The development of competence in practical work with big data in mastering the master's program in the field of "Business Informatics"

N.V. Morozov, A.N. Norkina, I.V. Prokhorov, A.A. Filatov

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)

The article provides an analysis of the history of the development of the concept of “big data”, the mastering of technologies based on big data, and the evaluation of educational programs on the qualification of Data Scientist. The author has assessed the competence of “Data Scientist” qualifications, analyzed educational programs for big data on the learning of Data Scientist in the world and in Russia. Based on the life cycle analysis of the technology of extracting, processing and analyzing big data, the authors showed the importance of the practical competencies of the Data Scientist qualification. For the application of big data technologies in business, in marketing research, in management, the authors have shown the need to master practical competences in the field of big data for undergraduates of the master's program "Business Informatics". In the article, the authors provide an overview of academic disciplines for the study of big data in the master's program of the specialty “Business Informatics” and describe and characterize the cycle of laboratory work on practical work with big data. The cycle of laboratory work on practical work with big data includes the following works: "Learning query languages of Internet search engines and using them in marketing information analysis", "Learning search query analytics services Google.Trends, Yandex.WordStat"; “Methods of an API application for extracting big data”; “Using loops to extract big data”; “The use of big data in marketing research”, “Study of the API of the statistical information server of the Moscow Exchange”. The authors analyze the experience of testing the cycle of laboratory work. It is planned to master the "Yandex.Metrika" technology to evaluate the site’s performance using counters of the number and composition of visitors, transitions from competing sites, popular content, time spent on the site, etc. Work is being done on big data as part of educational research, master’s and PhD dissertations.

Keywords: big data, data scientist, extraction of data, data mining, processing and analysis of big data, internet, Yandex, Google, application programming interface, API, social networks

At the present stage of Russia's transition to a “digital economy”, the most important condition for successful activity is the ability to process large arrays and information flows (big data). In this regard, there is a huge need for training specialists - data researchers (Data Scientist), focused on working with big data. Research, modeling and development of information technologies for search, extraction, processing, presentation, visualization and analysis of "big data" for use in business, especially in marketing and management, became possible by ensuring a high level of parallelism in modern data centers. Moreover, it is believed that the development of modern marketing is in the same stage as physics in the Middle Ages, when the transition from descriptive physics models to mathematical models was made. And marketing is now experiencing a period of rapid development based on big data technologies. Decision making in business, in government, in project management is becoming more and more reasonable. The important features of big data are their really large amount, weak structuredness and heterogeneity, as well as the need to process them very, very quickly. Thus, the task is to quickly process large arrays of heterogeneous (multimedia) data, extracting useful infor-

mation for making decisions.

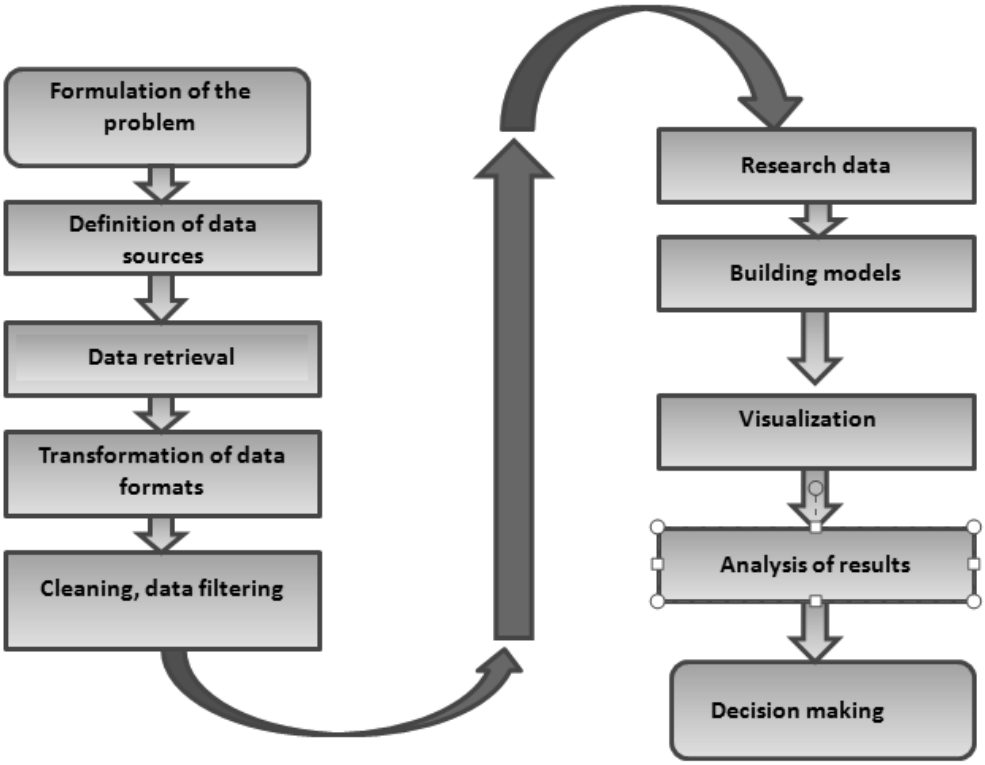

Fig. 1. The life cycle of technology for extracting, processing and analyzing big data

Figure 1 shows the life cycle of a technology (technological sequence of stages) of extracting and processing data from setting an informational task to analyzing the results. At the same time, at the stages of data research and model building, new extraction and analysis of the model of "big data" are often used.

The practical objective is to create a laboratory research base for the formation of competence and proposals on professional standards of the profession «Data scientist», the main which is the practical competence of exploration, mining, extraction, processing, analysis and visualization of big data. This will create a laboratory base for the training of specialists, PhD theses and master's theses in the new Master's program "Processing and Analysis of Big Data" in the specialty of "Business Informatics" with the qualification "Big Data Analyst" or "Data Scientist".

Big Data Issues

The current state of information and communication technologies is characterized by three concepts: virtualization, cloud technologies and "big data". The inevitability of using these information technologies, big data information models, and private and public clouds for data storage is becoming apparent in the modern world, which is often referred to as the term VUCA, which is volatile, unpredictable, complex and ambiguous. VUCA is characterized by the following features:

- Volatility-high variability of information, deliberate inadequacy of information,
- Uncertainty- unpredictability, decision making under uncertainty,
- Complexity - high complexity and non-obviousness of information interpretation,
- Ambiguity - ambiguity of information

Areas of use of big data.

In marketing, this is [1]:

- Customer flow modeling
- Recommender systems

• Market segmentation
• Analysis of social media (websites, blogs, forums, social networks, interacting with the community through multimedia content (text, video, photo).

In the banking and insurance business:
• Credit Risk Analysis
• Predictive analytics
• Fraud prevention
• Definition of abnormal behavior
• The accumulation and use of MDM data (master data)
• Management of risks
• Simulation of insurance claims

Predictive analytics in the banking sector is a technology of informational forecasting of actions of companies and individuals in order to identify groups of citizens at risk of committing financial crimes, suppressing tax fraud, preventing theft and using personal data. The accumulation and use of MDM data (master data) [2] is used for underwriting (studying the solvency of a potential bank borrower). Predictive analytics in the insurance industry is the identification of false insurance claims.

Data analysis at NRNU MEPhI has been engaged for a long time (including at the department of financial monitoring, the “Data analysis” discipline is being studied) [3,4,5,6]. However, the analysis is limited to data volumes, measured in megabytes, and at best a few gigabytes. At the same time it is - tabular data. Data Science Specialists work not only with structured data, but also with semi-structured and unstructured data of huge volume, measured in terabytes and petabytes.

Tools for working with big data.

For the analysis of big data, specialized tools and models are created using such products as Hadoop MapReduce, HP IDOL, and others. Hadoop [7] is a framework designed for building distributed applications that work with very large data. Hadoop is implemented on the basis of the MapReduce computational paradigm, according to which the application is distributed to many similar elementary tasks performed on the server cluster nodes and then the results of the calculations are combined into a single final result. The HP IDOL Information Platform (Intelligent Data Operating Layer) [8] is a unified platform for working with multimedia information (audio, video, text, social resources, e-mail and web content) and structured machine data (transaction logs or meter readings) . The platform is based on the Autonomy software for automatic processing of unstructured data, and the high-performance module for analyzing structured data from Vertica, a company included in 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. New HP solutions include HP Big Data Solutions, HP Social Media Solutions, HP IDOL OnDemand. The version offers users many Web services Big Data that developers and customers can use for analyzing multimedia data of various types (images, social network data, text, video, etc.). HP IDOL OnDemand uses the HP IDOL data analysis platform, which supports such functions as contextual search, mood analysis and face recognition. HP IDOL OnDemand is available to users in the form of a web service [8].

The complex of practical classes on the search, extraction, processing and analysis of big data in marketing research and business management.

The complex of practical classes on the search, extraction, processing and analysis of big data in marketing research and business management is intended for the training of specialists in the master's program in the specialty of training "Business Informatics". The complex consists of a series of practical sessions in the computer lab:

• The study of query languages of Internet search engines and their use in marketing information analysis;
• Study of analytics services GOOGLETRENDS, Yandex.WordStat;
• Application API methods for extracting big data;
• Using loops to extract big data;

• The use of big data in marketing research;
• Methods of processing big data using HADOOP;
• Study of the API of the statistical information server of the Moscow Exchange [9];
• It is planned to master the Yandex.Metrika technology to evaluate the site's performance using counters of the number and composition of visitors, referrals from competing sites, popular content, time spent on the site, etc.

The complex of practical exercises is planned to be further developed in the direction of using the HP IDOL data analysis platform [8] and the development of visualization (as, for example, in the publication [6]). For three years, the department carried out approbation of the practical training cycle. Tasks on the application of big data in the framework of educational research works, master's and PhD theses were carried out.

• Practical exercise "Learning the languages of Internet search engine requests and their use in marketing information analysis.

The practical exercise was developed on the basis of the previously developed laboratory work "Request Languages for Information Retrieval Systems of the Internet on the example of Yandex [4]. Students learn the formal query languages of search engines Yandex, Google and mail.ru, as well as the query language, close to natural language. Students compare the capabilities of search engines, study information retrieval strategies and apply information analysis in the marketing of modern information technologies.

• Practical exercise: "The study of intelligence GOOGLETRENDS service"

Using Google Trends service to the lab helps students keep track of the dynamics of popular trend, built on user searches. The analysis of requests occurs both by time interval and by region where the request was created. The service provides a visual representation of search queries, the ability to compare them among themselves and assess the popularity of marketing areas based on user requests (Fig. 2). The main advantage of the service is convenience and ease of use.

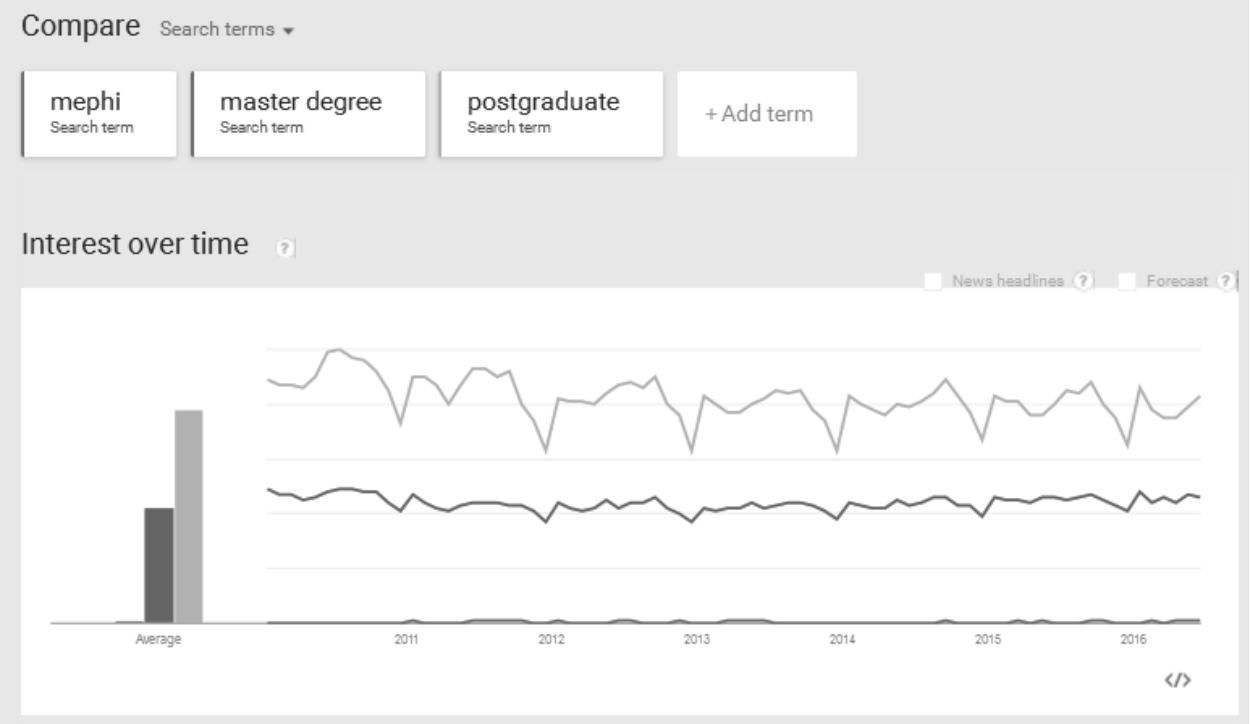

Fig.2. An example of a visual representation of the popularity of requests

• A series of practical exercises using an application software interface

The acronym 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, with which external applications and services allow you to extract data from WEB applications, such as social networks, cloud storages. The API allows you to use the data of such services in their own development. The API provides the provision of in-

formation from third-party WEB-applications or WEB-services and the rights to its automated use on a paid or free of charge basis.

At the lab, "Methods of application programming interface API to extract big data," students learn how to retrieve data in the form of text information from the social network VKontakte using API technology. The VKontakte API includes about 30 methods. The main ones are: Account, Friends, Groups, Messages, Photos, Video, Wall, Users and auxiliary: Database. The task is to retrieve data from the social network using an HTTP request using API methods. Also, this lab includes training in the use of the Yandex.Disk API. The task is to get ACCESS_TOKEN and work with files on Yandex.Disk from an application program. When retrieving large amounts of data from a social network, students write a program in Python that counts 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 limited to a sample of 1000 users. To do this, use the offset parameter, which will indicate the offset required to select a specific subset of participants. The data is accessed in json format. If there is a lot of data, for example, more than a million, then there is some time delay. To overcome the time delay of data acquisition, code optimization and more productive computing resources are used.

The practical lesson “Using big data in marketing research” teaches students how to process big data in order to use it in marketing research [1]. Students, using the API, receive information about the users of the social network group VKontakte, then process this information using clustering by age, geography and gender, and then estimate the financial costs of advertising and make a decision to open the relevant business. To get specific variables in the API query, students use the fields parameter. The result of the code execution will be the number of identical variables with the desired value.

The leading technology that belongs to the Big Data class is the HADOOP platform [7]. Practical lesson “Methods of processing big data using HADOOP” allows students to get acquainted with the virtual machine on which the ready-made HADOOP assembly is launched, which is called Cloudera. The idea of the job is to download and install Cloudera on a virtual machine. And then using the command line to start streaming - a task that calculates the number of identical words in articles.

Results

The Department of Financial Monitoring of the Institute of Financial and Economic Security of the National Research Nuclear University MEPhI has developed and tested a set of practical exercises on finding, extracting, processing and analyzing big data in marketing research and business management, as part of the program for training specialists in the master program in the Business Informatics program . A set of practical exercises clearly shows how to analyze in big data, extract data using an API, and process it using the Python language.

References

1. Use big data in marketing research http://www.ovtr.ru/stati/bolshie-dannye-big-data-v-marketingovyh-issledovaniyah
2. Igor Prokhorov, Nikolai Kolesnik. Development of a master data consolidation system model (on the example of the banking sector)// Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2018 (Ninth Annual Meeting of the BICA Society)// Procedia Computer Science 00 (2019) 000–000
3. Prokhorov I. V., Kochetkov O. T., and Filatov A. A..Practical Training of Students on the Extraction and Analysis of Big Data\\ KnE Social Sciences | III Network AML/CFT Institute International Scientific and Research Conference "FinTech and RegTech" | pages: 361-368 https://knepublishing.com/index.php/Kne-Social/article/view/1565/3697
4. Kletsova T.V., Prokhorov I.V. Information technology: spreadsheets and search engines. Laboratory practical work. National Research Nuclear University "MEPhI", Moscow, 2011
5. 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 num-

ber: 1797, Published: AIP Conference Proceedings Jan 5, 2017, http://aip.scitation.org/toc/apc/1797/1?expanded=1797
6. 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
7. Big Data from A to Z. Part 2: Hadoop, https://habrahabr.ru/company/dca/blog/268277/
8. Autonomy IDOL (Intelligent Data Operating Layer) // URL: http://www.tadviser.ru/index.php/Продукт:HP_Autonomy_IDOL_(Intelligent_Data_Operating_Layer) // 2014 г.
9. Software interface to the Moscow Exchange Information and Statistical Server http://www.moex.com/a2193