

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение  
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра математической кибернетики и компьютерных наук

**РАЗРАБОТКА БЫСТРОРАЗВОРАЧИВАЕМОГО  
ПРОГРАММНОГО ПРОДУКТА, ОБЕСПЕЧИВАЮЩЕГО  
ВИЗУАЛИЗАЦИЮ И МОНИТОРИНГ ДАННЫХ ПО НОВОЙ  
КОРОНАВИРУСНОЙ ИНФЕКЦИИ (COVID-19) НА БАЗЕ  
СЕРВИСОВ, ПРЕДОСТАВЛЯЕМЫХ ОБЛАЧНОЙ  
ПЛАТФОРМОЙ.**

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 411 группы  
направления 02.03.02 — Фундаментальная информатика и информационные  
технологии  
факультета КНиИТ  
Тепер Анны Леонидовны

Научный руководитель  
доцент, к. т. н.

\_\_\_\_\_

В. М. Соловьев

Заведующий кафедрой  
к. ф.-м. н., доцент

\_\_\_\_\_

А. С. Иванов

Саратов 2020

## ВВЕДЕНИЕ

Настоящая работа посвящена разработке информационных технологий и программных решений, позволяющих обеспечить мониторинг и визуализацию большого объема данных, поступающих в реальном времени, с помощью инструментов, предоставленных облачными платформами, и технологий Apache Airflow и Elasticsearch. В качестве примера больших данных использована информация по заболеваемости коронавирусом COVID-19 в России и Европе.

*Актуальность работы* обусловлена необходимостью постоянного мониторинга ситуации с продолжающейся пандемией COVID-19 с целью быстрого принятия эффективных управленческих решений.

Рассмотренные в данной работе технологии в будущем могут быть применены также в случаях других пандемий, природных и техногенных катастроф, других чрезвычайных событий, распространяемых на больших территориях и влияющих на большое количество людей, когда данные для анализа поступают в реальном времени.

*Была поставлена задача* изучить рынок облачных платформ, предоставляющих сервисы для работы с большими данными, выбрать подходящую платформу для реализации решения и, на базе облачных вычислений, разработать быстроразворачиваемый программный продукт, обеспечивающий визуализацию и мониторинг данных.

*Целью настоящей работы* является с использованием разработанного программного решения обеспечение визуализации и построение отчетных форм по данным о заболеваемости COVID-19, по количеству проведенных тестов, по выздоровлению и смертности от болезни, с использованием исходных данных, имеющих в открытом доступе.

Работа состоит из введения, двух теоретических и двух практических глав, заключения, списка использованных источников и приложений.

## 1 Основное содержание работы

Введение содержит основные положения: формулировку поставленной задачи и цель данной выпускной квалификационной работы.

В первой главе «**Предметная область и постановка задачи**» рассматривается предметная область работы – определение инфекции COVID-19 и общее описание основных технологий, использующихся для решения поставленной задачи: *Big Data* и *облачных вычислений*.

Коронавирусы — это оболочечные вирусы, содержащие одноцепочечную рибонуклеиновую кислоту. [1] В настоящее время известно шесть типов коронавируса, которые заражают человека и вызывают у него заболевания. Как правило, коронавирусы вызывают респираторные заболевания лёгкого течения, однако, возможны вспышки заболеваемости коронавирусной инфекцией с большим количеством летальных исходов. В декабре 2019 года были зарегистрированы первые случаи выявления пневмонии неизвестной этиологии в городе Ухань (КНР). [2] Впоследствии был выделен новый тип коронавируса под названием SARS-CoV-2. Всемирная организация здравоохранения (ВОЗ) 11 февраля 2020 г. присвоила официальное название инфекции, вызванной новым коронавирусом, – COVID-19 («Coronavirus disease 2019»).

В мире зарегистрировано более 5,2 миллионов случаев заражения, из них более 335 тыс. в России (по состоянию на 23 мая 2020 г.). Летальных исходов в мире более 339 тыс., в России – более 3,3 тыс. [3]

Появление COVID-19 поставило перед специалистами здравоохранения задачи, связанные с быстрой диагностикой и оказанием медицинской помощи больным, а перед государственными и муниципальными властями задачи своевременного наращивания в достаточном количестве мощностей медицинских инфекционных подразделений и обеспечения их оборудованием, материалами и квалифицированным медицинским персоналом.

В буквальном понимании термин *Big Data*, действительно, означает большие по объёму (в терабайтах, петабайтах и экзобайтах) данные. Но суть этого термина в другом — более точной формулировкой может служить «*сложные данные*».

*Большие данные* (англ. *Big Data*) — серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимае-

мых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети, сформировавшихся в конце 2000-х годов, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence.

В сущности, понятие больших данных подразумевает работу с информацией огромного объема и разнообразного состава, весьма часто обновляемой и находящейся в разных источниках в целях увеличения эффективности работы, создания новых продуктов и повышения конкурентоспособности. Консалтинговая компания Forrester дает краткую формулировку: «Большие данные объединяют техники и технологии, которые извлекают смысл из данных на экстремальном пределе практичности».

*Облачные вычисления* — это такой подход к размещению, предоставлению и потреблению приложений и компьютерных ресурсов, при котором приложения и ресурсы становятся доступны через Интернет виде сервисов, потребляемых на различных платформах и устройствах. Оплата таких сервисов осуществляется по их фактическому использованию.

Можно выделить следующие основные сервисы, предоставляемые облачными платформами:

- программное обеспечение как сервис (SaaS);
- платформа как сервис (PaaS);
- инфраструктура как сервис (IaaS).

Также в настоящее время выделяют три модели развертывания облаков:

- публичные;
- частные;
- гибридные.

В настоящей работе ставится задача визуализации и мониторинга в реальном времени данных по количеству тестов, заболеваемости COVID-19, по смертности и по выздоровлению, использующие базовые познавательные структуры человека, с целью увеличения объема воспринимаемой информации и повышения качества принимаемых на ее основе управленческих решений.

Информационная система, обеспечивающая сбор, мониторинг, анализ и визуализацию данных по тестированию, заболеваемости, смертности и вы-

здоровлению от COVID-19, должна обеспечивать решение следующих задач:

- планирование и запуск задач загрузки информации из источников данных;
- сбор, обработка и анализ потоковых данных в режиме реального времени;
- захват, преобразование и загрузка данные потоковой передачи в озера данных, хранилища и инструменты аналитики;
- хранение больших объемов данных, масштабирование их его в любом направлении в зависимости от изменений в потребностях;
- визуализация и исследование данных и временных рядов в разных видах (гистограммы, линейные графики, круговые диаграммы, тепловые карты и встроенная геопространственная поддержка), мониторинг приложений и использование в оперативных аналитических данных.

*Мониторинг данных* — это очень важная задача, позволяющая гарантировать работоспособность и высокую доступность к информации, данным и работы с ними. К сожалению, недостаточное внимание уделяется данной задаче, а это часто приводит к проблемам, неправильным управленческим решениям, иногда преступному бездействию.

В случае серьезных и продолжительных чрезвычайных ситуаций, таких как пандемии, крупные техногенные и природные катастрофы и другие – мониторинг просто жизненно необходим. Очень важно и нужно следить за текущими показателями, а также прогнозировать вероятные исходы и анализировать эти показатели, просчитывать риски появления определенных событий и так далее.

При этом система мониторинга должна обладать средствами визуализации результатов анализа в виде графиков и диаграмм, формирования отчетов для поддержки принятия решений; применения ГИС-технологий для реализации возможности визуализации результатов анализа и мониторинга данных.

Визуализация «больших данных» («Big Data Visualization») относится к областям как научной, так и информационной визуализации. В первом случае «большие данные» возникают в результате сложного компьютерного моделирования различных объектов и процессов. Во втором – имеет место визуальное описание и представление абстрактной информации, получаемой

в результате процесса сбора и обработки многокатегориальных данных, для анализа которых необходимо применение нескольких количественных и качественных мер оценки. [4]

Вторая глава **«Выбор платформы и инструментов для реализации решения задачи»** посвящена выбору облачной платформы для реализации программного продукта. Для этого изучен рынок облачных платформ и проведено сравнение трех самых крупных представителей рынка облачных платформ:

- *Amazon Web Services* Была создана в 2006 году и стала первооткрывателем в данной области, благодаря чему завоевала немалый рынок. С постоянными нововведениями и улучшениями на протяжении многих лет, AWS представила более 70 услуг с широким спектром покрытия по всему миру. Серверы доступны в 14 географических регионах. По данным отчета за 2020 год от Synergy Research Group [5] Amazon занимает 33% на мировом рынке.
- *Microsoft Azure* Система была запущена в 2010 году и развивается очень быстрыми темпами. Microsoft Azure сейчас представляет собой многогранную сложную систему, которая обеспечивает поддержку множества различных услуг, языков программирования и фреймворков. В составе облака более 60 служб и центров обработки данных в 38 различных географических регионах. В 2019 году Microsoft Azure занимала 18% рынка.
- *Google Cloud Platform* Представленная в 2011 году, Google Cloud Platform является самой молодой облачной платформой среди «тройки лидеров» и, в первую очередь, удовлетворяет потребности поиска Google и Youtube. В настоящее время у компании представлено более 50 услуг и 6 глобальных центров обработки данных. В 2017 году Google Cloud Platform на рынке облачных услуг имела 5% долю. [6]

Сравнение производилось по следующим характеристикам облачных платформ:

1. предоставляемые вычислительные мощности;
2. хранилища данных;
3. ценовая политика платформ.

Для разрабатываемого программного продукта была выбрана платфор-

ма **Amazon Web Services**. Этот выбор обусловлен следующими соображениями:

- лидирующее положение на рынке облачных платформ и, соответственно, большое количество доступных для изучения экспертных данных, полученных на основе опыта его эксплуатации;
- возможность использования готовых технологических решений для работы с данными, необходимыми для решения поставленной задачи, таких как Amazon Kinesis Data Firehose и Amazon Elasticsearch Service;
- относительно низкие цены на большинство используемых в решении сервисов, удобная ценовая политика.

В третьей главе «**Инструментальные средства, используемые для решения задачи**» описаны используемые в реализации программного продукта инструменты.

Учитывая выбор облачной платформы, для реализации решения поставленной задачи были выбраны инструменты двух видов:

- Инструменты на базе *Amazon Web Services (AWS)*:
  - средство сбора, анализа, преобразования и передачи потоки журналов, событий и метрик в различные сервисы AWS Amazon Kinesis Agent;
  - сервис для доставки потоковых данных в режиме реального времени *Amazon Kinesis Data Firehose*;
  - служба развертывания, эксплуатации и масштабирования кластеров в облаке *AWS Amazon Elasticsearch Service*;
  - приложение для поиска и визуализации данных *Kibana*.
- Инструменты для обеспечения работы системы:
  - платформа для программной разработки и мониторинга конвейеров пакетных данных *Apache Airflow*;
  - инструмент для автоматизированного развертывания приложений в контейнерах *Docker*;
  - инструмент для безопасного и эффективного построения, изменения и создания версий инфраструктуры *Terraform*;
  - средство для конфигурирования информационных систем, развертывания программного обеспечения *Ansible*.

В четвертой главе «**Структура проекта**» описываются взаимодей-

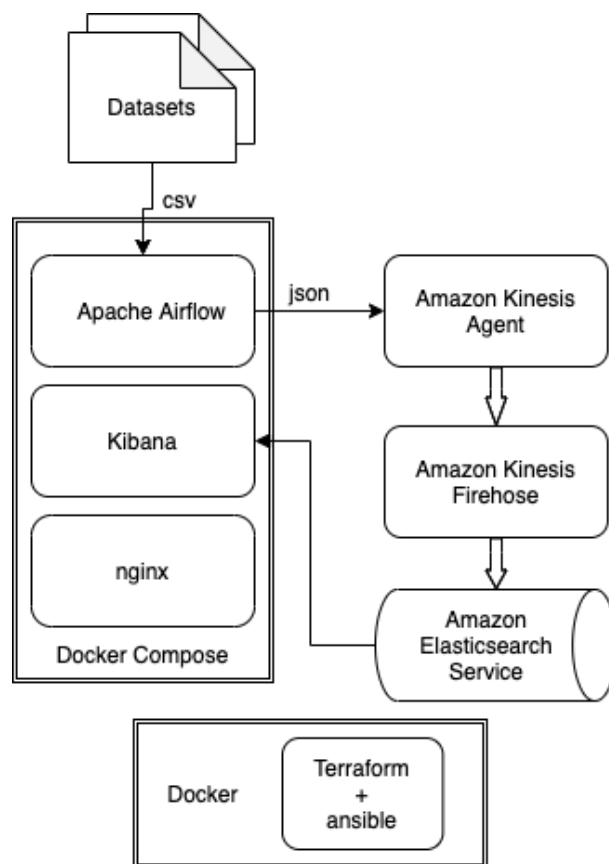


Рисунок 1 – Схема взаимодействия инструментальных средств

ствия инструментальных средств и потоков данных в программном продукте, источники обрабатываемых данных и полученные результаты работы программного продукта.

Схема показана на рисунке 1.

*Apache Airflow* обеспечивает планирование и выполнение в заданные моменты времени задач по считыванию исходных данных, производит их декомпозицию и преобразование, формирует и передает поток данных для дальнейшей обработки в формате JSON. При этом исходные данные могут быть получены из разных источников и иметь разный формат. Разворачивание приложения на сервере производится через *Docker Compose*.

*Kinesis Agent* получает поток данных и перенаправляет их в *Kinesis Firehose*, который в свою очередь преобразует данные и подготавливает для хранения. Затем данные передаются на хранение в хранилище в *Amazon Elasticsearch Service*.

Последним в цепочке используемых программных инструментов является *Kibana*, которая, используя обширные возможности *Elasticsearch* по поиску, фильтрации и обработки информации, получает необходимые данные,

их визуализирует на экране и строит по ним заданные отчетные формы.

Сервисы *Amazon Kinesis Agent*, *Amazon Kinesis Firehose*, *AWS Elasticsearch Service* и *Kibana* имеют web-интерфейс и работают под управлением web-сервера *nginx*. Разворачивание всего приложения производится с помощью *Docker* с использованием *Terraform* и *Ansible*.

*Источниками данных* являются два отдельных датасета представляющие данные по России и Европе соответственно, а также файл `countries.csv`, являющийся словарем географического расположения стран – один из образцов, подготовленных в рамках краткой инструкции по работе с Google Dataset Publishing Language. Данные хранятся в формате CSV.

Источником данных по России является датасет «*COVID-19 Russia regions cases*» на портале Kaggle.com [7].

Источник содержит данные о распространении вируса COVID-19 (коронавируса) в России по всем регионам. Набор данных содержит всё количество зарегистрированных (подтвержденных) случаев заражения, количество умерших и количество выздоровевших в России, начиная с января 2020 года по настоящее время. В таблицах содержатся данные как по дням, так и совокупные случаи. Данные ежедневно обновляются.

Информация в датасет в основном собрана с официальных правительственных ресурсов:

- [stopcoronavirus.rf](https://stopcoronavirus.rf) — официальный интернет-ресурс для информирования населения по вопросам коронавируса (COVID-19) [8];
- страница новостей о коронавирусе на сайте федеральной службы по надзору в сфере защиты прав потребителей и благополучия человека [9].

Некоторая информация взята из публичных ресурсов и новостей:

- Онлайн-карта распространения коронавируса [10]
- Индекс самоизоляции [11];
- Публичный дашборд Yandex DataLens [12];

Источником данных по Европе является датасет «*COVID-19 Coronavirus data*», собранный на сайте *EU Open Data Portal* [13]

Набор данных содержит последние общедоступные данные о COVID-19, включая ежедневное обновление ситуации, эпидемиологическую кривую и глобальное географическое распределение (ЕС и Великобритания, и весь

мир).

С начала пандемии коронавируса команда из Европейского центра профилактики и контроля заболеваний ежедневно собирала количество случаев и смертей от COVID-19 на основании отчетов, представленных органами здравоохранения во всем мире. Чтобы обеспечить точность и достоверность данных, процесс сбора данных постоянно совершенствуется. Это помогает отслеживать и интерпретировать динамику пандемии COVID-19 не только в Европейском союзе (ЕС), Европейском экономическом пространстве, но и во всем мире.

Поскольку в данном датасете отсутствуют географические данные стран, необходимые для визуализации данных в виде карт, в качестве словаря географического расположения стран использован файл `countries.csv`, являющийся одним из образцов, подготовленных в рамках краткой инструкции по работе с Google Dataset Publishing Language. [?]

*В результате*, разработанный программный продукт обеспечивает загрузку имеющихся исходных данных по COVID-19, их обработку и визуализацию в различных формах. Формируется также ряд отчетных форм, в том числе с использованием ГИС-технологий. Также возможно сохранение объектов как в формате CSV, так и в качестве объекта, который можно включить в html-страницу как независимый объект.

Пример формируемых отчетных форм приведен на рис. 2.

В пятой главе «**Описание кода программы**» дано подробное описание кода модулей программного продукта и выполняемых каждым модулем функций.

В **приложениях** приводится полностью код описанных модулей.

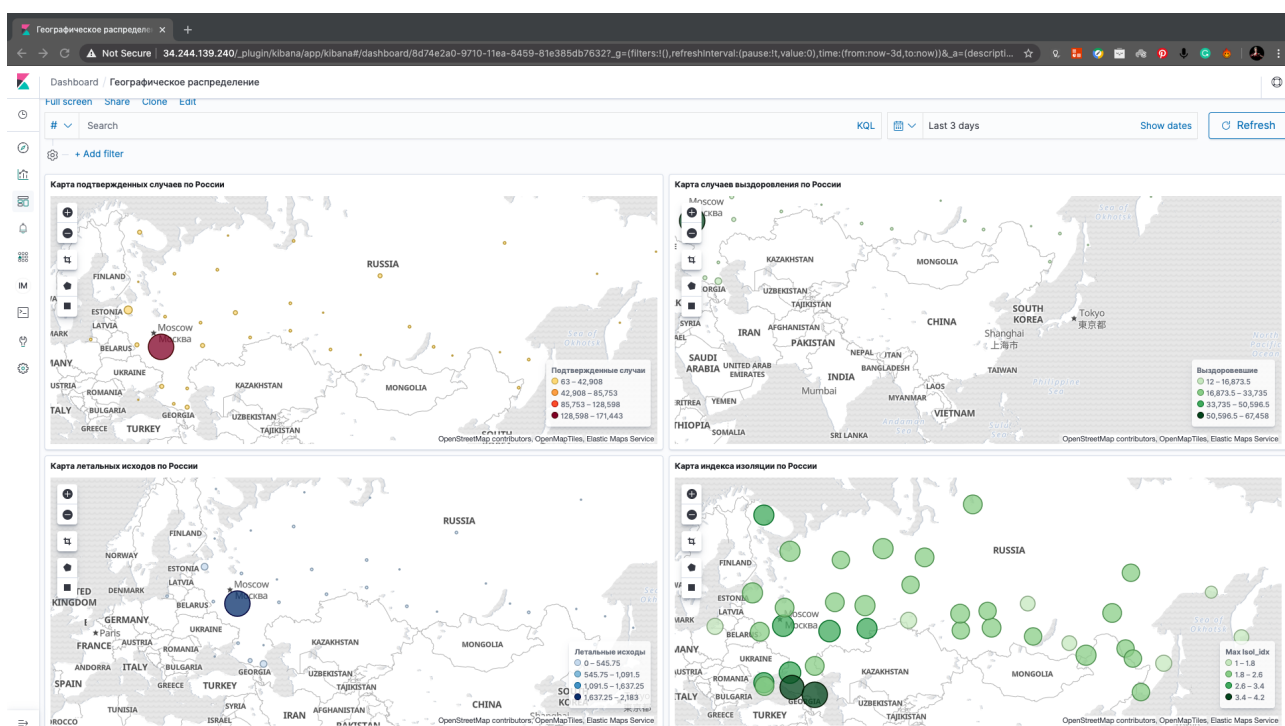


Рисунок 2 – Информационная доска «Географическое распределение»

## ЗАКЛЮЧЕНИЕ

В настоящей работе создан быстроразворачиваемый программный продукт на базе современных технологий с использованием сервисов AWS, который обеспечивает визуализацию и мониторинг данных, поступающих как в реальном времени, так и исторических данных. С использованием разработанного программного инструментария решена задача мониторинга и визуализации данных по количеству тестов, заболеваемости COVID-19, по смертности и по выздоровлению от болезни. Программное решение обеспечивает также построение необходимых отчетных форм. Источниками данных являются открытые наборы данных по COVID-19, в том числе с данными по Российской Федерации.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 *Weiss, S. R.* Coronavirus pathogenesis. / S. R. Weiss, J. L. Leibowitz // *Advances in virus research*. — 2011. — Vol. 81. — P. 85–164.
- 2 *World Health Organization*,. Pneumonia of Unknown Cause – China. [Электронный ресурс]. — URL: <https://www.who.int/csr/don/05-january-2020-pneumonia-of-unknown-cause-china/en/> (Дата обращения 20.05.2020). Загл. с экр. Яз. англ.
- 3 *Johns Hopkins University*,. CORONAVIRUS RESOURCE CENTER [Электронный ресурс]. — URL: <https://coronavirus.jhu.edu/> (Дата обращения 20.05.2020). Загл. с экр. Яз. англ.
- 4 Анализ и визуализация «больших данных» / Труды международной научной конференции «Параллельные Вычислительные Технологии» (ПаВТ'2015). — Издательский центр ЮУрГУ, 31 марта - 2 апреля 2015. — URL: <http://www.cv.imm.uran.ru/e/3241533-analiz-i-vizualizatsiya-bolshih-dannyih> (Дата обращения 20.05.2020). Загл. с экр. Яз. рус.
- 5 *Synergy*,. Incremental Growth in Cloud Spending Hits a New High while Amazon and Microsoft Maintain a Clear Lead [Электронный ресурс]. — February, 2020. — URL: <https://www.srgresearch.com/articles/incremental-growth-cloud-spending-hits-new-high-while-amazon-and-microsoft-maintain-clear-lead-reno-nv-february-4-2020> (Дата обращения 29.05.2020). Загл. с экр. Яз. англ.
- 6 *artur.baranok*,. Сравнение услуг облачных провайдеров: Microsoft Azure, AWS или Google Cloud [Электронный ресурс]. — January 2017. — URL: <https://la.by/blog/sravnenie-uslug-oblachnyh-provayderov-microsoft-azure-aws-ili-google-cloud> (Дата обращения 29.05.2020). Загл. с экр. Яз. рус.
- 7 *Kapral42*,. COVID-19 Russia regions cases [Электронный ресурс]. — URL: <https://www.kaggle.com/kapral42/covid19-russia-regions-cases/data> (Дата обращения 20.05.2020). Загл. с экр. Яз. англ.
- 8 *Правительство Российской Федерации*,. Стопкоронавирус.рф — Официальный интернет-ресурс для информирования населения по вопросам коронавируса (COVID-19 [Электронный ресурс]. —

- URL: <https://стопкоронавирус.рф/information/> (Дата обращения 20.05.2020). Загл. с экр. Яз. рус.
- 9 *Федеральная служба по надзору в сфере защиты прав потребителей и благополучия человека*,. Новости о коронавирусе [Электронный ресурс]. — URL: <https://www.rosпотребнадзор.ru/about/info/news/> (Дата обращения 20.05.2020). Загл. с экр. Яз. рус.
- 10 *Онлайн-карта распространения коронавируса*,. CORONAVIRUS (COVID-19) [Электронный ресурс]. — URL: <https://coronavirus-monitor.ru/> (Дата обращения 20.05.2020). Загл. с экр. Яз. рус.
- 11 *ООО «ЯНДЕКС»*,. Индекс самоизоляции [Электронный ресурс]. — URL: <https://yandex.ru/company/researches/2020/podomam> (Дата обращения 20.05.2020). Загл. с экр. Яз. рус.
- 12 *ООО «Яндекс.Облако»*,. Коронавирус: дашборд. Публичный дашборд Yandex DataLens [Электронный ресурс]. — URL: <https://datalens.yandex/7o7is1q6ikh23?tab=q6> (Дата обращения 20.05.2020). Загл. с экр. Яз. рус.
- 13 *European Centre for Disease Prevention and Control*,. COVID-19 Coronavirus data [Электронный ресурс]. — URL: <https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data> (Дата обращения 20.05.2020). Загл. с экр. Яз. англ.