

Михайлова Анастасия Михайловна
студент, ФГБОУ ВО «Финансовый
университет при Правительстве Российской
Федерации», г. Москва
e-mail: Coolgirly9769@yandex.ru

«БОЛЬШИЕ ДАННЫЕ»: КАК ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ МОГУТ ПОМОЧЬ СТАТИСТИЧЕСКОЙ СЛУЖБЕ ПОВЫСИТЬ ЭФФЕКТИВНОСТЬ РАСЧЕТА ИНДЕКСА ПОТРЕБИТЕЛЬСКИХ ЦЕН

Аннотация. Представлены результаты исследования применения технологий Big Data («Большие данные») при расчете индекса потребительских цен (далее – ИПЦ). Объект исследования – государственное учреждение «Федеральная служба государственной статистики» (далее – Росстат), предмет исследования – процесс расчета ИПЦ на основе данных, поступающих в Росстат. Цель исследования – разработка рекомендаций по использованию «Больших данных» в статистике потребительских цен. Бизнес- и информационно-технологическая (ИТ-) архитектура Росстата разработана с помощью пошаговой инструкции по разработке архитектуры предприятия TOGAF (The Open Group Architecture Framework).

Ключевые слова: индекс потребительских цен, Большие данные, Big Data, TOGAF, бизнес- и ИТ-архитектура.

Mikhajlova Anastasia
Student, Financial University Under
the Government of the Russian
Federation, Moscow
e-mail: Coolgirly9769@yandex.ru

BIG DATA: HOW CAN INFORMATION TECHNOLOGIES HELP THE STATISTICS SERVICE TO INCREASE EFFICIENCY OF CALCULATION OF THE CONSUMER PRICE INDEX

Abstract. The results of the study of the application of the Big Data technologies in calculating the consumer price index (CPI) has been presented. The object of the study is the state institution «Federal State Statistics Service» (ROSSTAT), the subject of the study – the process of calculating the CPI based on data coming to the ROSSTAT. The purpose of the study is to develop recommendations on the application of Big Data in the statistics of consumer prices. The business and information technology (IT) architecture of the ROSSTAT were developed with the help of the step-by-step instruction for the development of architecture TOGAF (The Open Group Architecture Framework).

Keywords: consumer price index, Big Data, hadoop, TOGAF, business and IT architecture.

Индекс потребительских цен – это индекс, который измеряет динамику цен на различные потребительские товары и услуги за определенный промежуток времени. Исчисление индекса происходит в два этапа [1]:

- расчет элементарных индексов цен для элементарных агрегатов расходов;
- расчет среднего из элементарных индексов цен с целью получить индексы высокого уровня агрегирования,

используя относительные значения полученных агрегатов расходов в качестве весов.

Возникновение систематических ошибок при подсчете индекса – следствие использования устаревшей информации о корзине товаров и услуг и методики составления сбора и выборки сведений о ценах. Объем данных, поступающих в Федеральную службу государственной статистики (далее – Росстат) из разрозненных источников (датчики, excel-файлы, дискеты), оценивают в десятки терабайт. Тенденция к непрерывному росту источников данных ведет к использованию как структурированных, так и полу- и неструктурированных данных. Для решения указанных проблем выбраны технологии «Большие данные» – технологии хранения, вычисления и сервисные услуги для будущего потенциального пользователя.

Для создания информационно-технологической (ИТ-) и бизнес-архитектуры Росстата выбран подход для разработки, планирования, а также внедрения и управления архитектурой предприятия – TOGAF [9]. Выходом на этапе анализа разработки архитектуры предприятия являются диаграммы «Бизнес-слой», «Слой приложений» и «Технологический слой».

Бизнес-слой служит для описания деятельности Росстата и ее развития. Основной упор будет сделан на бизнес-процессы «сбор статистических данных о ценах на товары и услуги», «обработка полученных статистических данных» и «хранение обработанных данных» [6].

Слой приложений служит для описания приложений, функциональности, а также отношений между ними. Единая система сбора, обработки, хранения и представления статистической информации (ЕССО) состоит из следующих подсистем: подсистема единой нормативно-справочной информации, объединенная система регистров, подсистема сбора, обработки статистической информации, хранилище статистических данных и универсальная транспортная система [7]. Подсистема сбора и обработки – это ядро системы ЕССО. Данная система обеспечивает не только процесс производства, но и выпуска статистических данных, поддерживает полный цикл работы со статистическими данными, начиная с получения неструктурированных данных от респондентов и заканчивая уже размещением первичных данных в хранилище. В его состав входит несколько блоков – это хранилище данных регионального уровня (ХДРУ) и хранилище данных федерального уровня (ХДФУ) [5].

Модель технологического слоя сегмента архитектуры РОССТАТа представляет собой детализированное описание аппаратных средств и системного программного обеспечения (далее – ПО). Согласно закону о создании реестра отечественного ПО и политики импортозамещения вычислительной техники и микроэлектроники РОССТАТ придерживается ограничения на закупку ПО, вычислительного, коммуникационного оборудования госзаказчиками. Выбор РОССТАТа сделан в пользу развития отечественной вычислительной технологии, уникальной российской архитектуры микропроцессоров «Эльбрус» [5]. Система управления базами данных (далее – СУБД) «Линтер Стандарт», успешно применяемая для решения задач управления статистическими данными, имеет второй класс защиты от несанкционированного доступа с собственным синтаксисом языка хранимых процедур с возможностью репликации с другими базами данных.

Целевая архитектура РОССТАТа будет представлена в контексте подхода для разработки, планирования, а также внедрения и управления архитектурой предприятия – TOGAF, но уже с некоторыми корректировками.

В результате применения технологий «Больших данных» произойдут некоторые изменения в реализации основных бизнес-процессов [2, рис. 4 «Обработка полученных статистических данных» и «Хранение обработанных данных»]. Данные бизнес-процессы будут усовершенствованы в связи с внедрением технологий «Большие данные». Внедрение технологий «Больших данных» приведет к изменению текущей структуры слоя приложений архитектуры РОССТАТа. Вместо ЕССО с несколькими компонентами единого хранилища данных (регионального и федерального уровня) будет использован проект фонда Apache Software Foundation – Apache Hadoop.

Состав системы Hadoop:

- распределенная файловая система HDFS;
- система распределенных файловых вычислений MapReduce;
- платформа Pig для анализа больших объемов данных на Hadoop;
- система гарантированной доставки сообщений Kafka.

HDFS (Hadoop Distributed File System) – распределенная файловая система, которая используется в проекте Hadoop. HDFS-кластер включает NameNode-серверы и DataNode-серверы. Предназначение этих серверов – хранение данных. NameNode сервер управляет пространством имен указанной файловой системы и доступом клиентов к данным [3]. Статистические данные будут размещаться в последовательных блоках данных фиксированного размера. Каждая отдельная группа товаров или услуг будет располагаться в своем файле в конкретном блоке.

MapReduce – способ организации алгоритмов, когда пользователь определяет процедуры для обработки данных. Запуски функций могут работать параллельно, на различных машинах. Предназначение системы распределенных параллельных вычислений – электронная обработка статистических данных.

Apache Kafka – это распределенная система обмена сообщениями, промежуточное звено между продюсерами и консьюмерами данных. Система собирает данные у продюсеров, хранит у себя в распределенном хранилище по топикам, а далее раздает эти данные консьюмерам по подписке. В статистике потребительских цен Apache Kafka используют для сбора и агрегации пользовательской активности на сайтах: открытие веб-страниц, данные запросов на сайте, пользовательская кликабельность.

Pig – это платформа для анализа «Больших данных», включающая язык высокого уровня в целях написания программ анализа данных, а также инфраструктуру для запуска этих программ. Написанные сценарии скрытым образом преобразуются в различные задачи MapReduce, которых исполняются в кластере Hadoop. В статистике потребительских цен платформа Pig используется для анализа полученных статистических данных и формирования агрегатных цен на основе указанных данных.

Система Hadoop будет рассматриваться как одна из составляющих универсальной платформы данных Arenadata Hadoop. Arenadata Hadoop – интегрированный набор различных компонентов корпоративного уровня на базе определенных решений с открытым исходным кодом. Дистрибутив Arenadata Hadoop распределенной платформы хранения неструктурированных и полуструктурированных данных на базе Apache Hadoop выбран на основе анализа российского аппаратного и программного обеспечения, произведенного Федеральной налоговой службой и нанятой ей консалтинговой фирмой N.

Изменения произойдут не только в бизнес-слое, слое приложений, но и в текущей технологической инфраструктуре архитектуры РОССТАТа. В зависимости от распределения типа нагрузки к имеющимся серверам отечественного производства будут добавлены сервера трех типов: сервер типа Slaves, NameNode и Resource Manager. На серверах этого типа поддерживается среда разработки компании Oracle Corporation JDK 1.8 (64-bit) с кроссплатформенным интерактивным пакетом для разработчиков java-приложений (например, java-приложение для wordcount – подсчет количества встречающихся наименований проданных товаров/услуг в элементарном агрегате, по которому будет собрана информация) [6]. Одновременно используют два вида СУБД – это Apache Hbase и Hive. Apache Hbase – это СУБД класса NoSQL-систем, которая благодаря объединению удобства пакетной обработки, обновления и произвольного доступа выигрывает у своих внутренних конкурентов – СУБД типа ключ-значение (масштабируемые распределенные хранилища). Главная характеристика – работа с данными, объемы которых измеряют терабайтами [8]. Выбор в пользу нереляционной СУБД сделан на основе специфики расчета потребительских цен. В этих СУБД размещение множества объектов с произвольным набором атрибутов разбито не на нескольких взаимосвязанных таблиц, а напротив, все данные расположены в одной сущности (например, одна группа товаров/услуг содержит множество классов, подклассов, элементарных агрегатов, репрезентативных продуктов, атрибуты которых находятся в одной сущности) [8].

Благодаря предложенному решению, основанному на технологии «Больших данных», будут преодолены систематические ошибки в подсчете индекса потребительских цен. Использование неструктурированных источников информации станет основной составляющей при подсчете ИПЦ.

Библиографический список

1. Приказ Росстата от 20.08.2012 г. № 454 (последняя редакция) «Об утверждении Методологических рекомендаций по организации наблюдения за потребительскими ценами и тарифами на отдельные виды услуг». [Электронный ресурс]. – Режим доступа: Справочная правовая система «КонсультантПлюс» (дата обращения: 15.01.2018).
2. Зараменских, Е. П. Основы бизнес-информатики. – М.: Юрайт. – 2016. – 407 с.
3. Клеппман, М. Высоконагруженные приложения. Программирование, масштабирование, поддержка. – С.: Питер. – 2018. – 740 с.
4. Марц, Н. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени / Н. Марц, Дж. Уоррен. – М.: Вильямс. – 2016. – 356 с.
5. Парфенов, Ю. Проектирование реляционных хранилищ данных. Учебное пособие. – М.: Юрайт. – 2017. – 122 с.
6. Силен, Д., Мейсман А., Али, М. Основы Data Science и Big Data. Python и наука о данных / Д. Силен, А. Мейсман, М. Али. – М.: Питер. – 2017. – 336 с.
7. Тельнов, Ю. Ф., Федоров, И. Г. Инжиниринг предприятия и управление бизнес-процессами. Методология и технология: учеб. пособие / Ю. Ф. Тельнов, И. Г. Федоров. – М.: ЮНИТИ-ДАНА. – 2015. – 207 с.
8. Фаулер, М. NoSQL. Новая методология разработки нереляционных баз данных / М. Фаулер, П. Дж. Садаладж. – М.: Вильямс. – 2017. – 192 с.
9. TOGAF Version 9.1. [Электронный ресурс]. – Режим доступа: <https://www.opengroup.org/togaf/> (дата обращения: 02.07.2017).

References

1. Prikaz Rosstata ot 20.08.2012 № 454 (poslednjaja redakcija) «Ob utverzhdenii Metodologicheskikh rekomendacij po organizacii nabljudenija za potrebitel'skimi cenami i tarifami na otdel'nye vidy uslug» [Order of Rosstat dated 20.08.2012 N 454 (last revised) «On approval of Methodological recommendations on the organization of monitoring of consumer prices and tariffs for certain types of services»]. Available at: Spravochnaja pravovaja sistema «Konsul'tantPljus» [ConsultantPlus legal reference system] (accessed 15.01.2018).

2. Zaramenskih E. P. Osnovy biznes-informatiki [*Basics of business informatics*]. Moscow, Jurajt Publ., 2016. 407 p.
3. Kleppman M. Vysokonagruzhennye prilozhenija. Programmirovanie, masshtabirovanie, podderzhka [*Highly loaded applications. Programming, scaling, support*]. Saint Petersburg, Piter Publ., 2018. 740 p.
4. Marc N., Uorren Dzh. Bol'shie dannye. Principy i praktika postroenija masshtabiruemyh sistem obrabotki dannyh v real'nom vremeni [*Big Data. Principles and practice of building scalable data processing systems in real time*]. Moscow, Vil'jams Publ., 2016. 356 p.
5. Parfenov Ju. Proektirovanie reljacionnyh hranilishh dannyh. Uchebnoe posobie [*Designing of relational data warehouses. Tutorial*]. Moscow, Jurajt Publ., 2017. 122 p.
6. Silen D., Mejsman A., Ali M. Osnovy Data Science i Big Data. Python i nauka o dannyh [*Basics of Data Science and Big Data. Python and the science of data*]. Saint Petersburg, Piter Publ., 2017. 336 p.
7. Tel'nov Ju. F., Fedorov I. G. Inzhiniring predprijatija i upravlenie biznes-processami. Metodologija i tehnologija: ucheb. Posobie [*Enterprise engineering and business process management. Methodology and technology: tutorial*]. Moscow, JuNITI-DANA Publ., 2015. 207 p.
8. Fauler M., Sadaladzh P. Dzh. NoSQL. Novaja metodologija razrabotki nereljacionnyh baz dannyh [*NoSQL. New methodology for developing non-relational databases*]. Moscow, Vil'jams Publ., 2017. 192 p.
9. TOGAF Version 9.1. Available at: <https://www.opengroup.org/togaf/> (accessed 02.07.2017).