S.V. Zapechnikov, S.Ya. Nagibin, M.Yu. Senatorov, E.V. Frolkov, A.V. Shmid

An Approach to the Evaluation and Prediction of the Outflow of Bank Deposits Dynamics Using the Big Data Technologies

Keywords: big data, unstructured data, outflow of deposits.

The results of implementation of pilot-project to test IBM Big Data platform for automation of processes of collection, processing and analysis of unstructured big data applied to the individual functional unit of the banking supervision of the Central Bank of Russia is presented.

С.В. Запечников, С.Я. Нагибин, М.Ю. Сенаторов, Е.В. Фролков, А.В. Шмид

ПОДХОД К ОЦЕНКЕ И ПРОГНОЗИРОВАНИЮ ДИНАМИКИ ОТТОКА БАНКОВСКИХ ВКЛАДОВ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ БОЛЬШИХ ДАННЫХ

Анализ тенденций развития информационных технологий, в том числе по оценкам ведущих исследовательских компаний в IT-индустрии (Gartner, InternationalData-Corporation и других), свидетельствует о неуклонном росте рынка технологий больших данных (BigData) и их востребованности для крупных компаний в сфере финансов, телекоммуникаций, розничной торговли и других для достижения конкурентных преимуществ. Одна из основных сфер применения технологий больших данных – сфера обеспечения информационной, финансовой и экономической безопасности. Одной из важных составляющих банковской безопасности и устойчивости финансового рынка в целом является возможность своевременного прогнозирования динамики изменения объема банковских вкладов населения, особенно предсказание ситуаций, связанных с резким оттоком вкладов.

Для решения обозначенной задачи необходим учет не только чисто экономических факторов с использованием традиционных методов эконометрики, но и анализ содержания большого массива информационных источников Интернета, поскольку на отток вкладов серьезное вличние могут оказать такие внеэкономические факторы, как новостная информация в СМИ, распространение слухов, панических настроений, преднамеренный вброс дезинформации и прочие факторы. Для этого, как показывает практика, лучше всего подходят методы интеллектуального анализа данных, уже апробированные в сфере информационной безопасности при решении задач фильтрации пакетного трафика, анализа электронной почты, обнаружения корреляции событий информационной безопасности, предотвращения утечек конфиденциальной информации и пр.

Ежегодно объемы хранимой информации вырастают на 40 %. Согласно прогнозам IDC, сегмент BigData будет расти примерно в семь раз быстрее рынка информационных и телекоммуникационных технологий в целом, и к 2015 году его мировой объем достигнет 16,9 млрд дол. Таких же прогнозов придерживается и Gartner, считающая, что BigData – одно из самых ключевых технологических направлений IT-отрасли.

В настоящее время ведущими производителями специализированных программно-аппаратных систем (IBM, Oracle, HP, Teradata, EMCGreenplum и другими) предлагаются решения для сбора, обработки, анализа и хранения данных класса BigData (далее в тесте используется понятие «технологии BigData»). При этом декларируемые возможности программно-аппаратных решений позволяют существенно повысить эффективность работы пользователей с большими объемами как структурированных, так и неструктурированных данных.

Данные, которыми оперируют предприятия и организации, в том числе из банковского сектора, всё больше становятся неструктурированными, а их объемы быстро и резко возрастают. Средств на основе традиционных систем бизнес-аналитики (Businessintelligence – BI) уже недостаточно для полноценного анализа [1]. Прогноз роста объема «цифрового мира» к 2016 году составит порядка семи зеттабайт. При этом существенный вклад в этот рост вносят новые источники неструктурированной информации, такие как социальные сети, мобильные устройства, медиасреда и т.п., на долю которых приходится свыше 80 % всех хранимых данных [1]. По разным оценкам, влияние социальных сетей на общество растет существенно быстрее, чем у традиционных средств массовой информации, что побуждает многие компании учитывать этот аспект при принятии бизнес-решений.

За рубежом технологии BigData наиболее востребованы в финансовом секторе (в частности, среди инвестиционных банков, которые работают с большими потоками информации с торговых площадок в режиме реального времени). По данным опроса CNews, российские вендоры пока не торопятся инвестировать большие средства в разработки в области BigData из-за крайне низкого спроса, но положительно оценивают перспективность таких решений в целом [2].

Возрастающая потребность функциональных подразделений надзорного блока Банка России в обработке больших объемов неструктурированной информации вкупе с необходимостью постоянного совершенствования мер оперативного предотвращения ситуаций быстрого ухудшения финансового положения кредитных организаций, особенно банков – участников системы страхования вкладов, обусловило целесообразность проведения исследований по решению *задачи сбора, обработки и анализа неструктурированной информации из внешних источников по банковским депозитным продуктам для физических лиц в целях оценки и прогнозирования динамики оттока вкладов*. Под внешними источниками неструктурированной информации понимаются: новостные сайты, сайты с форумами и отзывами пользователей по предоставляемым банковским услугам, социальные сети.

Исследования проводились с использованием платформы компании IBM на стенде одного из ведущих российских системных интеграторов, ЗАО «ЕС-лизинг», в структуре которого успешно функционирует Центр компетенции по технологиям IBMBigData.

На рис. 1 приведена общая схема платформы с основными функциональными компонентами [3]:

- IBMInfoSphereBigInsights – компонент, основанный на программном обеспечении с открытым кодом ApacheHadoop, для хранения и обработки квазиструктурированной и неструктурированной информации;
- IBMInfoSphereStreams – компонент для обработки потоковых данных;
- IBMPureData (Netezza) – программно-аппаратный комплекс, предназначенный для быстрого и глубокого анализа больших объемов структурированных данных;
- IBMInfoSphereDataExplorer – инструментальное средство корпоративного поиска, в котором, кроме традиционного поиска по ключевым словам, доступны средства динамической категоризации, визуализации и персонализации результатов.

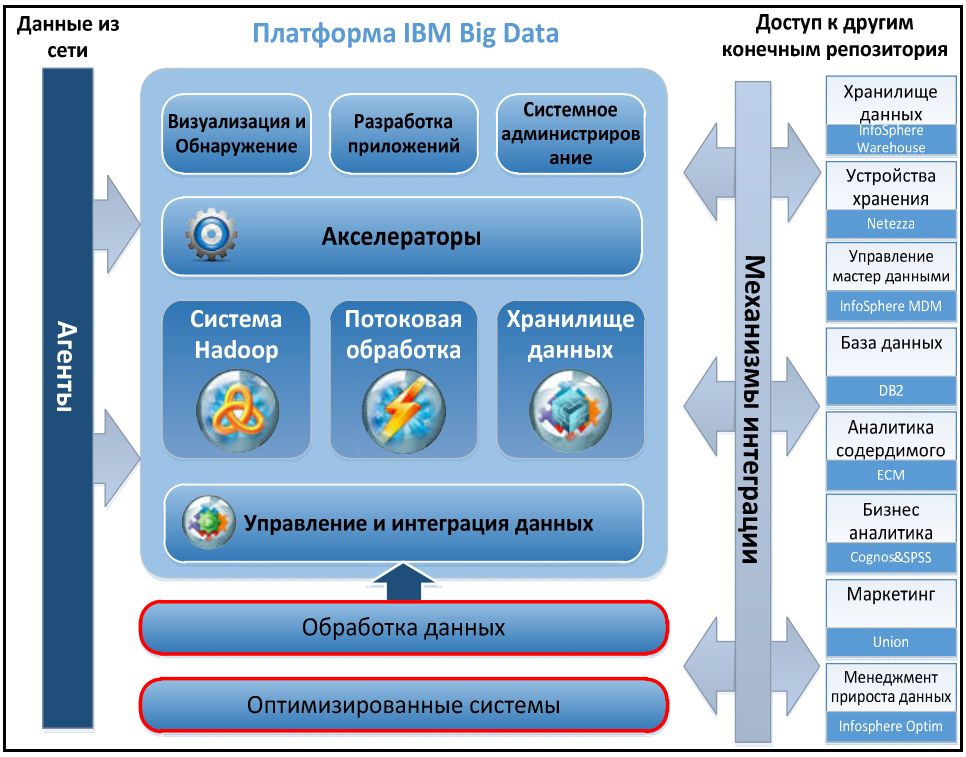

Рис. 1. Общая схема платформы IBM BigData

Помимо вышеперечисленных компонент, платформа содержит еще так называемые акселераторы, которые представляют собой готовые механизмы извлечения и обработки данных:

- машинных данных (логов, системных журналов серверов и т.д.);
- социальных данных с проприетарными сервисами на основе платной подписки;
- телекоммуникационных данных для анализа записей разговоров.

На рис. 2 приведен вариант архитектуры созданного опытного участка для исследования задачи по сбору, обработке и анализу данных внешнего контента в целях оценки и прогнозирования динамики оттока вкладов в коммерческих банках и кредитных организациях.

Несмотря на заявленные возможности платформы IBM BigData по сбору данных внешнего контента с использованием встроенных коннекторов, в рамках пилотного проекта был использован асинхронный способ получения данных с помощью созданных краулеров на основе технологии Node.js. Функционал Node.js позволяет в краткие сроки адаптировать сканеры неструктурированной информации и не относится к классу проприетарных решений. Он позволяет достаточно быстро проводить структуризацию информации по различным признакам, выявляемым в ходе анализа поступающей информации, а также провести предобработку информации для обрезки паразитных частей получаемых страниц, и обфускации текстовой информации на страницах, не подвергаемых аналитической обработке.

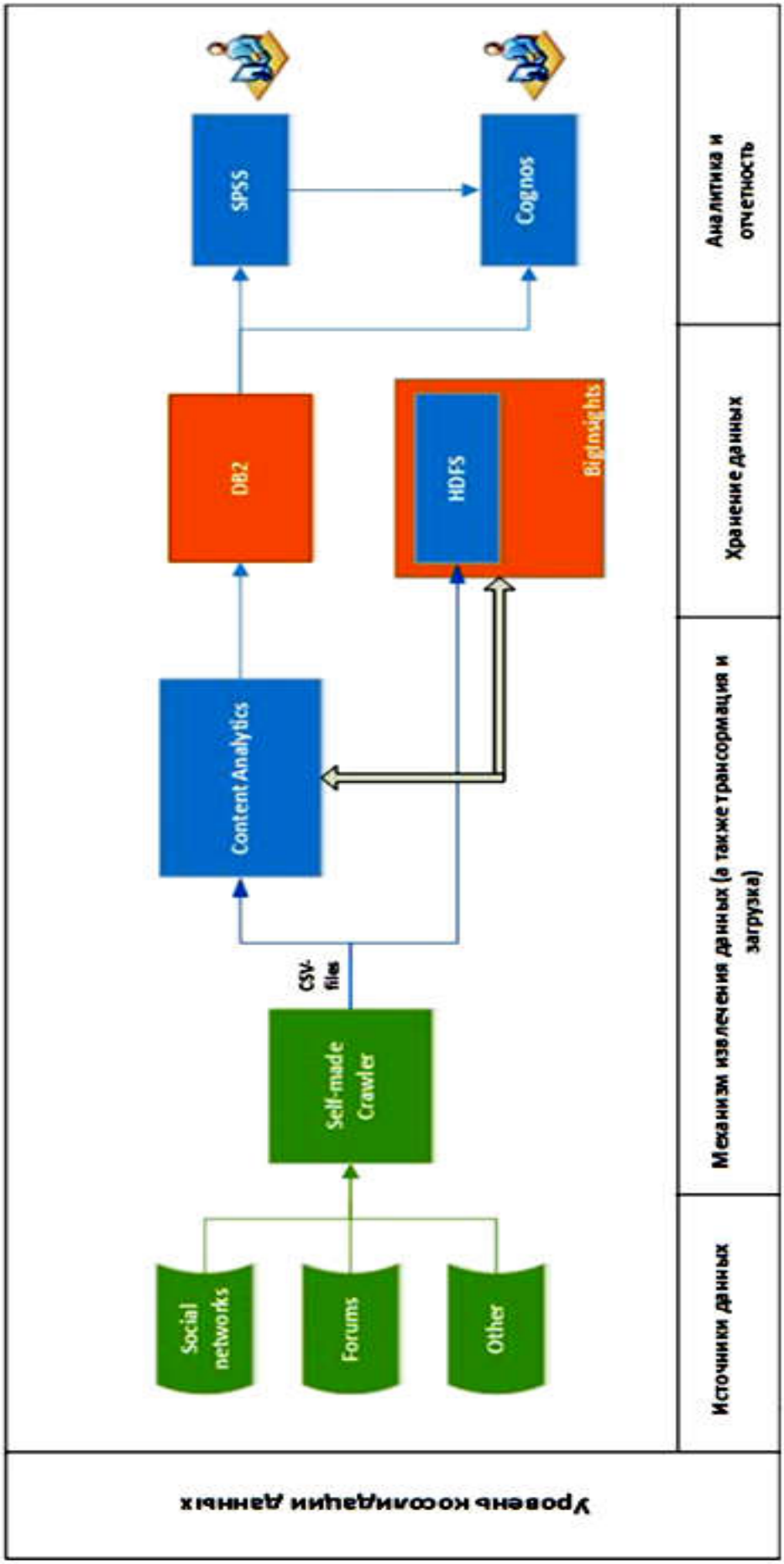

Рис. 2. Вариант архитектуры опытного участка в рамках пилотного проекта

В табл. 1 обобщены данные по количеству собранных сообщений из внешних источников неструктурированных данных в ходе решения задачи пилотного проекта.

Таблица 1. Перечень источников неструктурированных данных

Источник	Количество сообщений
Сообщения форума Bankir.ru	2 109 309
Сообщения форума Banki.ru	1 137 753
Отзывы с Banki.ru	135 550
Сообщения социальной сети Facebook	133 871
Сообщения социальной сети ВКонтакте	111 081
Сообщения социальной сети Twitter	72 997
Всего сообщений	3 700 561

Статистика распределения обработанных сообщений по годам применительно к депозитным банковским продуктам представлена на рис. 3 и характеризует практически экспоненциальный рост активности клиентов кредитных организаций в Интернете.

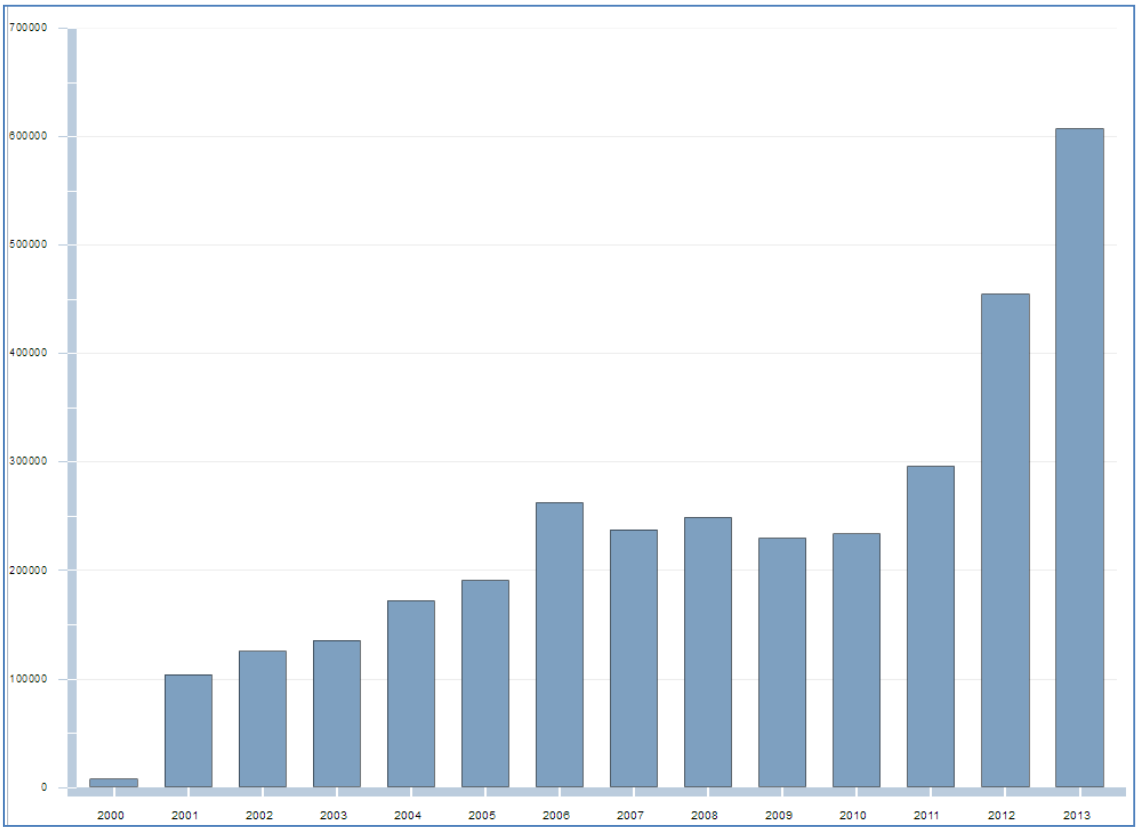

Рис. 3. Статистика обработанных сообщений по банковской тематике

При наличии в линейке продуктов IBM BigData специализированной поисковой системы DataExplorer решение задачи категоризации, кластеризации и визуализации результатов анализа внешнего контента на основе заданных критериев и созданных фасетов осуществлялось с использованием программного продукта IBM Content Analytics with Enterprise Search, обеспечившего лучшие возможности по просмотру, синтаксическому разбору и анализу содержимого с созданием пригодного для поиска индекса. Функционал данного программного продукта позволяет также импортировать полученную структурированную статистическую информацию в базу данных для проведе-

ния предиктивного анализа с использованием специализированных программных средств.

Следует также отметить, что на данном этапе полученные сообщения не разделялись на позитивные и негативные, поскольку это требует применения специальных технологий извлечения полезной информации из текста и правильной ее интерпретации (TextMining), которые реализуются методами компьютерной лингвистики. Обработка всего многообразия форматов социальных данных сопряжена с рядом трудностей, что делает практически нереализуемым комплексный автоматизированный анализ данных социальных СМИ. Например, сообщения в Twitter очень короткие, что затрудняет их контекстный анализ. К тому же социальные СМИ обычно развиваются динамически, порождая новые языковые формы (сленг, аббревиатуры и т.д.), которые трудно обрабатывать с помощью существующих методов анализа текста [4].

Учитывая отсутствие строгих регулятивных критериев понятия оттока вкладов физических лиц, в рамках пилотного проекта в качестве прогнозируемого события было принято изменение объёма депозитов банка ниже среднерыночного значения. Прогнозирование данного события реализовано с использованием известных статистических и эконометрических моделей и алгоритмов в составе программных продуктов IBM SPSS Statistics и IBM SPSS Modeler.

Для учёта активности пользователей социальных сетей и интернет-форумов в число объясняющих переменных было включено количество сообщений, посвящённых депозитным банковским продуктам, за предшествующие 12 месяцев с заданным ограничением на выборку не менее 40 сообщений.

Из внутренних источников структурированной информации в качестве объясняющих переменных использованы:

- данные оборотной ведомости по счетам бухгалтерского учета кредитных организаций применительно к исследуемым банковским продуктам – депозитам физических лиц;
- отдельные макроэкономические показатели в региональном разрезе: уровень безработицы и средний доход на душу населения;
- процентные ставки по банковским депозитам.

Для более полного представления о точности прогнозирования полученных моделей проведено исследование с оценкой по неполным данным. Вся выборка, используемая для оценивания, была разбита на две подвыборки: первая составляла 80 % всей выборки и использовалась для оценивания моделей; оставшиеся 20 % использовались только для подсчёта долей правильно идентифицированных наблюдений. Результаты оценки точности прогнозирования, представленные в табл. 2, позволяют сделать вывод об удовлетворительной погрешности прогнозирования события, связанного с изменением объёма депозитов банка ниже среднерыночного значения.

В табл. 3 представлен фрагмент варианта реализации рейтинга банков на 1 ноября 2013 г. по вероятности изменения объема депозитов ниже среднерыночного, рассчитанный с использованием logit-модели IBM SPSS Statistics по вероятности наступления события A_{it} – изменения объема депозитов ниже среднерыночного:

$$A_{it} = \{\Delta_{it} < \overline{\Delta_t}\},$$

где Δ_{it} – отношение значения объёма депозитов физических лиц у банка i на конец периода t к значению на начало периода t, $\overline{\Delta_t}$ – отношение значения суммарного объёма депозитов всех банков на конец периода t к значению на начало периода t.

Таблица 2. Результаты оценки точности прогнозирования события с использованием моделей и алгоритмов IBMSPSS

Наименование модели (алгоритма)	Точность прогнозирования модели (алгоритма), %
Байесовская сеть	69
Линейная дискриминантная модель ЛДМ1	66
Линейная дискриминантная модель ЛДМ2	66
Нейронная сеть	70
Logit	68
CHAID	72
C5.0	81
Quest	69

Таблица 3. Фрагмент варианта реализации рейтинга банков на 1 ноября 2013 г. по вероятности изменения объема депозитов ниже среднерыночного

	Наименование банка	Вероятность изменения объема депозита ниже среднерыночного
1	КБ Ренессанс Капитал (ООО)	0,77692549
2	ЗАО Смартбанк	0,76096783
3	ОАО Московский кредитный банк	0,70463116
4	ЗАО Банк Русский Стандарт	0,68428136
5	ООО ХКФ Банк	0,66999822
6	ЗАО Райффайзенбанк	0,6560539
7	ОАО Уралсиб	0,6529727
8	ЗАО ЮниКредит Банк	0,62443364
9	ОАО Сбербанк России	0,61402411
10	ОАО Банк Москвы	0,6109419
11	Связной Банк (ЗАО)	0,60476635
12	ВТБ 24 (ЗАО)	0,60156455
13	НБ Траст (ОАО)	0,59762466
14	ОАО Промсвязьбанк	0,59312503
15	ОАО КБ Восточный	0,58818823

Для оценивания вероятности события A_{it} вводится индикатор этого события y_{it}: $y_{it} = 1$, если банк i не столкнулся с оттоком вкладов в период t; $y_{it} = 0$ – в противном случае. Согласно logit-модели [5]:

$$(y_{it} = 1) = \frac{\exp(x'_{it}\beta)}{1 + \exp(x'_{it}\beta)},$$

где x_{it} – значения некоторых объясняющих переменных для банка i в период t; β – неизвестные параметры, которые подлежат определению. Оценки $\hat{\beta}$ параметров β находятся методом частного максимального правдоподобия (partialmaximumlikelihood), стандартные ошибки вычислены в предположении зависимости наблюдений для одного банка и независимости наблюдений для разных банков [5].

Таким образом, в рамках пилотного проекта создан существенный практический задел по использованию компонентов платформы IBM BigData, а также других смежных программных средств IBM для решения задачи сбора, обработки и анализа данных

из внешних источников неструктурированной информации в интересах подразделений надзорного блока Банка России.

Учитывая экспоненциальный рост данных внешнего контента, в том числе по банковской тематике, дальнейшие исследования и внедрение технологий больших данных применительно к решению задач Банка России предполагается осуществлять по следующим основным направлениям:

- совершенствование модели прогнозирования оттока вкладов физических лиц за счет привлечения данных из форм отчетности, формируемых с периодичностью 1–5 дней;
- выявление признаков искажения представляемых кредитными организациями форм отчетности по косвенным данным – информации, порождаемой клиентами банков в сети Интернет (социальных сетях, форумах и т.п.). Предполагается, что объем порождаемого контента по банковской тематике (вклады, кредиты и т.п.) может быть связан с аккумулированием средств на соответствующих счетах корреспондентского счета кредитной организации, а также другими показателями банковской отчетности и макроэкономики (данные Росстата);
- определение по данным из внешних источников неструктурированной информации различных взаимосвязей субъектов, оказывающих существенное влияние на решения, принимаемые органами управления банка, в целях выявления их деятельности в интересах третьих лиц, установления банковских групп и т.п.

СПИСОК ЛИТЕРАТУРЫ:

1. Черняк Л. Аналитика неструктурированных данных // Открытые системы. 2012. № 6. С. 30 – 34.
2. BigData: возможность или необходимость/ Аналитический бюллетень // CNewsAnalytics. 2013. 11 с.
3. Paul C. Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, David Corrigan, James Giles Harness the Power of Big Data. The IBM Big Data Platform. – The McGraw-Hill Companies, 2013. 281 p. ISBN: 978-0-07180818-7.
4. Шрек Т., Кейм Д. Визуальный анализ данных из СМИ // Открытые системы. 2013. № 6. С. 18 – 22.
5. Wooldridge, J. M. Econometric Analysis of Cross Section and Panel Data.– Cambridge, Massachusetts. MIT Press. 2010. 1096 pp.

REFERENCES:

1. Chernyak L. Analitikanestrukturirovannyhdannyh // Otkrytyesystemy. 2012. No. 6. Pp. 30 – 34. (In Russian)
2. Big Data: vozmozhnostilineobhodimost / Analiticheskiybulleten // CNews Analytics. 2013. 11 pp. (in Russian)
3. Paul C. Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, David Corrigan, James Giles Harness the Power of Big Data. The IBM Big Data Platform. – The McGraw-Hill Companies, 2013. – 281 p. – ISBN: 978-0-07180818-7.
4. Shrek T.,Keim D. Visualniyanalyzdannyhiz SMI // Otkrytyesystemy. 2013. No. 6. Pp. 18 – 22.
5. Wooldridge, J. M. Econometric Analysis of Cross Section and Panel Data.– Cambridge, Massachusetts: MIT Press, 2010. 1096 pp.