

УДК 004.62, 004.042

ПРИМЕНЕНИЕ ТЕХНОЛОГИЙ БОЛЬШИХ ДАННЫХ ДЛЯ ОРГАНИЗАЦИИ СБОРА, ПОТОКОВОЙ ОБРАБОТКИ И ХРАНЕНИЯ ИНФОРМАЦИИ О КОМПАНИЯХ-НЕРЕЗИДЕНТАХ

Папоян Владимир Владимирович¹, Кореньков Владимир Васильевич²,
Кадочников Иван Сергеевич³

¹Студент;
ГБОУ ВО МО Университет «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: vlapoyan@jinr.ru.

²Директор, доктор технических наук, профессор;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
ГБОУ ВО МО Университет «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: korenkov@jinr.ru.

³Старший научный сотрудник, инженер-программист;
Российский экономический университет им. Г.В. Плеханова,
Лаборатория «Облачных технологий и аналитики Больших данных»;
117997, г. Москва, Стремянный пер., 36;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
e-mail: kadivas@jinr.ru

Перед банками стоит задача установить, ведут ли их клиенты реальную деятельность, или они являются техническими компаниями, которые могут использоваться для сомнительных операций. Одним из решений вышеуказанной проблемы является разработка национального информационного ресурса контроллинга деятельности компаний-нерезидентов. Разработка эффективного ресурса предполагает использование технологий больших данных. В рамках настоящей статьи рассматривается процесс сбора, потоковой обработки и хранения данных о компаниях из национальных реестров. Применение описанных в настоящей статье технологий и методов позволяет добиться стабильности функционирования системы и упрощения ее поддержки. Реализация была протестирована на двух национальных реестрах компаний Companies House и The Insolvency Service.

Ключевые слова: большие данные, синтаксический анализ, потоковая обработка данных, Apache Kafka, Apache NiFi.

THE APPLICATION OF BIG DATA TECHNOLOGIES FOR ORGANIZATION OF EXTRACTING, FLOW AND STORE DATA ABOUT NON-RESIDENT COMPANIES

Papoyan Vladimir¹, Korenkov Vladimir², Kadochnikov Ivan³

¹Student;
Dubna State University,
Institute of the system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
e-mail: vlapoyan@jinr.ru.

²Director, Doctor of Technical Science, professor;
Joint Institute for Nuclear Research,
Laboratory of Information Technologies;

141980, Dubna, Moscow reg., Joliot-Curie, 6;
Dubna State University,
Institute of the system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
e-mail: korenkov@jinr.ru.

³Senior researcher, engineer-programmer;
PLEKHANOV Russian University of Economics,
117997, Moscow, Stremyanny lane, 36,
Joint Institute for Nuclear Research,
Laboratory of Information Technologies;
141980, Dubna, Moscow reg., Joliot-Curie, 6;
e-mail: kadivas@jinr.ru

Banks need to establish if their clients are tax evasion companies or run real business. The development of national information resource for retrieval and analysis of information on non-resident companies is one of the key to solve the problem. The application of Big Data technologies is necessary for implementation of the efficient resource. Therefore, problems such as extracting, flow and stores data about companies from national registers are considered in the article. The application of technologies and approaches described in this article allow to achieve stable performance and support facilitate of the system. The implementation is tested on two registers of companies The Insolvency Service and Companies House.

Keywords: Big Data, web scraping, data flow, Apache Kafka, Apache NiFi.

Введение

В современных условиях оценка реальности деятельности клиентов-нерезидентов имеет все большее значение для Центрального банка Российской Федерации. Банки и контролирующие органы сталкиваются с серьезными трудностями в определении реальности деятельности клиентов-нерезидентов.

Для решения сложившейся ситуации банковские учреждения совместно с научно-исследовательской лабораторией «Облачных технологий и аналитики Больших данных» Российского экономического университета имени Г.В. Плеханова достигли соглашения о разработке информационного ресурса контроллинга деятельности компаний-нерезидентов.

На одном из этапов разработки настоящего ресурса требуется организовать систему, задачами которой является сбор и сохранение информации о компаниях-нерезидентах из национальных реестров в различных юрисдикциях.

Реестр компаний – это реестр организаций в юрисдикции, под которой они работают [1]. В зависимости от страны реестр компаний имеет различные типы регистраторов, содержания, назначения и общедоступности. Реестры могут находиться под управлением судов, правительственных учреждений или торговых палат, представлены в электронном виде. Доступ к ним осуществляется посредством перехода на специально созданные веб-сайты.

Общее количество действующих компаний, зарегистрированных в различных юрисдикциях, порядка сотни миллионов. Для организации потоковой обработки и хранения рассматриваемого количества информации необходимо применение технологий Больших данных. В связи с этим используются такие свободно распространяемые программные продукты, как *Apache Kafka*, *Apache NiFi* и *HDFS*, ставшие промышленным стандартом для организации систем обработки больших объемов данных.

Помимо организации потоковой обработки и хранения информации о компаниях-нерезидентах, необходимо осуществить сбор и структурирование данных из национальных реестров компаний. Задача сбора и структурирования данных, добываемых из Интернета, делится на две отдельные фазы:

1. перебор веб-страниц (англ. *crawling*);
2. извлечение данных из веб-страниц (англ. *scraping*) [2];

Таким образом, первым этапом в организации системы является добыча данных из веб-ориентированных источников. В качестве инструментов для сбора данных из веб-страниц выступает технология *Scrapy* и библиотеки языка программирования *Python: Requests* и *BeautifulSoup*.

1. Сбор данных о компаниях-нерезидентах из национальных реестров

В ходе анализа национальных реестров компаний был сделан вывод о том, что их можно классифицировать по способу доступа к данным. В этом случае реестры делятся на две группы:

1. реестры, в которых данные хранятся в виде файла;
2. реестры, в которых данные представлены в формате *HTML*.

Таким образом, метод сбора зависит от способов доступа, поддерживаемых источником.

Первый способ – использование фреймворка *Scrapy*. Данный способ применяется для реестров компаний, в которых данные представлены в формате *HTML*. Для того чтобы собрать необходимую информацию из данного типа источника, требуется программа, которая способна перебирать веб-страницы и извлекать из них данные. Для такой задачи наиболее подходит фреймворк *Scrapy*. С помощью пользовательских классов, разработанных на языке программирования *Python*, осуществляется управление ядром *Scrapy*. Данные классы необходимы для структурного анализа и извлечения элементов из веб-страниц.

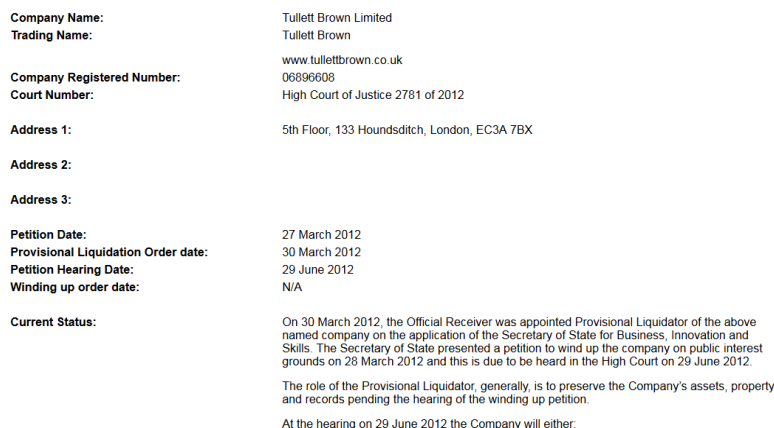
На текущем этапе рассматривается пример работы программы сбора данных для реестра компаний *The Insolvency Service* [3]. Веб-страницы данного реестра представлены на рисунке 1 и рисунке 2. Рисунок 1 отображает начальную страницу данного веб-сайта. На начальной странице *Insolvency Service* содержится список компаний в данном реестре компаний. Рисунок 2 отображает веб-страницу компании *Tullett Brown Limited*. Все остальные компании в данном реестре имеют такую же структуру предоставляемой о них информации, как показано на рисунке 2.



The screenshot shows the logo of The Insolvency Service and a section titled "Current company winding up or provisional liquidations". Below this, there are three entries, each with a table of details:

Find out more about provisional liquidation or view the results differently	
Company Name:	Tullett Brown Limited
Further Details:	Tullett Brown Limited
Court Number:	High Court of Justice 2781 of 2012
Company Registered Number:	06896608
Company Name:	MR Investment Club Limited
Further Details:	MR Investment Club Limited
Court Number:	High Court of Justice 1872 of 2012
Company Registered Number:	07747729
Company Name:	Manor Rose Limited
Further Details:	Manor Rose Limited
Court Number:	High Court of Justice 1869 of 2012
Company Registered Number:	06913882

Рис. 1. Начальная веб-страница реестра *The Insolvency Service*



The screenshot shows the details for Tullett Brown Limited, including company name, trading name, website, registered number, court number, address, petition date, and current status.

Company Name:	Tullett Brown Limited
Trading Name:	Tullett Brown
	www.tullettbrown.co.uk
Company Registered Number:	06896608
Court Number:	High Court of Justice 2781 of 2012
Address 1:	5th Floor, 133 Houndsditch, London, EC3A 7BX
Address 2:	
Address 3:	
Petition Date:	27 March 2012
Provisional Liquidation Order date:	30 March 2012
Petition Hearing Date:	29 June 2012
Winding up order date:	N/A
Current Status:	On 30 March 2012, the Official Receiver was appointed Provisional Liquidator of the above named company on the application of the Secretary of State for Business, Innovation and Skills. The Secretary of State presented a petition to wind up the company on public interest grounds on 28 March 2012 and this is due to be heard in the High Court on 29 June 2012. The role of the Provisional Liquidator, generally, is to preserve the Company's assets, property and records pending the hearing of the winding up petition. At the hearing on 29 June 2012 the Company will either:

Рис. 2. Веб-страница компании *Tullett Brown Limited*

Таким образом, появляется необходимость в написании пользовательского класса, с помощью которого программа по сбору будет иметь возможность обхода всех содержащихся компаний и извлечения всей предоставляемой информации для каждой из них.

Алгоритм программы сбора данных для реестра *Insolvency Service* представлен на рисунке 3.

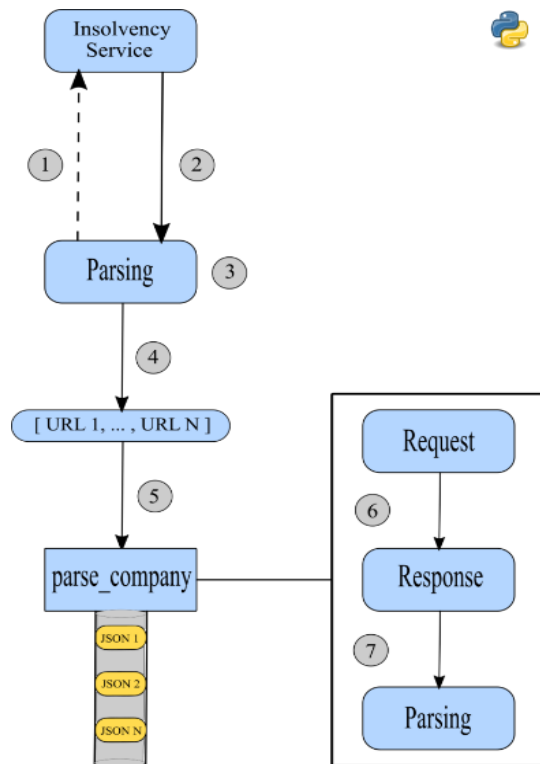


Рис. 3. Алгоритм программы по сбору данных для реестров типа HTML

Процесс работы пользовательского класса выглядит следующим образом:

1. *HTTP* запрос по заданному *URL*-адресу реестра *The Insolvency Service*;
2. загрузка данной веб-страницы;
3. обработка начальной страницы.
4. результат шага номер 3: список из *URL*-адресов страниц компаний;
5. для каждого элемента вызывается функция *parse_company*;
6. для текущего элемента происходит загрузка веб-страницы;
7. осуществляется синтаксический анализ текущей страницы. Результат данного шага: данные о текущей компании;
8. результат работы пользовательского класса – поток *JSON* файлов с данными о каждой компании в отдельности.

За счет того, что *Scrapy* построен на парадигме асинхронного программирования, процесс синтаксического анализа веб-страниц существенно быстрее, если сравнивать с синхронным подходом.

Для каждого реестра компаний данного типа создается собственный пользовательский класс, так как веб-сайты имеют свою специфику описания разметки на языке *HTML*, отличную друг от друга.

Второй способ – разработка специализированного скрипта на языке программирования *Python*. Для того чтобы получить данные о компаниях, в случае когда информация в реестре представлена в виде файла, необходимо произвести загрузку данного файла. Целью разработанного скрипта является получение *URL*-адреса для скачивания файла с данными о компаниях в реестре. Для этого используются широко распространенные библиотеки в языке программирования *Python*: *Requests* и *BeautifulSoup*.

На текущем этапе рассматривается пример работы *Python* скрипта для получения *URL*-адреса загрузки файла с данными о компаниях из реестра *Companies House* [4]. Веб-страница данного реестра отображена на рисунке 4. Из рисунка видно, что в реестре *Companies House* представлена возможность скачать данные о компаниях одним файлом *BasicCompanyDataAsOneFile-2019-03-01.zip*.

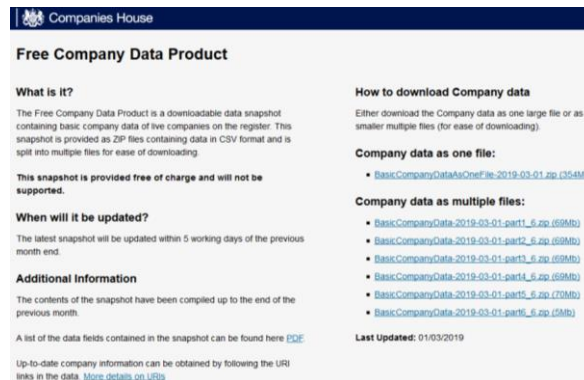


Рис. 4. Веб-страница реестра *Companies House*

Таким образом, необходимо разработать *Python* скрипт, с помощью которого будет получен *URL*-адрес для скачивания файла с данными о компаниях из реестра *Companies House*. Алгоритм работы данного скрипта представлен на рисунке 5.

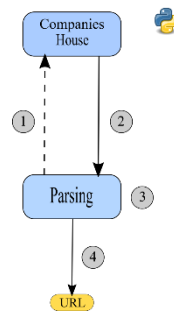


Рис. 5. Алгоритм скрипта для получения *URL*-адреса

Процесс работы скрипта выглядит следующим образом:

1. *HTTP* запрос по заданному *URL*-адресу реестра *Companies House*;
2. загрузка данной веб-страницы;
3. синтаксический анализ текущей веб-страницы.
4. результат шага номер 3 – *URL*-адрес для скачивания файла с данными о компаниях.

Для каждого реестра компаний данного типа создается собственный *Python* скрипт, так как веб-сайты имеют свою специфику описания разметки на языке *HTML*, отличную друг от друга.

Таким образом, организован сбор данных. Для случая, когда данные в реестре представлены в формате *HTML*-страниц, разработаны пользовательские классы, с помощью которых осуществляется управление ядром в фреймворке *Scrapy*. В результате, с помощью *Scrapy* происходит извлечение данных о компаниях. Для случая, когда данные в реестре представлены в виде файла, разработаны *Python* скрипты. С помощью данных скриптов осуществляется извлечение *URL*-адресов для загрузки файлов с данными о компаниях.

Данный этап является точкой входа в подсистему потоковой обработки информационной системы сбора данных. Полученные результаты в процессе сбора впоследствии используются следующими компонентами системы в процессе потока данных.

2. Организация потока и хранения данных о компаниях-нерезидентах

В результате сбора данных необходимо обеспечить прием и интеграцию потока данных из различных источников в хранилище данных. Организация потока данных осуществляется посредством использования современных технологий больших данных: *Apache Kafka* и *Apache NiFi*. Данные технологии предлагают масштабируемую и отказоустойчивую структуру приема и интеграции данных. Также как и в случае со сбором данных, задача организации потока данных делится на два случая в зависимости от типа доступа к данным в реестре компаний.

Результатом сбора данных из реестров компаний, в которых информация представлена в виде файла, являются *URL*-адреса для загрузки данных файлов. Таким образом, для данного случая необходимо осуществить загрузку файлов в хранилище данных – *HDFS*. Для того чтобы информационная система была масштабируема и справлялась с высокими нагрузками, результаты работы *Python* скриптов направляются в очередь сообщений *Kafka*. С практической точки зрения это будет выглядеть так, как представлено на рисунке 6.

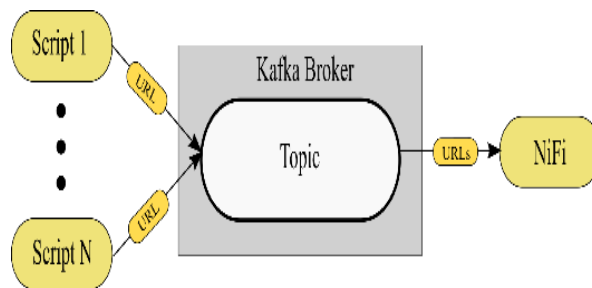


Рис. 6. Обмен сообщениями между скриптами и NiFi

Множество *Python* скриптов выступают в качестве издателей для брокера *Kafka*. На брокере *Kafka* создается топик. Далее издатели опубликовывают свои сообщения (*URL*-адреса) в заданном топике. В роли потребителя выступает платформа автоматизации потока данных между системами – *Apache NiFi*. После того как *NiFi* подписался на существующий топик, сообщения становятся доступными для дальнейшей обработки.

На рисунке 7 представлена схема работы *Apache NiFi*.

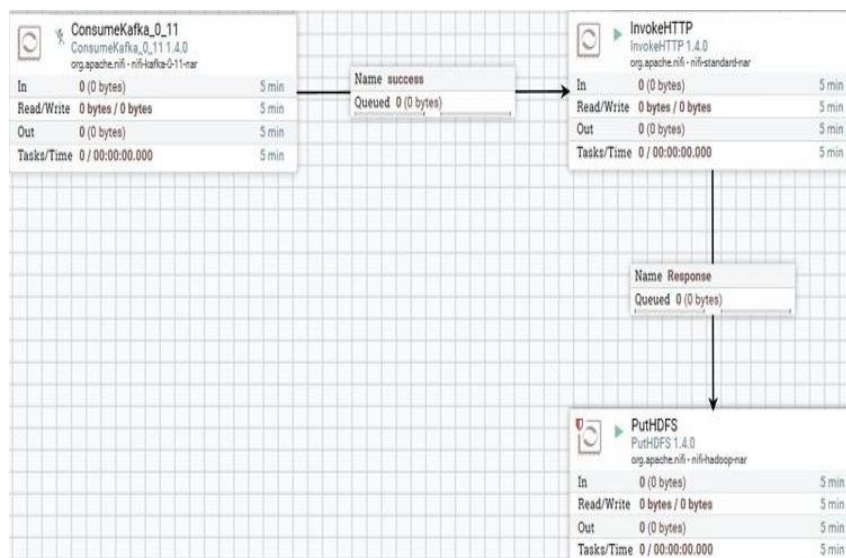


Рис. 7. Скачивание файла и сохранение его в HDFS с помощью NiFi

На начальном этапе (блок *ConsumeKafka*) осуществляется прием сообщений (*URL*-адресов) из *Kafka*. Далее (блок *InvokeHTTP*) по полученному *URL*-адресу производится загрузка файлов с данными о компаниях. Заключительным этапом (блок *PutHDFS*) является сохранение файла с информацией о компаниях в хранилище данных – *HDFS*.

Таким образом, данные о компаниях скачиваются по *URL*-адресу полученному на этапе сбора, после чего сохраняются в хранилище данных посредством организации потока данных с помощью *Apache Kafka* и *NiFi*.

Второй случай организации потока данных предназначен для национальных реестров, в которых информация о компаниях представлена в формате *HTML*. Результатом сбора в данном случае для одного реестра является поток файлов с данными о компаниях в формате *JSON*. Для того чтобы система была масштабируема и справлялась с высокими нагрузками, результаты работы пользовательских классов направляются в очередь сообщений *Kafka*. Схематически данный поток представлен на рисунке 8.

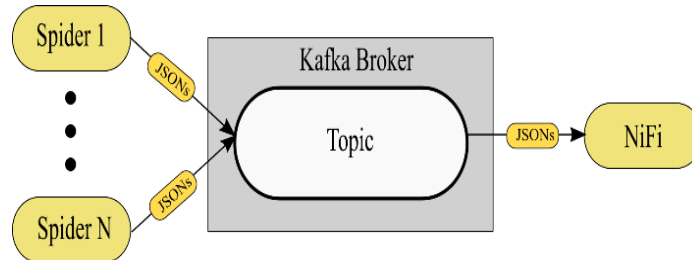


Рис. 8. Обмен сообщениями между пользовательскими классами *Scrapy* и *NiFi*

Множество пользовательских классов (англ. *Spider*) выступают в качестве издателей для брокера *Kafka*. На брокере *Kafka* создается топик. Далее издатели направляют поток файлов с данными о компаниях в заданный топик. В роли потребителя выступает платформа автоматизации потока данных между системами – *Apache NiFi*. После того как *NiFi* подписался на существующий топик, сообщения стали доступны для дальнейшей обработки.

На рисунке 9 представлен процесс работы *Apache NiFi*.



Рис. 9. Объединение файлов и их сохранение в *HDFS* с помощью *NiFi*

На начальном этапе (блок *ConsumeKafka*) осуществляется прием сообщений (поток *JSON* файлов, в котором содержится запись об одной компании) из *Kafka*. Далее (блок *MergeContent*) производится объединение *JSON* файлов в один. Заключительным этапом (блок *PutHDFS*) является сохранение *JSON* файла с информацией о компаниях в хранилище данных – *HDFS*.

Так, данные о компаниях, полученные на этапе сбора, были объединены в один файл, после чего сохранены в хранилище данных посредством организации потока данных с помощью *Apache Kafka* и *NiFi*.

Поток данных организован таким образом, что в случае добавления новых программ по сбору, система продолжит стабильно функционировать за счет использования очереди сообщений *Apache Kafka* и платформы автоматизации потока данных *Apache NiFi*. В общем случае схема потоковой обработки данных представлена на рисунке 10.

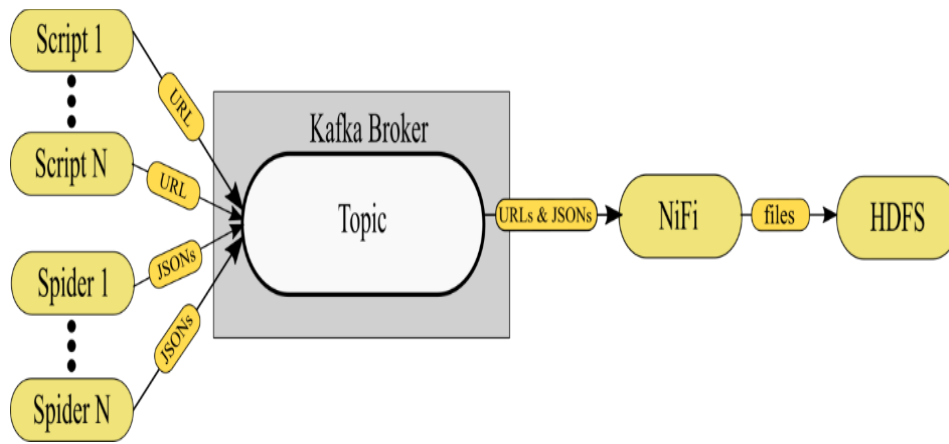


Рис. 10. Схема потоковой обработки данных

Заключение

В результате организации сбора и потоковой обработки данных в *HDFS* хранится набор файлов различных расширений (*CSV*, *XML*, *JSON* и т. п.). Каждому файлу соответствует национальный реестр компаний в сети Интернет. После того как данные сохранены в *HDFS*, они доступны для дальнейшей обработки.

Применение описанных в настоящей статье технологий и методов позволяет добиться стабильности функционирования системы сбора и хранения данных о компаниях-нерезидентах и упрощения ее поддержки. Использование таких продуктов, как *Apache Kafka*, *Apache NiFi* и *HDFS*, является промышленным стандартом для организации систем потоковой обработки больших данных.

Таким образом, в рамках разработки информационного ресурса контроллинга деятельности компаний-нерезидентов организованы сбор, потоковая обработка и хранение информации о компаниях-нерезидентах из национальных реестров.

Подходы по организации сбора и потоковой обработки данных, описанные в настоящей статье, применим не только к информации из реестров компаний, но и к аналогичным, схожим по своей структуре источникам данных, отличающимся по содержанию в зависимости от рассматриваемой проблематики. Используемые подходы могут быть полезны и на других этапах разработки информационного ресурса контроллинга деятельности компаний-нерезидентов.

Список литературы

1. World Bank. *Doing Business 2015: Going Beyond Efficiency*: World Bank Publications, 2014. — С. 47. — ISBN 978-1-4648-0352-9.
2. Kate Matsudaira. *Communications of the ACM*. — 2014. — Vol. 57. — No. 3. — Pp. 10-11. — DOI=10.1145/2567664.
3. The Insolvency Service // *Liquidation companies*. — [Электронный ресурс]. URL: <https://www.insolvencydirect.bis.gov.uk/piudb/viewqryl.asp>. — (дата обращения 01.07.2019).
4. Companies House // *Free Company Data Product*. — [Электронный ресурс] URL: http://download.companieshouse.gov.uk/en_output.html. — (дата обращения 01.07.2019).