CHECKING FOREIGN COUNTERPARTY COMPANIES USING BIG DATA

L.A. Badalov [1], S.D. Belov [1,2, a], I.S. Kadochnikov[1,2]

[1] *Plekhanov Russian University of Economics, 36 Stremyanny per., Moscow, 117997, Russia*

[2] *Laboratory of Information Technologies, Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russia*

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

The project aims to create a database of companies and company data and an automated analytical system based on this data. The development of the system will allow credit institutions to obtain information about the links between companies, to carry out a policy of "Know your customer" - to identify the final beneficiaries, to assess risks, to identify relationships between customers. It could the need of banks to fulfill the requirements on national authorities, laws on offshore tax evasion and FATCA, the recommendations of the Group of development of financial measures of struggle against money-laundering (FATF), the Basel Committee on banking supervision. For the moment, there are some projects like OpenCorporates having global databases of companies collected from many jurisdictions. But at the same they don't cover neither all the national registries, nor other useful data sources (courts, customs, press, etc.). Also, the existing services have rather sketchy abilities on searching for relations between companies, which are not always direct. The project we present is about to overcome main of these deficiencies. Number of companies worldwide is more than 150 million. Having company information from many sources, there is no other reasonable way to process it using Big Data technologies. In the research we use such technologies along with machine learning and graph databases.

Keywords: companies controlling, money laundering, finances, Big Data

1. Introduction

The integration of the Russian financial system into the international system has led to the emergence of a wide range of regulatory documents of the Bank of Russia on disclosure of information about the ultimate beneficiaries of non-resident clients, as well as other requirements. In addition, deoffshorization laws and FATCA [1] have made it necessary for banks to have information about the ultimate beneficiaries. The American law on taxation of foreign accounts – (Foreign Account Tax Compliance Act-FATCA) not only obliges Russian banks to identify American taxpayers among their customers, but also threatens large fines in case of non-execution. The risks associated with possible sanctions by the us internal Revenue service (IRS) are extremely high banks can be fined in the form of compulsory withdrawal of 30% of the amounts of international transfers. Thus, us law establishes obligations for credit institutions and measures for non-performance or improper performance of such obligations but does not provide the tools with which to comply with such requirements.

Another responsibility that is imposed on credit institutions is the need for regular re-examination of the entire customer base in order to identify non-resident customers and customers-foreign taxpayers, which in the absence of information about the beneficial owners is equivalent, in fact, re-identification.

The main result of these responsibilities of banks is the ability to confirm or refute the reality of non-resident clients, that is, to establish whether they are real activities or technical companies that can be used for questionable transactions.

2. Information sources

2.1. Classification of the open data sources

The work on the analysis of the main available sources of information on several jurisdictions, the definition of the nature and completeness of the data contained in them, technical ways of access to them. Number of companies worldwide is more than 150 million [2].

Currently, the system for analysis considers information from the following types of open sources:

- National register of companies;
- Portals aggregating various companies' data (like OpenCorporates [3]);
- Financial oversight services databases;
- Global database of legal entity identifiers (GLEIF [4]).

In the future, it is also planned to use the following sources:

- Tax service data;
- Customs information;
- Court decisions databases;
- Data leakage of information from offshore, the results of investigative journalism;
- Databases of distressed companies (temporarily managed, in the process of closing, etc.));
- Other sources of commercial information (persons, beneficiaries, public documents on administrative proceedings, etc.).

At the moment, there are 40 considered main data sources.

2.2. Data consolidation

Data collection and processing is carried out based on modern methods and technologies for obtaining information from web-based sources.

Depending on the data source, three main types of queries are available:

- Retrieving all or part of the information as archive files;
- Programmatic access to sources using HTTP requests;
- Search for information about a company or companies using the source website.

The first method, the use of archives, makes it possible to quickly obtain significant amounts of fairly complete information about companies.

The second method, programmatic access to sources, involves primarily the search for data about a particular company. At the same time, obtaining information about all companies available in the register requires additional scanning and considerable time.

The third method, the extraction of information from web site source, is the most time-consuming web sites are designed for visual perception by a person, machine-readable information varies greatly from source to source, and can be assumed to be additional software processing on the client side, etc. So often to retrieve information from sources of this type requires the development of separate specialized modules that is implemented in the framework of the project.

2.3. Data gathering and pre-processing

The system uses two types of information gathering from the sources: scheduled and on request. The scheduled one is implemented as a periodically run job, performing the downloading of the archives, scanning the sources for information about the largest possible number of companies, as well as the updating of information on subjects already entered into the database. Obtaining information on request is carried out when the user accesses the system to obtain relevant information from sources and save it in the database. Special modules have been developed to collect the information from the sources on the Internet.

At the stage of pre-processing information is structured, highlights the main fields necessary for further preservation and analysis of relationships with other companies.

Basic information about the company includes:

- Name (including previous names);
- The IDs of the company in the registers;
- Jurisdiction;
- Company status;
- Form of organization;
- Registration date;
- Legal address, other contact information;
- CEOs and other management officials;
- Founders, owners, subsidiaries;
- Links to the source of the source information.

3. Information processing and analysis

3.1. Revealing links between companies

To identify the affiliation of the companies, in addition to direct comparison of relationships through the founders and owners, the analysis of indirect signs is used. We consider companies that have a coincidence in several positions. First, the fragments of the name, officers, founders, registration address, contact information, owners, subsidiaries, historical ties, similarity of the names and company profiles, etc. in addition, it uses the previously found relations.

Discovered information about those or other links of the companies is stored in a graph database, entries in which both the company and the other object types (officers, founders, registration address, contact information). This approach allows for more flexible link analysis and complex search queries.

For the analysis and storage of the revealed connections the graph base Neo4j [5] is used. This database also allows you to visualize the graph links using built-in tools.

2.3. Data gathering and pre-processing

In the implementation of the software infrastructure of the system used a stack of software products and tools that have become de facto industry standards in their areas: Spark [6, 7], Hadoop, Kafka, Flume, Marathon, Docker, Elasticsearch. The deployment of clusters of these basic components allows for scalability and high availability of the system. The technologies and algorithms used make it easy to develop collection and analysis tools and to connect new structured and unstructured data sources.

Aggregation and buffering input data and processed data are the means used to organize data flows software package, Apache Flume and system data buffering Kafka. These products have shown their reliability and stability when working in high-load systems.

The data collection tools are developed in Python using public open source software libraries. Django software platform is used as a basis for the organization of the graphical user interface in the form of a website. All user requests are made on the server side.

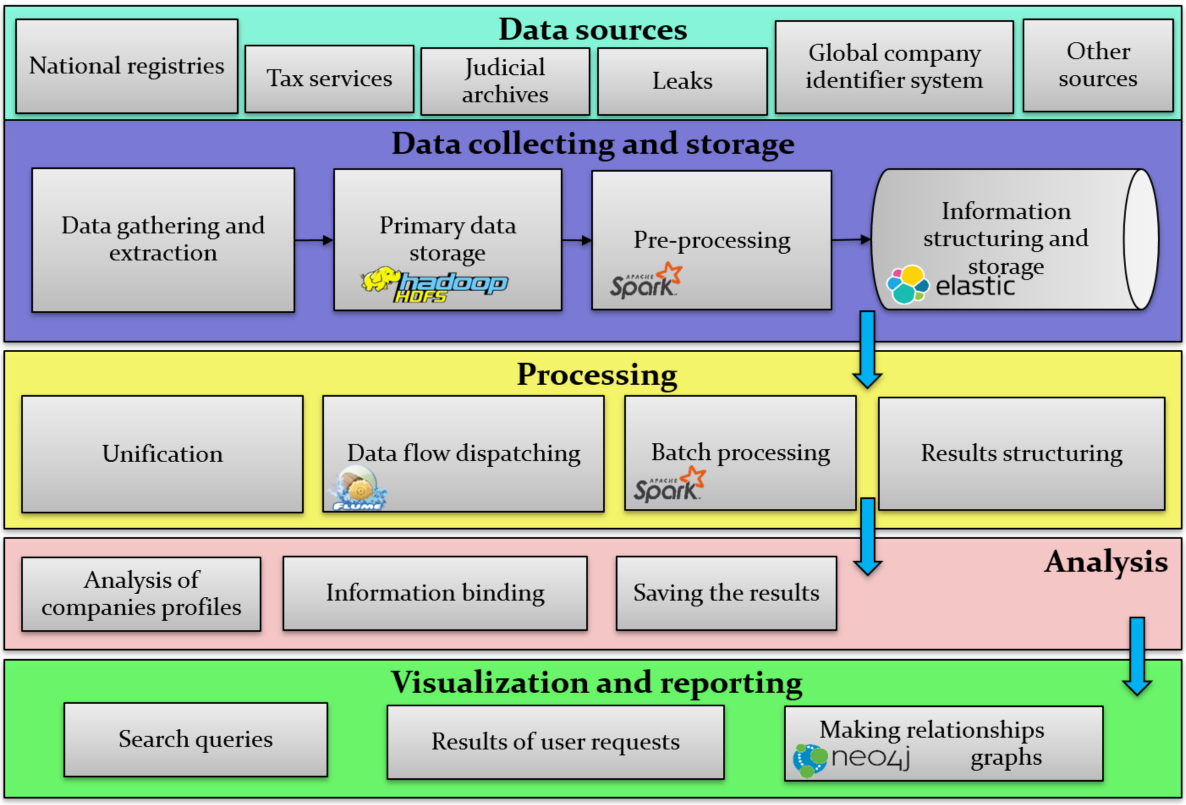

Figure 1. Information processing workflow and tools

The scheme of interaction of software infrastructure components is shown in Figure 1. For the data processing, the original program code developed during the project implementation are used.

4. Acknowledgement

The work of Sergey Belov and Ivan Kadochnikov was partially supported by the Russian Foundation for Basic Research (RFBR), grant 18-07-01359 "Development of information-analytical system of monitoring and analysis of labour market's needs for graduates of Universities on the basis of Big Data analytics".

5. Conclusion

Modern banking is impossible without the use of information systems. Banks are actively using several modern IT program systems to solve current problems. Bank controlling is no exception and advanced IT solutions are needed to effectively implement its functions. This will provide banking supervision with significant coverage of credit institutions of different scale-from non-Bank credit institutions and banks with basic licenses to banks with universal licenses.

We believe that the gradual and well-thought-out implementation of all the measures proposed in the work will contribute to the growth of the quality and role of Bank controlling, and, as a result, will entail an increase in the stability and reliability of the entire banking sector.

In the framework of the study, considering the regulatory framework of regulators, approaches and methods were developed to automate the receipt of information about non-resident companies operating in the territory of the Russian Federation, as well as to make decisions on companies-contractors.

As a result of the work, an information and analytical system was created that allows:

- Search for information on non-resident companies using open sources in various jurisdictions;
- Identify links between companies based on different information about them;
- Identify the ultimate beneficiaries, if possible.

During the operation of the system, the internal database is also increased when receiving data from external sources. In addition to be valuable, the company information database, which is compiled from many different sources, also provides a unique opportunity to automate the acquisition of new knowledge, such as links between companies registered in different jurisdictions around the world.

Based on the developed system can be deployed information service that provides a range of services for controlling non-resident companies and decision-making on specific companies in the key requirements of regulators. The software infrastructure of the system, created based on open-source products of industrial level and used in various sectors of economic activity, makes it possible both to expand the functionality of the created system and to scale it under large volumes of processed data and requests of users of services.

References

[1] Foreign account tax compliance act FATCA. Available at: https://www.irs.gov/businesses/corporations/foreign-account-tax-compliance-act-fatca

[2] World Bank, number of companies worldwide. Available at: https://data.worldbank.org/indicator/CM.MKT.LDOM.NO

[3] OpenCorporates database. Available at: https://opencorporates.com

[4] Global Legal Entity Identifier (GLEIF). Available at: https://www.gleif.org/

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

[6] M. Zaharia et al., Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. NSDI 2012. April 2012.

[7] M. Armbrust et al., Spark SQL: Relational Data Processing in Spark. SIGMOD 2015. June 2015.