of the work is to gain the skill of working with structured data and unstructured data.

Progress. It is necessary to find the most viewed products in the online store; find out if they are the best selling

Since Hadoop can store unstructured and structured data without changing the entire database, you can also receive, store, and process the web event log. This allows you to find out what site visitors actually viewed most often.

For this we need data on web visits. The most common way to track site navigation is to use Apache Flume. Flume is a scalable tool that allows real-time tracking of routes, filtering, combining and performing “mini-operations” on the data of these paths. For convenience, some sample access data has already been preloaded in Cloudera in log files.

Laboratory work 3. Strong analysis of relationships using Spark

The objective is familiarization with the service Spark to speed up and simplify data processing.

Progress. The Spark service uses a similar concept of the ‘map’ and ‘reduce’ operations (the ‘join’ and ‘groupBy’ operations are just special variations of the ‘reduce’ operation). The key advantage of using Spark is that the code takes up less space and intermediate results can be stored in memory, which generally allows for iterations to be much faster.

Using MapReduce is a good option for tasks that use data that cannot fit in memory (for example, petabytes of data). This work uses Spark-on-YARN, which means that MapReduce and Spark (like many components of CDH) have a common resource manager, which makes it easier to allocate resources among a large number of users.

It is necessary to analyze the relationship with the use of Spark and to determine the products most often ordered together. A tool in CDH for quick analysis of object relationships is Apache Spark. For this work, Sparkjob is used to give an idea of the relationships of objects.

Laboratory work 4. Interactive study of event logs

The objective is index data using any of the indexing options provided by Cloudera Search.

Working process. You can choose to batch index data using the MapReduce Indexing tool or, as in our example below, expand the Apache Flume configuration, which already received web log data and event placement in Apache Solr for real-time indexing.

Web log data is a standard web server log.

Solr organizes data in the same way as a SQL database. Each record is called a document and consists of fields defined by the schema: the same as a row in a database table. The difference is that the data in Solr are usually more poorly structured. You can start indexing in real time with Cloudera Search and Flume on the web server’s log data and use the Hue search user interface to explore it by creating a search index. Typically, when you deploy a new search pattern, four steps are taken:

- Create an empty configuration. In the Cloudera virtual machine, you will not need to perform steps 1 or 2, since the configuration and the schema file are already included in the cluster.
- Editing the scheme. The most common area that may be of interest is the <fields><fields /> section. From this area we can determine the fields that are present and searchable in the index.
- Download configuration. This operation may take several minutes.
- Creating a collection.

Laboratory work 5. Data Visualization

The objective is to display graphically the correlations found in previous laboratory works.

Familiarize yourself with the Hue web interface for visualizing data and creating your own dashboard.

Create a dashboard and use it to visualize the results obtained in previous works.

V. COURSE WORKS

At the end of the training discipline “Intellectual databases and data warehouses,” students performed term papers. They were given a new unfamiliar data set. Undergraduates could take the proposed data for work or find their own, but in this case they had to compile competent, interesting requests to their data. All calculations must be implemented using MapReduce [8]. The proposed input data contains the following fields: creation-date, education, employment, experience, industry, jobname, location, salary_max, salary_min, schedule. It is also worth noting that the data set includes 422434 rows with information on vacancies.

Example job assignment for a course work on working with big data:

1. Based on the “schedule” field, calculate the percentage for each “work schedule”. To do this, it is necessary to calculate the number of vacancies related to each “work schedule”, then build a pie chart. Comment on the result.

2. Based on the “location” field, count how many vacancies are related to each area. This post consists of the full address of the employer, for example: <location> Smolensk region, Dorogobuzh district, Dorogobuzh, Sovetskaya street, 1 </ location>. However, to accomplish the task, it suffices to single out the region (region, territory, republic, etc.), that is, the information located up to the “first comma”. In the above example will remain: "Smolensk region.” The distribution of the number of vacancies by region should be displayed in the table. In addition, print the total number of vacancies. Comment on the result.

3. Based on the “location” and “salary_min”, “salary_max” fields, calculate the minimum, maximum and average salary in the region. The obtained data must be displayed in the table. Comment on the result.

4. On the basis of the “location” field for the “Sverdlovsk Region”, bring out the top 10 most sought-after professions in the “jobname” field. The obtained data must be displayed in the table. Comment on the result.

5. Determine the maximum salary for the “Ecologist” of the “jobname” field if the “education” field is “Highest” and the experience of the “experience” field is more than or equal to 3 years.

If a field does not contain information, that is, is empty, it is not necessary to take it into account. This note does not apply to the “schedule” and “employment” fields.

One of the undergraduates for the course work was selected data from a social survey conducted in the United States [12]. When questioning people asked for different information: gender, age, education, marital status, type of employment, number of working hours per week, salary and other parameters. The question was raised about the average length of the working week depending on age. During the decision, the student obtained an interesting result (Fig. 2).

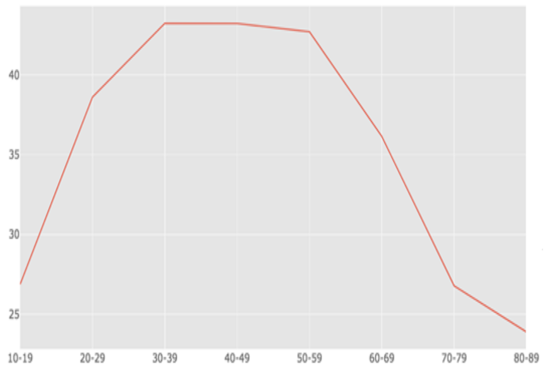

Fig. 2. The average duration of the work week depending on age

The resulting graph is similar to the image of the Gaussian distribution and says that the level of employment depends on the knowledge, experience of the specialist, as well as on the physical condition of citizens. Interest can cause the level of employment of people from 80 to 90 years. A smoother descent can mean that busy people at this age are high professionals and the employer needs such personnel, despite the age of these employees.

VI. CONCLUSIONS

The article presents the program of the discipline “Intelligent databases and data warehouses”, which gives undergraduates practical skills to work with the toolkit of big data processing in a virtual machine from Cloudera.

Using Hadoop with MapReduce and Hive requires virtually no deep knowledge of programming. Using modern tools, you can perform analysis and forecasting, as well as use visualization to simulate problems, which significantly speeds up and facilitates work with big data [1; 8; 11].

Training in the discipline of such a work program is one of the stages in preparing students for the direction 15.04.04 process automation and production process to participate in the WorldSkills IT standards championship in terms of machine learning and big data and Industrial Internet of Things.

VII. REFERENCES

1. Горячев Д.В., Воронов В.И. Большие данные и машинное обучение // Технологии информационного общества: Матер. XII Междунар. отраслевой науч.-техн. конф. М., 2018. С. 327–328. [Goryachev D.V. Big Data and Machine Learning // Information Society Technologies: Proceedings of 12th the international scientific-practical conference. Moscow, 2018. P. 327–328].

2. Voronov V.I., Voronova L.I. Features of realization master’s program “Automation of technological processes and manufactures” // International Journal of Applied and Fundamental Research. 2016. № 2. URL: www.science-sd.com/464-25196.

3. Уонт Р., Шилит Б., Дженсон С. Механизмы Интернета вещей. // Открытые системы. СУБД. 2015. № 1. [Wont R., Shilit B., Jenson S. The mechanisms of the Internet of Things. // Open systems. DBMS. 2015. N 1.]. URL: https://www.osp.ru/os/2015/01/13045328/.

4. Безумнов Д.Н., Воронова Л.И. О развитии и стандартизации технологии Интернета вещей // Технологии информационного общества: Матер. XII Междунар. отраслевой науч.-техн. конф. М., 2018. С. 293–294. [Bezumnov D.N., Voronova L.I. On the development and standardization of the Internet of things technology // Information Society Technologies: Proceedings of XII Intern. branch scientific.-tech. conf. Moscow, 2018. P. 293–294].

5. Применение IoT в энергетике: опыт МОЭК // IoT World Russia Summit. [The use of IoT in the energy sector: the experience of the JSC «MIPC» // IoT World Russia Summit]. URL: http://iotworldsummit.ru/novosti/primenenie-iot-v-energetike.

6. АК «Алроса» завершила проект по внедрению информационной системы управления производством Айхальского ГОК // CNews. [AK “Alrosa” completed the project on the implementation of the production management information system of the Aikhalsky GOK // CNews]. URL: http://www.cnews.ru/news/line/2018-09-05_ak_alrosa_zavershila_proekt_po_vnedreniyu_informatsionnoj.

7. Воронов В.И., Воронова Л.И., Генчель К.В. Применение параллельных алгоритмов в нейронной сети для распознавания жестового языка // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2018). VII Междунар. науч.-техн. и науч.-метод. конф.: Сб. науч. ст.: В 4 т. / Под ред. С.В. Бачевского. М., 2018. С. 207–212. [Voronov V.I., Voronova L.I., Genchel K.V. The use of parallel algorithms in the neural network to recognize the sign language // Actual problems of information and telecommunications in science and education (APINO 2018). VII International scientific tech. and scientific method. Conf.: In 4 t. / Ed. S.V. Bachevsky. M., 2018. P. 207–212].

8. Воронова Л.И., Воронов В.И. Big Data. Методы и средства анализа: Учеб. пос. М., 2016. [Voronova K.I., Voronov V.I. Big Data. methods and tools for analysis. Moscow, 2016]

9. Intro to Hadoop and MapReduce // Udacity. URL: https://classroom.udacity.com/courses/ud617.

10. Downloading a Cloudera QuickStart VM. // Cloudera. URL: https://www.cloudera.com/ downloads/quickstart_vms/5-13.html.

11. Воронов В.И., Усачев В.А. Компетенция “машинное обучение и большие данные” // Приоритетные направления развития науки и образования / Под общ. ред. Г.Ю. Гуляева. Пенза, 2017. С. 97–108. [Usachev V.A., Voronov V.I. The competence “Machine learning and Big Data” // Priorities for development of science and education / Ed. G.Yu. Gulayev. Pensa, 2017. P. 97–108].

12. Adult Data Set // UCI Machine Learning Repository. URL: https://archive.ics.uci.edu/ml/datasets/adult.