Knowledge E
Engaging minds

Conference Paper

Practical Training of Students on the Extraction and Analysis of Big Data

Prokhorov I. V., Kochetkov O. T., and Filatov A. A.

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe shosse 31, Moscow, 115409, Russia

Abstract

The article deals with questions of studies, development and practical use in teaching complex laboratory work on extracting and analyzing big data to train specialists in the specialty 10.05.04 "Information and Analytical Security Systems", direction of training "Information sSecurity of Financial and Economic Structures" in the framework of the educational discipline "Distributed Automated Information Systems".

Keywords: big data, data scientist, extraction, processing and analysis of big data, information security of financial and economic structures, the Internet, Yandex, Google, application programming interface –API.

Corresponding Author:
Prokhorov I. V.
ivprokhorov@mephi.ru

Received: 11 December 2017
Accepted: 20 January 2018
Published: 13 February 2018

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the FinTech and RegTech: Possibilities, Threats and Risks of Financial Technologies Conference Committee.

1. Introduction

At the present stage of development of economic and financial security, the most important condition for a successful activity is the ability to process huge arrays and information flows (big data). In connection with the transition of the country to the "digital economy" there is a huge need for training specialists - Data Scientist, focused on working with big data. Research, modeling and development of information technologies for searching, extracting, processing, presenting, visualizing and analyzing "big data" for countering money laundering and terrorism became possible due to ensuring a high level of parallelism in modern data centers.

The characteristic features of big data are their extremely large volume, weak structure and heterogeneity and the need to process them very quickly. Figure 1 shows the technological sequence of the stages of extraction and processing of data from the setting of the information problem to the analysis of the results. At the same time, new models of extraction and analysis of "big data" (including data for Rosfinmonitoring) are often used at the stages of data research and model building (Fig.1).

The long-term practical task is the creation of a laboratory research base for the formation of competences and proposals on the professional standard of the profession

How to cite this article: Prokhorov I. V., Kochetkov O. T., and Filatov A. A., (2018), "Practical Training of Students on the Extraction and Analysis of Big Data" in *III Network AML/CFT Institute International Scientific and Research Conference "FinTech and RegTech: Possibilities, Threats and Risks of Financial Technologies"*, KnE Social Sciences, pages 361–368. DOI 10.18502/kss.v3i2.1565

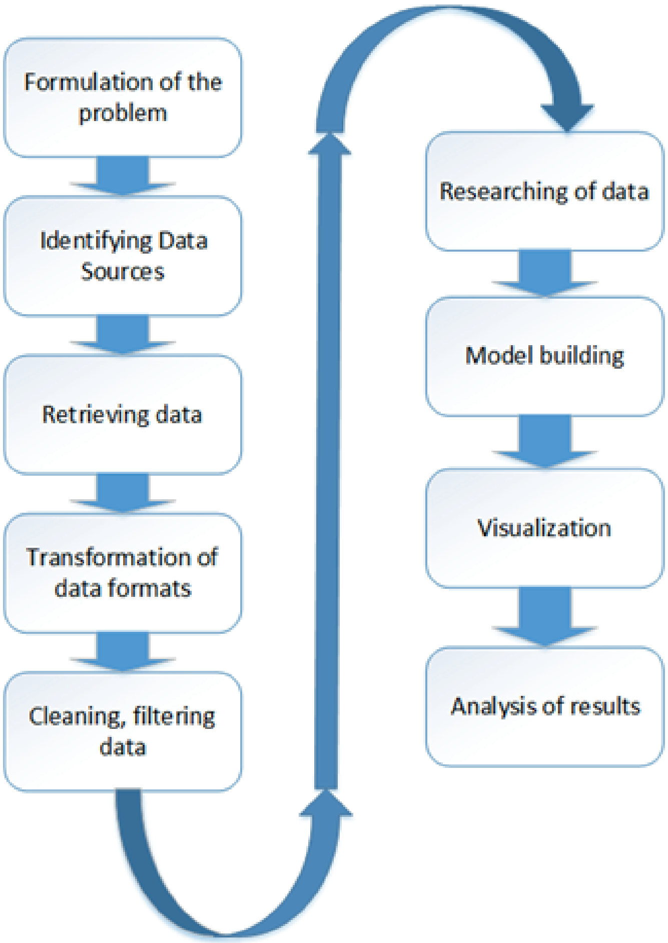


Figure 1: Technology for extraction, processing and analysis of big data.

”Data scientist”, where the main one is the competence to search for, to extract, to process, to analyze and to visualize big data. This will create a laboratory base for the training of specialists, Ph.D. theses and master’s theses on the new master’s program ”Processing and Analysis of Big Data” in the direction of training ”Business Informatics” with the qualification ”Analyst of Big data” or ”Data Scientist” at the Department of Financial Monitoring of the Institute for Financial and Economic Security of the NRNU MEPhI.

2. Big Data

The modern development of information and communication technologies is characterized by three concepts: virtualization, cloud technologies and ”big data”. These information technologies and models become inevitable in the modern world, volatile, unpredictable, complex and ambiguous world, which is often denoted by the term VUCA, where the decision-making process is characterized by:

Volatility - variability, deliberate inadequacy of information,

Uncertainty - unpredictability, multidimensionality of cause-and-effect relationships,

Complexity - complexity and non-obvious interpretation of any important information,

3. Ambiguity

Areas of using big data are quite wide. In marketing it is:

- Market segmentation
- Modeling the flow of customers
- Reference systems
- Analysis of social media (websites, blogs, forums, social networks, interacting with the community through multimedia content (text, video, photos).

In the financial and insurance spheres:

- Prevention of fraud
- Definition of abnormal behavior
- Accumulation and use of MDM data (master data)
- Analysis of credit risks
- Modeling of insured events
- Predictive analytics

Predictive analytics in the financial sphere is information technology predicting the actions of companies and entities in order to identify groups of citizens at risk of financial crimes, fraud, theft and the use of personal data, false insurance claims, tax fraud, as well as to carry out underwriting (studying the solvency of a potential borrower of the bank).

The analysis of the data has been carried out for a long time (including at the Department of Financial Monitoring of the NRNU MEPhI, the discipline ”Data Analysis” is taught). However, the analysis is limited to the volumes measured in megabytes, and at best several gigabytes. In doing so, it is tabular data. Data Science specialists work not only with structured, but also with weakly structured and unstructured data of huge volume, measured in terabytes and petabytes. For analysis, specialized models and tools are created using products such as Hadoop MapReduce, HP IDOL, etc.

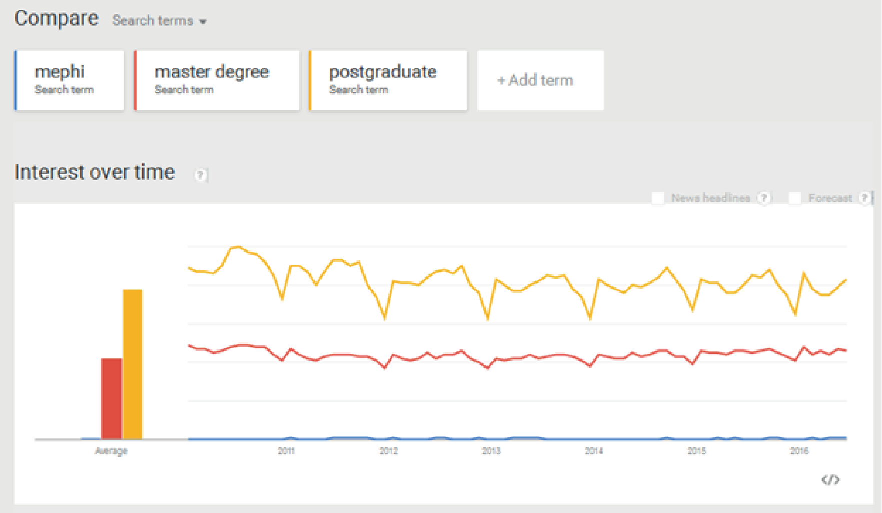
Hadoop [1] is a framework for building distributed applications that work with very big data. Hadoop is implemented on the basis of the computational paradigm MapReduce, according to which the application is distributed to many similar elementary tasks performed on the nodes of the server cluster and then the results of the calculations are combined into a single final result. Information platform HP IDOL (Intelligent Data Operating Layer) [2] is a single platform for working with multimedia information (audio, video, text, social resources, e-mail and web content) and structured machine data (transaction logs or metrics). The platform is based on the software product company Autonomy for the automatic processing of unstructured data, and a high-performance analysis module structured data of company Vertica, which is a part of HP. The HP IDOL software includes functions for processing unstructured data, such as automatic entity extraction (based on machine learning), conceptual data analysis (identifying relationships between data in different systems), data array visualization, cluster analysis. Among the new HP solutions are “HP Big Data Solutions”, “HP Social Media Solutions”, “HP IDOL OnDemand”. The version of “HP IDOL OnDemand” offers users a lot of Big Data web services that developers and clients can use to analyze multimedia data of various types (images, social networks, text, video, etc.). “HP IDOL OnDemand” uses the HP IDOL data analysis platform, which supports functions such as context search, mood analysis and face recognition. “The HP IDOL OnDemand” solution is available to users in the form of a web service.

4. Laboratory workshop ”Extraction and processing of big data”

The laboratory workshop ”Extraction and processing of big data” is intended for training specialists in the specialty 10.05.04 ”Information and Analytical Security Systems”, direction of training: ”Information Security of Financial and Economic Structures” within the framework of the ”Distributed Automated Information Systems” to conduct laboratory work.

The complex consists of a series of laboratory works:

- Languages of queries for information retrieval systems of the Internet on the example of the Yandex system (developed earlier [3])
- Study of analytics services GOOGLETRENDS[4]
- Methods for retrieving big data,
- The use of big data in marketing research,



mephi Search term

master degree Search term

postgraduate Search term

Figure 2: An example of a visual representation of the popularity of queries.

- Methods for processing of big data with HADOOP.
- Search in big data

The complex will be further developed in the direction of using the HP IDOL data analysis platform and the development of visualization, an example of this approach is the publication [5].

5. Exploring the GOOGLETRENDS analytics service

Using the Google Trends service in the laboratory work helps students to track the dynamics of popularity on the search query. The analysis of requests occurs both on a time interval and on regions where the query has been created. The service provides a visual representation of search queries, the possibility of comparing them with each other and evaluating the popularity of queries. The main advantage of the service is convenience and ease of use (Fig. 2).

6. The cycle of laboratory works using the application programming interface

The Reduction API (Application Programming Interface) is an application programming interface that is supported by almost any software product. The API provides support for sets of various functions, classes, structures, libraries and services that allow external applications and services to extract data from WEB applications, such as social networks, cloud storage. The API allows you to apply the data of such services in your own developments. The API provides the provision of information from third-party WEB-applications or WEB-services and the rights to its automated use on a paid or gratuitous basis.

The laboratory work "Methods of extracting big data" teaches students to extract data in the form of textual information from the social network VKontakte using the API technology.

The VKontakte API has about 30 methods. The main ones are: Account, Friends, Groups, Messages, Photos, Video, Wall, Users and auxiliary: Database. The essence of the task is that by using the HTTP request, using the API methods, to get data from the social network. Also, this laboratory work includes the training using the Yandex.Disk API. The essence of the job is to get ACCESS_TOKEN and to work with files on Yandex.Disk. And the last task in this lab is to write a program in Python that calculates the number of people in the community who are currently online. To do this, you need to use a loop, since the getMembers method is constrained to a sample of 1000 users. To do this, you must use the offset parameter, which will indicate the offset required to retrieve a particular subset of the participants. And access to the data takes place in the json format. The peculiarity of obtaining data in this way is that if there are many of them (for example, more than a million), then some delay occurs. For example, 3.7 million of these users were received in 50 minutes. Among the basic data were id, name, surname, online status. To overcome the time delay in obtaining data, code optimization and supercomputers with higher computational performance are used [7].

The laboratory work "Using big data in marketing research" teaches students to process big data for the purpose of using them in marketing research. The essence of the task is to use the API to return information about users of the social networking group VKontakte, then to process this information for clustering by age, geography and gender, and then to evaluate the financial efforts for advertising and make a decision to open a corresponding business. To get specific variables in the API query, use the fields parameter. All required variables are separated by commas. The result of executing the code will be the number of identical variables with the desired value.

The leading technology, which belongs to the Big Data class, is the HADOOP platform [1]. Therefore, the laboratory work "Methods of processing big data with the help of HADOOP" was created. It allows students to be acquainted with a virtual machine that runs the finished assembly HADOOP, called Cloudera. The idea of the task is to download and install Cloudera on a virtual machine. And then, using the command line, run streaming, - a task that counts the number of identical words in the articles.

In order to successfully apply big data, it is necessary to organize a search on them. In laboratory works, there are used replies to VKontakte API requests, which are tagged, but in text format. These data can be accessed in json format as objects. By accessing data, you can apply the search function to them. But such an approach with a large amount of data can lead to a long delay, so it makes sense to upload this data into the SQL database and access the data via an SQL query using the LIKE operator. The students can learn this from the following lab work, which is called "Search in big data." When performing a keyword search, the LIKE statement must have a keyword. For a user input of keywords, you can create a web interface, where you will find a form for entering a keyword and a search button, after clicking which you will go to another page with results. This interaction between SQL and the web interface is done using the following languages: PHP and HTML. The search is carried out by the content of the marked data fields. The fields are the name, surname, city, gender. That is, if the entered keywords are not found in the name column, then the search will go by the last name column and so on. Using SQL, the search will be executed almost instantly.

7. Results

At the Department of Financial Monitoring of the Institute of Financial and Economic Security, NRNU MEPhI, the laboratory workshop "Extraction and processing of big data" was developed and is being tested in the framework of the program for training specialists in the specialty 10.05.04 "Information and Analytical Security Systems", direction of training: "Information Safety of Financial and Economic Structures" within the framework of the educational discipline "Distributed Automated Information Systems". Laboratory works clearly show how to analyze data in big data, extract data using the API, and process it using the Python language.

References

[1] Big Data from A to Z. Part 2: Hadoop, https://habrahabr.ru/company/dca/blog/268277/

[2] Autonomy IDOL (Intelligent Data Operating Layer) // URL: http://www.tadviser.ru/index.php/Продукт :HP_Autonomy_IDOL_(Intelligent _Data_Operating_Layer) // 2014 г.

[3] Kletsova T.V., Prokhorov I.V. Information technology: spreadsheets and search engines. Laboratory practical work. National Research Nuclear University "MEPhI",

Moscow, 2011

[4] O. T. Kochetkov, I. V. Prokhorov, The Research of Approaches of Applying the Results of Big Data Analysis in Higher Education. INFORMATION TECHNOLOGIES IN EDUCATION OF THE XXI CENTURY (ITE-XXI): Proceedings of the International Scientific-Practical Conference "Information Technologies in Education of the XXI Century" Moscow, Russia 7–8 December 2015, ISBN: 978-0-7354-1463-1, Editors: Boris G. Kiselev and Oleg A. Panin Volume number: 1797, Published: AIP Conference Proceedings Jan 5, 2017, http://aip.scitation.org/toc/apc/1797/1?expanded=1797

[5] V. D. Kolychev, I. V. Prokhorov, Application of IT-technologies in visualization of innovation project life-cycle stages during the study of the course "Management of innovation projects". INFORMATION TECHNOLOGIES IN EDUCATION OF THE XXI CENTURY (ITE-XXI): Proceedings of the International Scientific-Practical Conference "Information Technologies in Education of the XXI Century" Moscow, Russia 7–8 December 2015, ISBN: 978-0-7354-1463-1, Editors: Boris G. Kiselev and Oleg A. Panin Volume number: 1797, Published: AIP Conference Proceedings Jan 5, 2017. http://aip.scitation.org/toc/apc/1797/1?expanded=1797

[6] Use big data in marketing research http://www.ovtr.ru/stati/bolshie-dannye-big-data-v-marketingovyh-issledovaniyah

[7] Why Supercomputers Are Needed, http://smb.ixbt.com/__buggypage/