УДК: 330.47, 004.67, 004.75

БОЛЬШИЕ ДАННЫЕ И ЦИФРОВАЯ ЭКОНОМИКА

Белов Сергей Дмитриевич[1], Зрелова Дарья Петровна[2]
Кореньков Владимир Васильевич[3]

[1]Ведущий программист;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
Российский экономический университет имени Г.В. Плеханова;
117997, РФ, г.Москва, Стремянный пер., 36;
e-mail:belov@jinr.ru.

[2]Инженер-программист;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
e-mail: zrelova@me.com.

[3]Директор, 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;
Российский экономический университет имени Г.В. Плеханова;
117997, РФ, г.Москва, Стремянный пер., 36;
e-mail: korenkov@cv.jinr.ru.

В данной работе Большие данные рассматриваются как «зонтичный» термин, объединяющий различные понятия, технологии и методы обработки данных в распределенных информационных системах, обеспечивающие получение качественно новой полезной информации (новых знаний). Описываются этапы «стандартного» исследования в подходе Больших данных. Дается краткое описание экосистемы Больших данных, состоящей из нескольких основных категорий. Рассматриваются различные проекты и инициативы национального и межгосударственного уровня, примеры применения Больших данных в бизнесе, экономике, обществе. В качестве конкретных примеров построения и использования аналитических платформ Больших данных приведены успешные социально-экономические исследования, выполненные авторами в составе научных коллективов на базе Российского экономического университета им. Г.В. Плеханова.

Метафора Больших данных удачна, поскольку естественным образом связывает комплекс понятий, технологий и методов Больших данных с экономикой посредством намека на связь с другими известными метафорами – «Большой нефтью», «Большой рудой» и т.д.

Ключевые слова: Большие данные, цифровая экономика, информационные системы.

BIG DATA AND DIGITAL ECONOMY

Belov Sergey[1], Zrelova Daria[2], Korenkov Vladimir[3]

[1]Lead programmer;
Joint institute for nuclear research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, Joliot-Curie, 6;
Plekhanov Russian University of Economics;
36 Stremyanny per., Moscow, 117997, Russia;
e-mail: belov@.jinr.ru.

[2]*Programmer engineer;*
Joint institute for nuclear research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, Joliot-Curie, 6;
e-mail: zrelova@me.com.

[3]*Director, Doctor of Technical Science, professor;*
Joint institute for nuclear research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, Joliot-Curie, 6;
Dubna State University;
Institute of the system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
Plekhanov Russian University of Economics;
36 Stremyanny per., Moscow, 117997, Russia;
e-mail: korenkov@cv.jinr.ru.

In this paper, Big Data is considered as an "umbrella" term that combines various concepts, technologies and methods of data processing in distributed information systems that provide a qualitatively new useful information (new knowledge). The stages of "standard" research in the Big Data approach are described. A brief description of the Big Data ecosystem, which consists of several main categories, is given. Various projects and initiatives at the national and international levels are considered, as well as examples of the use of Big Data in business, economy, and society. As concrete examples of the construction and use of analytical platforms for Big Data, successful socio-economic research carried out by the authors as part of research teams at the Plekhanov Russian University of Economics is presented.

The Big data metaphor is certainly successful, since it naturally connects a complex of concepts, technologies and methods of Big data with the economy by hinting at a connection with other well-known metaphors – "Big oil", "Big ore", etc.

Keywords: Big data, digital economy, information systems.

Введение

Большие данные стали термином, повсеместно используемым во многих отраслях промышленности, государственного управления, в научных кругах. Существует множество определений Больших данных, используемых в разных контекстах и с разными целями, что как в притче о слепых мудрецах и слоне, способно привести к однобокому, лишенному структурной целостности, пониманию. Поэтому до принятия «стандартного определения» с целью минимизации возможной путаницы, разумно использовать так называемое консенсусное определение – «Большие данные представляют собой информационные активы, характеризующиеся таким большим объемом, скоростью и разнообразием, что для их преобразования в ценность требуются специальные технологии и аналитические методы» [1].

Данные, прежде всего, Большие данные – это новый актив, тренд и ключевой фактор цифровой экономики (современное направление экономической теории), применяемый в новых целях и для реализации новых идей. Большинство технологий цифровой экономики развивается в рамках методологической основы Больших данных. Технологии больших данных обеспечивают возможность анализа самых разных источников данных, быстро обрабатывать и анализировать информацию и, тем самым, обеспечивать оперативное принятие хорошо обоснованных решений. Технологии больших данных обеспечивают более эффективные способы ведения бизнеса.

В связи с тем, что Большие данные помогают снижать затраты, лучше и быстрее принимать решения, создавать новые продукты и услуги технологии Больших данных востребованы в целом ряда отраслей и видов деятельности. Умение оперировать большими объемами информации, анализировать взаимосвязи между ними и принимать взвешенные решения несет потенциал для увеличения показателей доходности и прибыльности, повышения эффективности. В финансовых организациях – это анализ кредитоспособности, управление активами и рисками, предотвращение мошенничества. В энергетике – управление режимами энергосистемы, мониторинг состояния оборудования, прогноз потребления энергии. В торговле – продвижение товаров, управление товарными запасами, прогнозирование спроса, логистика.

Большие данные востребованы в государственном и муниципальном управлении, медицине, здравоохранении, науке и образовании. «Умный город», «Интернет вещей» и «Промышленный интернет» – это тоже Большие данные.

Многие страны приступили к осуществлению планов и инициатив в области исследований и прикладных программ в сфере Больших данных. В 2012 года администрация президента США официально запустила инициативу по исследованию и развитию Больших данных. Помимо США, Великобритания, Франция, Австралия и Япония также представили свои инициативы в области Больших данных. Европейская комиссия в своей рамочной программе исследований и инноваций Horizon 2020 инвестировала в промышленные исследования и приложения, связанные с Большими данными.

В России в 2017 году утверждена правительственная программа «Цифровая экономика Российской Федерации», Большие данные входят в список 9 основных цифровых технологий, указанных в программе.

Существуют не только национальные, но и наднациональные, международные программы. Одной из самых значимых является инициатива ООН в области Больших данных – Global Pulse.

Самый яркий пример последнего времени, иллюстрирующий возможности «работы» Больших данных – это многочисленные и разнообразные исследования в области COVID-19, в том числе – создание аналитических систем уровня «near real-time» с интерактивным доступом к самой разной информации о пандемии и предоставлением средства для анализа и прогноза посредством веб-интерфейса... для всех желающих.

1. Краткий обзор Больших данных

Большие данные (Big Data) часто определяются как данные, которые по объему, скорости и разнообразию превышают возможности традиционных систем хранения и обработки данных. При решении этой проблемы возникает необходимость разработки новых методов сбора, обработки, анализа данных; необходимость в обновленном инструментарии поиска, совместного использования, хранения, передачи, визуализации данных; выработки новых алгоритмов запросов, обновления и конфиденциальности информации. Технологии Больших данных делают возможным обработку и анализ огромного количества данных, в некоторых случаях – всех данных, касающиеся того или иного явления в их первозданном виде – структурированном, неструктурированном, потоковом и позволяют раскрыть потенциал больших массивов данных за счет выявления скрытых закономерностей и фактов.

Основные вехи. 1997 - ученые NASA Michael Cox и David Ellsworth опубликовали статью «The problem of Big Data» [2]. 2001 - Doug Laney опубликовал статью «3D data management: controlling data Volume, Velocity & Variety» [3]. 2004 – Jeffrey Dean и Sanjay Ghemawat опубликовали статью «Simplified Data Processing on Large Cluster» (MapReduce) [4].

Считается, что широкие обсуждения начались в 2005 году в изданиях компании O'Reilly media в связи с необходимостью хоть как-то определить те данные, с которыми традиционные технологии управления и обработки данных не справлялись в силу их сложности и большого объема.

В 2008 году журнал Nature в специальном номере, посвященном теме Big Data, опубликовал статью Клиффорда Линча «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объемами данных?» [5]. В номере были собраны материалы о феномене взрывного роста объемов и многообразия обрабатываемых данных. Там же обсуждались технологические перспективы в парадигме вероятного скачка «от количества к качеству».

В 2009 году термин широко распространился в деловой прессе, а к 2010 году относят появление первых продуктов и решений. К 2011 году большинство крупнейших поставщиков информационных технологий уже используют понятие Больших данных, в том числе IBM, Oracle, Microsoft, Hewlett-Packard, EMC.

В 2011 году компания Gartner отметила Большие данные как тренд номер два в информационно-технологической инфраструктуре (после виртуализации) и дала прогноз, что внедрение технологий Больших данных окажет значительное влияние на подходы в области информационных технологий в производстве, здравоохранении, торговле и государственном управлении.

В том же 2011 году компания McKinsey опубликовала доклад «Big Data: The next frontier for innovation, competition, and productivity» [6].

С 2013 года Большие данные как академический предмет появились в вузовских программах науки о данных (Data Science) и др.

Рост производства данных, наблюдающийся в течение последних 10-15 лет, подчиняется экспоненциальному закону. По приблизительным оценкам в 2020 году будет создано около 40 ZB данных, в 2025 – более 140 ZB. Хотя термин «очень большие наборы данных» не относятся к какой-либо конкретной величине, в настоящее время чаще всего подразумеваются значения от петабайт (1 PB = 10^{15} байт) до эксабайт (1 EB = 10^{18} байт). Эта условная граница будет постепенно сдвигаться в область зеттабайт (1 ZB = 10^{21} байт), йоттабайт (1 YB = 10^{24} байт) и далее.

Где производится такое количество данных? В современном мире повсюду – в Интернете, науке, телекоммуникациях, социальных сетях, финансовых организациях, на фондовых, сырьевых и валютных рынках, в розничной торговле, транспорте, энергетике, промышленности, интернете вещей, городской среде, умных городах. За последние несколько лет было сгенерировано больше данных, чем за всю прошлую историю человечества. В одном только Google генерируется 40,000 поисковых запросов каждую секунду, что составляет 1,2 триллиона запросов в год. Пользователи Facebook в среднем каждую минуту отправляют 30 миллионов сообщений. Происходит быстрый рост объёмов видео и фото, более 300 часов видеофайлов каждую минуту загружаются на один лишь YouTube.

При этом возрастает вклад науки в рост данных. Это – физика высоких энергий, астрономия и астрофизика, биология, исследования климата, нанотехнологии и другие научные отрасли. Так, в процессе работы Большого адронного коллайдера (LHC – Large Hadron Collider) [7] в ЦЕРН столкновения частиц фиксируются миллионами датчиков, генерирующие огромные массивы данных. Совокупный объем информации, производимый четырьмя экспериментальными установками на LHC, составляет около 20 PB в год, объем накопленных к настоящему времени данных – около 200 PB. Эти объемы сопоставимы, а порой и превосходят объемы данных, накопленных в хранилищах традиционных генераторов и поставщиков Больших данных, таких как Facebook, чьи фото и видео хранилища оцениваются как минимум в 100 PB. Следует отметить, что потоки данных после реконструкции ускорителя (до уровня High Luminosity LHC) вырастут более, чем на порядок. Другим «поставщиком» быстро растущих объемов научных данных является астрономия. В настоящий момент это, прежде всего, проекты SDSS (Sloan Digital Sky Survey, Нью-Мексико, США) и VISTA (Visible and Infrared Telescope for Astronomy, Чили), в 2019 году должен начать работу телескоп-рефлектор LSST (Large Synoptic Survey Tele-scope, Чили) с потоком данных более 10 PB в год, в 2022 – тридцатиметровый телескоп TMT (Thirty Meter Telescope, Гавайи, США) с потоком данных более 30 PB в год. Рекордные объемы данных ожидаются от строящегося в ЮАР радиотелескопа SKA (Square Kilometer Array Radio Telescope) [8] – 300 PB в год, начиная с 2023 года, и 1 EB в год после 2025 года.

Комплекс понятий Больших данных, появившись в 2005 г., постепенно обрастая новыми смыслами и технологиями, в течение последних лет находится практически на максимальном уровне интереса, что подтверждается данными сервиса Google Trends, и о чем говорит постоянно растущее количество научных публикаций на тему Больших данных – например в журналах, индексируемых Scopus, в 2014 году было опубликовано около 5000 работ, а в 2019 – уже более 19000 (см. рис.1).

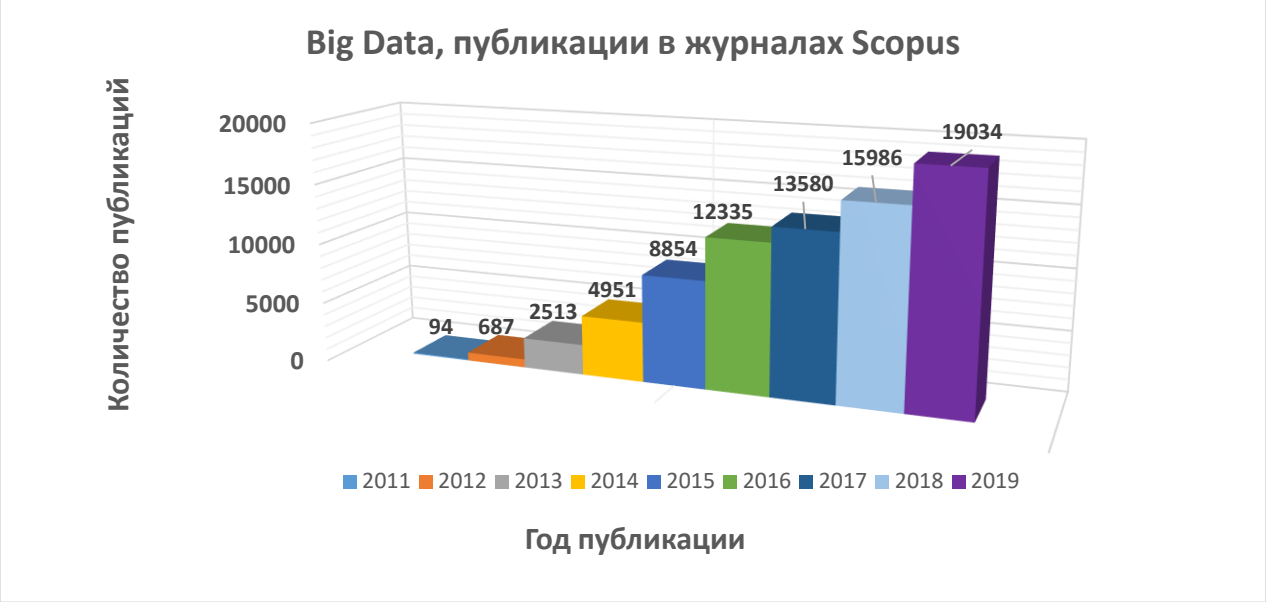

Рис. 1. Количество публикаций с 2011 по 2019 в журналах, индексируемых Scopus, в которых термин Big Data присутствует в названии, ключевых словах или аннотации

На рис.2 представлено распределение публикаций по отраслевым категориям, отражающее соотношение объема исследований в области Больших данных в различных отраслях знаний, в том числе – прикладных исследований в области бизнеса, управления, энергетики, материаловедения, социальных областях, медицине и т.д.

Большие данные – это «зонтичный» термин, объединяющий группу понятий, технологий и методов обработки очень больших объемов данных в распределенных информационных системах, обеспечивающих организацию качественно новой полезной информации (знаний). Большие данные – это не только сами данные, но и технологии их обработки и сценарии использования, которые можно объединить в несколько основных категорий: а) новые алгоритмы, пригодные для обеспечения поиска в огромных наборах данных и их обработки; б) новые технологии управления данными, позволяющие работать со сложными, неоднородными и распределенными источниками данных; в) высокопроизводительные вычислительных системы; г) архитектуры и алгоритмы, позволяющие обрабатывать потоковые данных, поступающие из высокоскоростных сетей, от приборов и датчиков; д) высокопроизводительные и высоконадежные распределенные файловые системы, способные работать с данными петабайтного диапазона; е) технологии, позволяющих интегрировать разнородные данные из разных источников – слияние и интеграция данных (data fusion and integration).

Аналитика Больших данных предусматривает следующие этапы исследования: а) изучение области применения (предварительные знания, цели), б) сбор данных из различных источников, в) создание целевого набора данных, г) очистка данных – процесс обнаружения и исправления поврежденных или неточных данных из набора записей, таблицы или базы данных, д) преобразование данных – подготовка данных для дальнейшего анализа с помощью приведения к стандартной форме или нормировки, е) сокращение объема – поиск значимых признаков, сокращение размерности, ж) описание природы данных, выбор необходимых инструментов (регрессия, кластеризация, классификация), исследование связей в данных, моделирование данных, проверка моделей, прогнозная аналитика, з) интеллектуальный анализ данных – поиск интересующих паттернов, и) обнаружение знаний – трансформация, визуализация, к) использование полученных знаний.

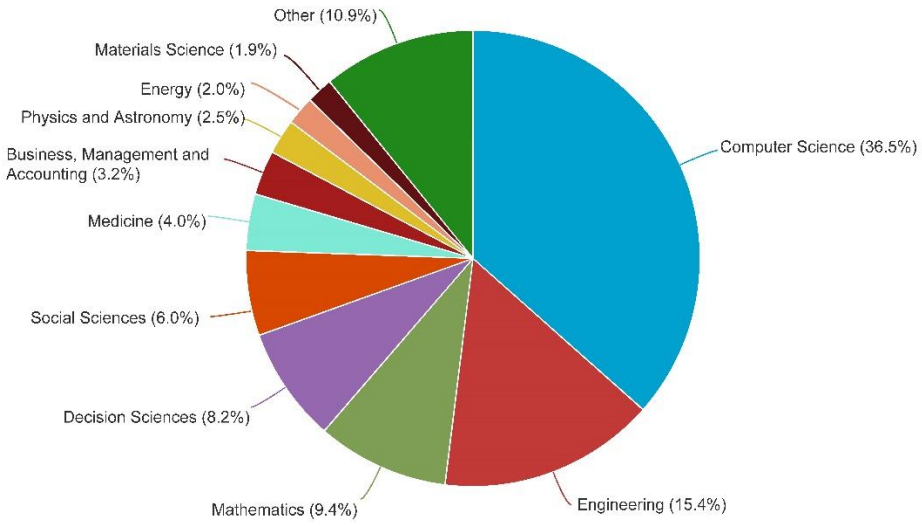

Рис. 2. Распределение публикаций по отраслям

Технологии Больших данных образуют большую экосистему, состоящую из нескольких основных категорий: а) базы данных NoSQL – MongoDB, MemCache, Hbase, Hy-perTable, Cassandra, Neo4J; NewSQL – Hive, Impala, BayesDB, Sensei, Drizzle; б) распределенные файловые системы (для масштабирования и работы на нескольких серверах сразу) – HDFS, Red Hat ClusterFS, Apache MapReduce, Apache Pig, Apache Spark; в) машинное обучение (извлечение полез-ных знаний из данных) – Mahout, WEKA, Spark, Python, R; г) планирование (для автоматизации повторяющихся операций и запуск заданий по событиям – Oozie, Falcon; д) развертывание системы (автоматизация установки и настройки компонентов в кластерах Больших данных) – Mesos, HUE, Ambari; е) программирование служб (обеспечение доступа к приложениям больших данных как к сервису) – Apache Thrift, Zookeper; е) безопасность (создание централизованной и высокоточной системы управления доступом к данным) – Sentry, Ranger; ж) интеграция данных (объединение данных, находящихся в разных источниках) – Apache Flume, Sqoop, Scribe.

Неотъемлемой частью Больших данных и важной компонентой их экосистемы являются методы, алгоритмы и технологии, образующие группу понятий машинного обучения, развивающихся на основе так называемых нейронных сетей глубокого обучения. Машинное обучение является предметом очень быстрого и успешного развития, находится на пике интереса не только информационных технологий, экономики, промышленности, других прикладных областей, но также широкой общественности и даже политики. Машинное обучение – средство поиска правил и связей в очень больших объемах информации. Технологии машинного обучения позволяют не просто проверять гипотезы, но искать неизвестные ранее факторы.

Области применения машинного обучения: а) системы компьютерного зрения, б) анализ и синтез изображений и видео, в) анализ и классификация текстов, г) распознавание и синтез речи, д) медицинские приложения, е) технические приложения, ж) научные приложения.

В настоящее время тема Больших данных достаточно полно отражена в мировой литературе, см., например, [9-16].

2. Роль Больших данных в цифровой экономике

Сложность современных социально-экономических систем порождает лавинообразный рост многообразия форм представления, структурной сложности и объема информации, связанный с происходящими процессами. Это качественно меняет требования как к подходам к анализу подобных масштабных систем, так и к созданию средств принятия решений по эффективному влиянию на процессы. В качестве примеров подобных систем можно привести частично регулируемые государством социаль-

ные институты (рынок труда, система высшего и среднего профессионального образования), банковский сектор, транспортную, коммунальную и социальную среду мегаполисов. Для решения задач, возникающих при анализе данного класса систем, а также для эффективного и своевременного принятия решений, необходимы новые научные и технологические подходы.

2.1. Применение Больших данных в бизнесе, экономике, обществе

Задачи, требующие применения технологий Больших данных, характерны для целого ряда отраслей и видов деятельности. Прежде всего это банковский сектор, страхование, ценные бумаги и инвестиции, производство, розничная и оптовая торговля, профессиональные и развлекательные услуги, здравоохранение, транспортные и коммуникационные услуги, СМИ, строительство, образование, управление и др. Умение оперировать большими объемами информации, анализировать взаимосвязи между ними и принимать взвешенные решения несет потенциал для увеличения показателей доходности, прибыльности и повышения эффективности. Внедрение технологий Больших данных помогают компаниям: а) снижать затраты, поскольку обеспечиваются более эффективные способы ведения бизнеса, б) создавать новые продукты и услуги на основе лучшего понимания потребностей клиента, в) лучше и быстрее принимать решения за счет быстрой обработки и анализа информации из самых разных источников данных.

В качестве примера использования технологий Больших данных в финансовой сфере можно указать такие виды деятельности как анализ кредитоспособности, управление рисками, предотвращение мошенничества, управление активами, электронные торговые площадки. В сфере энергетики это – управление режимами энергосистемы, мониторинг технического состояния электронно-сетевого оборудования, прогнозирование. В области транспорта – оптимизация и планирование расписания грузоперевозок, контроль дислокации и выполнения графиков и др. Ряд проектов в области Больших данных выходит за отраслевые рамки. Например, активно внедряемая по всему миру концепция «умного города» (Smart City), которая представляет собой интеграцию информационных и коммуникационных технологий для управления городским имуществом, объединяющая транспортные системы, школы, библиотеки, больницы, электростанции, водоснабжение и многое другое. McKinsey Global Institute предсказывает появление 600 «умных» городов к 2020 году. В настоящий момент уже есть города, где концепция реализуется на широкой основе: Сингапур, Фудзисава (Япония), Милтон-Кинс (Великобритания), Иньчуань (Китай).

По данным McKinsey, эффект применения от внедрения технологий Больших данных может достигать следующих величин: 300 млрд. долл. – ежегодный потенциал для здравоохранения в США, 250 млрд. евро – ежегодный потенциал управления государственным сектором EU, 600 млрд. долл. – ежегодный мировой потенциал увеличения покупательского спроса только за счет использования информации о перемещениях покупателей, 60% – потенциал увеличения выручки в розничной торговле.

2.2. Инициатива ООН в области Больших данных

Global Pulse [17] – проект ООН в области Больших данных. Миссия – ускорение разработок и масштабного внедрения инноваций в области Больших данных для устойчивого развития и гуманитарной деятельности. Инициатива была основана на признании того, что цифровые данные дают возможность лучше понять изменения в благосостоянии человека и получить обратную связь в режиме реального времени о том, отражающую эффективность принимаемых политических решений. Global Pulse разрабатывает аналитические инструменты и подходы в своих лабораториях, внедряет полезные инновации во всей системе ООН.

На сайте Global Pulse десятки проектов. Некоторые примеры – «Улучшение транспортного планирования с помощью анализа данных в реальном времени», «Изучение влияния экстремистского насилия на разжигание ненависти в интернете», «Информирование госструктур с помощью майнинга соцсетей».

2.3. Большие данные на государственном уровне

Многие страны приступили к осуществлению планов и инициатив в области исследований и прикладных программ в сфере Больших данных. В 2012 года администрация президента США официально

запустила инициативу по исследованию и развитию Больших данных с инвестициями более 200 миллионов долларов. В инициативе принимали участие несколько федеральных государственных учреждений – Министерство обороны (DoD), Агентство оборонных исследований (DARPA), Министерство энергетики (DoE), Национальный институт здравоохранения (NIH), Национальный научный фонд (NSF) и Геологическая служба США (USGS).

Помимо США, Великобритания, Франция, Австралия и Япония также представили свои инициативы в области Больших данных. Европейская комиссия в своей рамочной программе исследований и инноваций Horizon 2020 инвестировала около 120 миллионов евро в промышленные исследования и приложения, связанные с Большими данными.

В России в 2017 году утверждена правительственная программа «Цифровая экономика Российской Федерации», предусматривающая поддержку развития как уже существующих прорывных и перспективных сквозных цифровых платформ и технологий, так и создание условий для возникновения новых. Большие данные входят в список 9 основных цифровых технологий, указанных в программе.

2.4. Программа «Цифровая экономика Российской Федерации»

Программа утверждена 28.07.2017 распоряжением Правительства РФ №1632 и осуществляется в соответствии с целями, задачами, направлениями, объемами и сроками реализации основных мер государственной политики Российской Федерации по созданию необходимых условий для развития цифровой экономики Российской Федерации, в которой данные в цифровой форме являются ключевым фактором производства во всех сферах социально-экономической деятельности, что повышает конкурентоспособность страны, качество жизни граждан, обеспечивает экономический рост и национальный суверенитет.

Программа предусматривает развитие следующих сквозных технологий: большие данные, нейротехнологии и искусственный интеллект, системы распределенного реестра, квантовые технологии, новые производственные технологии, промышленный интернет, компоненты робототехники и сенсорика, технологии беспроводной связи, технологии виртуальной и дополненной реальностей.

Потенциальный эффект для ВВП страны от цифровизации экономики к 2025 году оценивается в 4,1-8,9 трлн. рублей, что составит от 19 до 34% общего увеличения ВВП. Общая цель – рост объемов цифровой экономики в 3 раза [18].

Наряду с нефтью, газом и другими природными ресурсами, данные становятся новым активом экономики [19], в которой технологии и платформы Больших данных будут играть одну из ключевых ролей.

2.5. Платформы Больших данных для задач цифровой экономики

В качестве примеров реализации платформ для задач цифровой экономики рассматриваются решения для системы высшего образования и кредитного рынка, разработанные в рамках проектов «Автоматизированная информационная система мониторинга и анализа кадровых потребностей по номенклатуре специальностей высшего учебного заведения» [20, 21] и «Автоматизированный анализ деятельности компаний-нерезидентов РФ» [22], выполненных по заказу РЭУ им. Г.В. Плеханова.

Система высшего образования должна отвечать потребностям рынка труда и быстро адаптироваться к его изменениям. Но на практике осуществление такой гибкой настройки для инертной и консервативной образовательной системы является очень сложной задачей. Целью проекта «Автоматизированная информационная система мониторинга и анализа кадровых потребностей по номенклатуре специальностей высшего учебного заведения» является предоставление дополнительных возможностей для выявления качественных и количественных взаимосвязей между образованием и рынком труда. Система определяет соответствие содержания образовательных программ вузов потребностям рынка труда на основе нейросетевой модели word2vec, реализующей идею дистрибутивной семантики. Система построена по модульному принципу и включает: а) модуль сбора текстовых данных (работающий в автоматическом режиме с использованием открытых источников, которыми являются интернет-порталы и кадровые агентства), б) модуль загрузки и хранения данных, состоящий из распределенного хранилища данных, обеспечивающего репликацию и архивирование, в) модуль автоматической

обработки, осуществляющий подготовку информации к анализу и автоматическое связывание требований рынка труда с компетенциями образовательного стандарта, а также машинное обучение, г) пользовательский интерфейс для создания и отображения отчетов на основе технологий бизнес-аналитики. Общая схема обработки данных приведена на рис. 3. Система предназначена для широкого круга пользователей и в первую очередь для руководителей регионов, вузов, компаний, кадровых агентств.

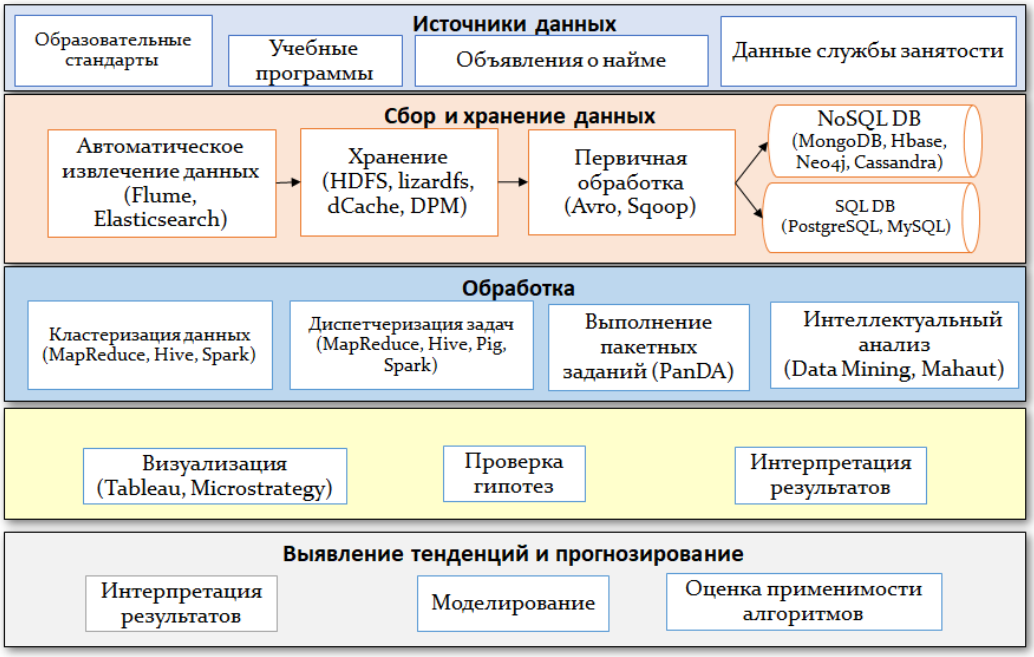

Рис. 3. Схема автоматизированной информационной системы мониторинга и анализа кадровых потребностей по номенклатуре специальностей высшего учебного заведения

Современный банковский бизнес невозможен без использования информационных систем. Банковский контроллинг не является исключением, и для эффективной реализации его функций необходимы передовые ИТ-решения. Это позволит обеспечить банковский надзор значительным охватом кредитных организаций различного масштаба – от небанковского кредитных учреждений и банков с базовыми лицензиями к банкам с универсальными лицензиями. В рамках проекта разработана автоматизированная система анализа деятельности компаний-нерезидентов РФ с целью выявления цепочки контрагентов и конечных бенефициаров. Система использует два типа сбора информации из источников: по расписанию и по запросу. Плановое задание реализуется в виде периодически выполняемого задания, выполняющего загрузку архивов, сканирование источников на предмет наличия информации о компаниях, а также обновление информации о субъектах, уже внесенных в базу данных. На базе разработанной системы может быть развернута информационная служба, предоставляющая комплекс услуг по контролю компаний-нерезидентов и принятию решений по конкретным компаниям в соответствии с ключевыми требованиями регуляторов. При реализации программной инфраструктуры системы был использован стек программных продуктов и инструментов с открытым исходным кодом промышленного уровня, ставших де-факто отраслевыми стандартами в своих областях: Spark, Hadoop, Kafka (буферизация данных), Flume (потоковая обработка), Marathon, Docker, Elasticsearch. Схема взаимодействия компонентов программной инфраструктуры представлена на рис.4. Используемые технологии и алгоритмы позволили реализовать сбор и совместный анализ структурированных и неструктурированных данных из разных источников. Развертывание кластеров этих базовых компонентов обеспечивает масштабируемость и высокую доступность системы.

Рис. 4. Схема взаимодействия компонентов программной инфраструктуры автоматизированной системы анализа деятельности компаний-нерезидентов РФ

Заключение

В связи с повсеместным внедрением и развитием цифровизации в различных отраслях экономики, обоснованным выглядит мнение, что данные теперь это «новая нефть». Одной из характерных особенностей цифровой экономики является необходимость получения, накопления и всестороннего анализа данных о различных процессах, от производственных до социальных, от управления бизнесом и техническими системами до управления государством. В данной статье обобщен опыт авторов по разработке программно-аппаратных аналитических платформ и их применению для решения практических задач цифровой экономики. Рассмотрена эволюция задач и подходов к аналитике Больших данных, а также рост масштабов применения подобных технологий и платформ в различных областях знания – от научных мегапроектов до государственных программ и инициатив Организации объединенных наций. В качестве конкретных примеров использования аналитических платформ Больших данных приведены успешные социально-экономические исследования, выполненные авторами в составе научных коллективов на базе Российского экономического университета им. Г.В. Плеханова.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда (проект № 19-71-30008).

Список литературы

1. Andrea De Mauro, Marco Greco and Michele Grimaldi. “What is Big Data? A Consensual Definition and a Review of Key Research Topics”. In “AIP Proceedings”2014, “4th International Conference on Integrated Information”.
2. Cox M., Ellsworth D. The problem of big data. Application-Controlled Demand Paging for Out-of-Core Visualization. Proceedings of the 8th IEEE Visualization '97 Conference.
3. D. Laney, “3D Data Management: Controlling Data Volume, Velocity & Variety”, META-Group, 2001.
4. J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters”, Commun. ACM, vol. 51, no. 1, Jan. 2008.
5. Lynch C. How do your data grow? Nature. — 2008. — Vol. 455. — № 7209. — P. 28–29.
6. Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Byers A.H. Big data: The next frontier for innovation, competition, and productivity: McKinsey Global Institute Report. — May 2011.
7. LHC (Large Hadron Collider). Web: https://home.cern/science/accelerators/large-hadron-collider
8. Square Kilometre Array. Public Website: https://www.skatelescope.org/.
9. Bernard Marr. “Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance”. John Wiley & Sons Ltd, 2015.
10. Дэви Силен, Арно Мейсман, Мохамед Али. Основы Data Science и Big Data. Python и наука о данных. — СПб.: Питер, 2017.
11. Натан Марц и Джеймс Уоррен. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени. — Издательский дом «Вильямс», 2016.
12. Big Data Analytics. Tools and Technology for Effective Planning. Edited by Arun K. Somani, Ganesh Chandra Deka. CRC Press. Taylor & Francis Group. — 2018.
13. Кэти О’Нил, Рэйчел Шатт. Data Science. — СПб.: Питер, 2019.
14. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets. — Cambridge University Press, 2012.
15. Анналин Ын, Кеннет Су. Теоретический минимум по Big Data. Все, что нужно знать о Больших данных. — СПб.: Питер, 2019.
16. Zhanpeng Huang, Pan Hui, Christoph Peylo. When Augmented Reality Meets Big Data. — arXiv:1407.7223v1.
17. https://www.unglobalpulse.org/about-new.
18. Аптекман А. и др., Цифровая Россия: новая реальность. — Digital/McKinsey. 2017.
19. Цифровая экономика. Видение 2025. — «Ростех», 2017.
20. Зрелов П., Кореньков В.В., Кутовский Н. Петросян А.Ш., Семенов Р. Филозова И.А. Мониторинг потребностей рынка труда в выпускниках вузов на основе аналитики с интенсивным использованием данных // Труды XVIII Межд. конф. DAMDID/RCDL. — М.: ТОРУС Пресс, 2016. — С. 124-131.
21. Belov S., Filozova I., Kadochnikov I., Korenkov V., Semenov R., Zrelov P. Automated system to monitor and predict matching of higher vocational education programs with labour market // Proc. of 26th International Symposium NEC-2017, Budva, 2017. — http://ceur-ws.org/Vol-2023. — P. 98-104.
22. Badalov L.A., Belov S.D., Kadochnikov I.S. Checking foreign counterparty companies using Big Data // Selected Papers of the 8th International Conference "Distributed Computing and Grid-technologies in Science and Education". — Vol. 2267 of CEUR Workshop Proceedings. — CEUR-WS.org, 2018. — P. 523-527. — http://ceur-ws.org/Vol-2267/523-527-paper-100.pdf.