ANALYTICAL PLATFORM FOR SOCIO-ECONOMIC STUDIES

S.D. Belov[1,2,a], A.V. Ilina[1], J.N. Javadzade[1,3], I.S. Kadochnikov[1,2], V.V. Korenkov[1,2], I.S. Pelevanyuk[1,2], V.A. Tarabrin[2], P.V. Zrelov[1,2] and R.N. Semenov[1,2]

[1] *Joint Institute for Nuclear Research, 6 Joliot-Curie st., Dubna, 141980, Russia*

[2] *Plekhanov Russian University of Economics, Stremyanny lane 36, Moscow, 117997, Russia*

E-mail: [a] belov@jinr.ru

Started in natural sciences, the high demand for analyzing a vast amount of complex data reached such research areas as economics and social sciences. Big Data methods and technologies provide new efficient tools for research. In this paper, we discuss the main principles and architecture of the digital analytical platform aimed to support socio-economic applications. Integrating specific open-source solutions, the platform intended to cover full-cycle data analysis and machine learning experiments, from data gathering to visualization. One of the system's primary goals is to deliver the advantage of the cloud and distributed computing and GPU accelerators with Big Data analysis techniques. The authors present the approach of building the platform from low-level services such as storage, virtual infrastructure, pass-through authentication, up to data flows processing, analysis experiments, and results representation.

Keywords: Big Data platform, socio-economic studies, machine learning

Sergey Belov, Anna Ilina, Javad Javadzade, Ivan Kadochnikov, Vladimir Korenkov,
Igor Pelevanyuk, Roman Semenov, Vitaliy Tarabrin, Petr Zrelov

1. Introduction

The processes of Big Data analysis in different areas, despite some peculiarities, are pretty similar. Analytics solutions and methods are widely used and can be successfully used in various fields of science. A platform-based approach for creating a software and hardware environment seems promising, in which there are both basic, infrastructure components common to information flows of all classes of tasks, and specialized services that improve the characteristics (for example, speed or quality) obtained in a particular area of scientific and practical results. A generalized architecture of an automated analytical system was proposed to solve problems requiring both streaming and batch processing of large amounts of data or having great internal complexity, including implicit connections. To build each functional level of the platform, the open-source software products were selected, primarily from the Big Data technology stack.

2. Labour market monitoring project

Recently, the prospects for the "digitalization" of economic processes have been actively discussed. This is a challenging task that cannot be solved within the framework of classical methods. The prospects for their qualitative development are illustrated in the article by the example of using big data analytics and text mining to assess the labor demand of regional labor markets. Another critical issue is the study of the interaction of the labor market and the vocational education system [1]. The problem was solved using the automated information system developed by the authors for monitoring the compliance of the personnel needs of employers with the level of training of specialists. The information base for collecting information is open source. The presented system creates additional opportunities for identifying qualitative and quantitative relationships between the education sector and the labor market. It is aimed at a wide range of users: authorities and administrations of regions and municipalities; management of universities, companies, recruiting agencies; graduates and graduates of universities.

The purpose of introducing information systems for monitoring and forecasting the situation in the labor market and analyzing staffing needs is to provide additional opportunities for identifying qualitative and quantitative relationships between education and the labor market. The system is designed for a wide range of users and is intended primarily for heads of regions, universities, companies, recruitment agencies. It is expected that the project will provide a closer connection between the education system in the country and the labor market, provide an opportunity to adjust curricula, open new educational programs, or adjust existing ones in accordance with the country's economic goals, and allows regions to implement effective recruitment and training. After that, it is assumed that the system will become a useful tool for young professionals who are starting to look for a job in their chosen profession, as well as for those choosing a profession.

The following Internet resources are used as a source of data on vacancies: the portal "Work in Russia" (information site of the Russian Labor Agency), portals of the staffing companies HeadHunter and SuperJob. In addition, the register of approved professional standards and Federal state educational standards of higher professional education are used as guiding documents [2]. The subject of a separate study is assessing how job advertisements reflect the real needs of the market.

The implemented prototype of an automated information system is a web application with an intuitive user interface that provides reliable data storage.

The system is built on a modular basis. Firstly, it is a text data collection module (working in automatic mode using open sources - Internet portals and recruiting agencies).

Secondly, the load module and data store, consisting of a distributed data store (provides replication and archiving).

Third, an automatic processing module that prepares information for analysis, automatic linking of requirements and competencies, and machine learning.

Fourth, user interfaces to generate and display reports based on business intelligence technologies.

Basic information about the state of the labor market is obtained by analyzing the database of collected vacancies. To obtain correct statistics, it is necessary, first of all, to solve the following tasks:

- Search for duplicate vacancies. Even if one is using one source, job advertisements can be duplicated, but such checks are necessary if one is using multiple sources.
- Classification of vacancies by industry.
- Analysis of the content of the job offer, analysis of individual requirements for skills and competencies.

The data processing schema if shown in [fig. 1].

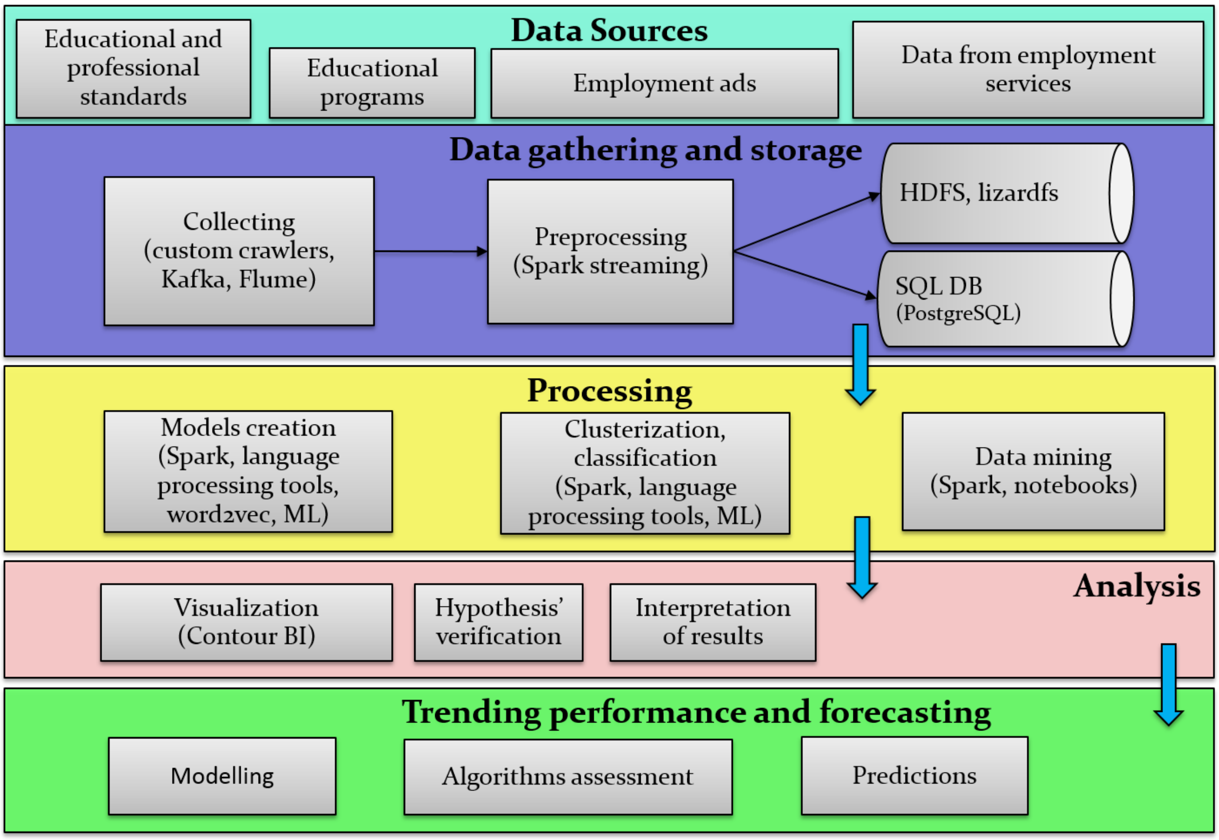
Data Sources
Educational and professional standards
Educational programs
Employment ads
Data from employment services
Data gathering and storage
Collecting (custom crawlers, Kafka, Flume)
Preprocessing (Spark streaming)
HDFS, lizardfs
SQL DB (PostgreSQL)
Processing
Models creation (Spark, language processing tools, word2vec, ML)
Clusterization, classification (Spark, language processing tools, ML)
Data mining (Spark, notebooks)
Analysis
Visualization (Contour BI)
Hypothesis' verification
Interpretation of results
Trending performance and forecasting
Modelling
Algorithms assessment
Predictions

Figure 1. Data processing for labour market analysis project

3. Analysis of links between companies

The project [3] aims to create a database of companies and data on companies and an automated analytical system based on these data. The development of the system will allow credit institutions to obtain information on relationships between companies, pursue the "Know Your Client" policy - to identify the ultimate beneficiaries, assess risks, and identify relationships between clients. This may be the need for banks to comply with the requirements of national authorities, laws on tax evasion in offshore and FATCA, recommendations of the Financial Action Task Force on Money Laundering (FATF), the Basel Committee on Banking Supervision. At the moment, there are some projects, such as OpenCorporates [4], which have global databases of companies collected from many jurisdictions. Nevertheless, at the same time, they do not cover all national registries or other helpful data sources (courts, customs, press, etc.). In addition, existing services have a relatively meager ability to find relationships between companies, which are not always straightforward. The project we are presenting aims to overcome the main of these shortcomings. The number of companies worldwide is over 150 million. With information about a company from many sources, there is no other reasonable way to process it using big data technologies. We use such technologies in our research along with machine learning and graphical databases.

To identify the affiliation of companies and the direct comparison of relationships through founders and owners, an analysis of indirect indicators is used. We are considering companies that have a match in several positions. First, fragments of name, officers, founders, registration address, contact information, owners, subsidiaries, historical ties, similarities in company names and profiles, etc., in addition, it uses previously found relationships. Discovered information about certain connections of companies is stored in a graphical database, in which records are both about the company and other types of objects (officials, founders, registration address, contact information). This approach allows for more flexible link analysis and complex search queries. The graph base Neo4j [5] is used to analyze and store the identified links. This database also allows one to visualize graphical relationships using built-in tools.

4. Analytical platform

A generalized architecture of an automated analytical system was proposed to solve problems requiring both streaming and batch processing of large amounts of data or having great internal complexity, including implicit connections. To build each functional level of the platform, a set of open-source software products was selected, primarily from the Big Data technology stack. The architecture of the proposed solution is shown in [fig. 2].

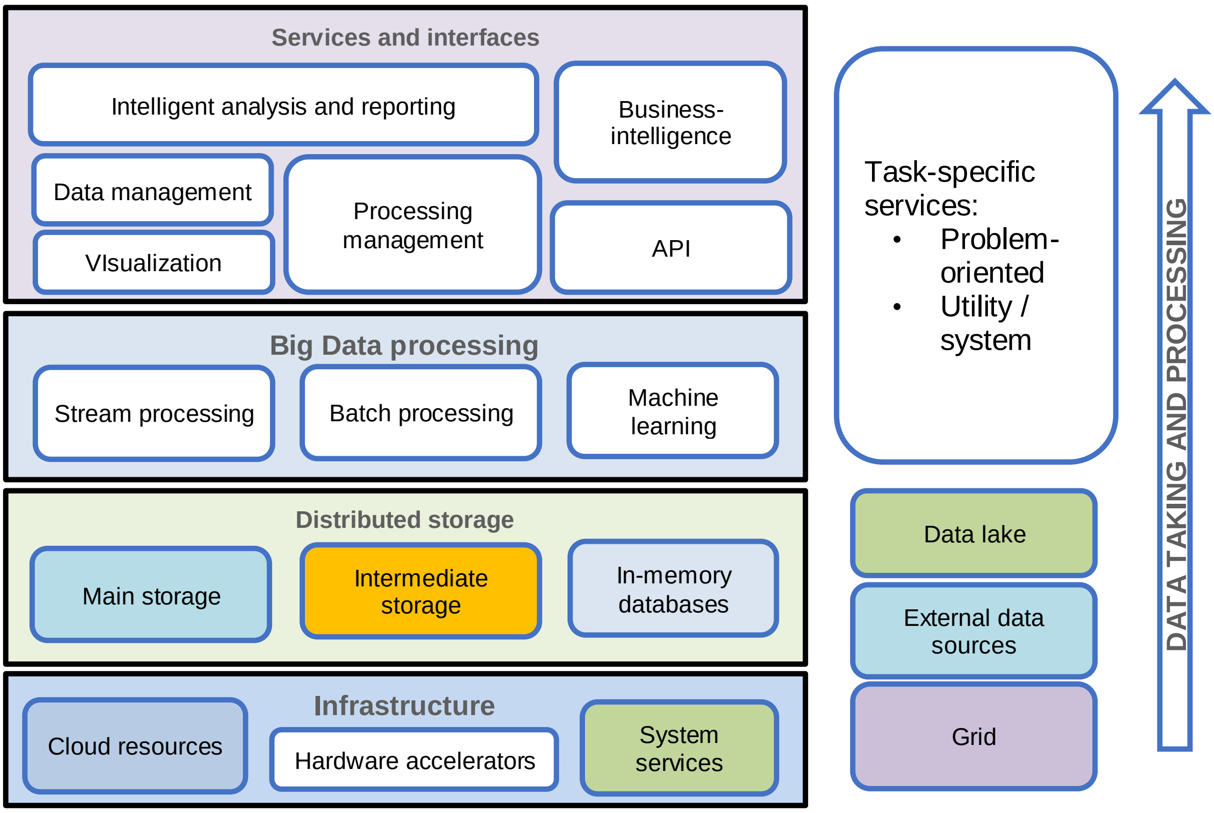

Figure 2. General scheme of the analytical platform

The platform is based on the open-source software solutions. Its modular structure allows replacing particular components if needed. Chosen packages are shown in the Table 1 below.

Table 1. Software stack of the platform.

Layer	**Software packages**
Visualization and system interfaces	Zeppelin, Jupyter (user interface) Graphana (reporting and graphical presentation of results) KrakenD (organization of software gateways for various components)
Distributed Big Data analytics	Apache Kylin
Computational Experiments in ML	MLflow
In-memory computations	Apache Spark, Dask, Hadoop
Organization of the process of data flow management and data collection	Apache Kafka, Apache Flume, Apache Airflow, Celery, Scrapy
Data vaults and specialized databases	CEPH, NFS (хранение и доступ к файлам) Elasticsearch (structured data indexing and analysis) Apache Ignite (in-memory database for fast access and caching) Russian Data Lake Apache Calcite (dynamic data management and integration)
Authentication and passthrough authorization, security	Free IPA, Vault
Computing infrastructure, resource management	OpenNebula, Kubernetes, Docker, Puppet, Git

5. Conclusion

Based on the experience of using big data technologies, a schematic of an analytical platform for performing socio-economic research was proposed. In addition, the selection of open-source software for building a modular analytical platform that allows analyzing Big Data using machine learning and hardware accelerators has been performed.

6. Acknowledgement

The study was carried out at the expense of the Russian Science Foundation grant (project No. 19-71-30008).

References

[1] A. Wolf, Review of Vocational Education // The Wolf Report, 2011

[2] Professional standards in Russia – [Web resource]. – http://profstandart.rosmintrud.ru

[3] Badalov L.A.et al., Checking foreign counterparty companies using Big Data, CEUR Workshop Proceedings, 2018, vol. 2267, pp. 523–527

[4] OpenCorporates: The Open Database Of The Corporate World — Available at: https://opencorporates.com/

[5] Neo4j graph database. Available at: https://neo4j.com/