ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

УДК 004.891

ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА EcoHealth ДЛЯ ХРАНЕНИЯ И АНАЛИЗА СТРУКТУРИРОВАННЫХ И НЕСТРУКТУРИРОВАННЫХ БОЛЬШИХ ДАННЫХ

И. Ю. Петрова*, С. В. Горянин**

**Астраханский государственный архитектурно-строительный университет*

***Астраханский государственный университет*

В настоящее время информационно-аналитические системы применяются в различных областях. В данной статье рассмотрена разработка архитектуры системы, архитектуры хранения данных и ее применение в сфере экологии и здравоохранения. Был осуществлен анализ существующих информационно-аналитических систем, их возможностей. На основе проведенного исследования автором предлагается создать информационно-аналитическую систему для корреляционного анализа структурированных и неструктурированных больших данных о загрязнении окружающей среды и состояния здоровья населения.

Ключевые слова: *информационно-аналитическая система, большие данные, корреляционный анализ, искусственный интеллект. экология, здравоохранение.*

INFORMATION-ANALYTICAL SYSTEM EcoHealth FOR STORAGE AND ANALYSIS STRUCTURED AND NONSTRUCTURED BIG DATA

I. Yu. Petrova*, S. V. Gorianin**

**Astrakhan State University of Architecture and Civil Engineering*

***Astrakhan State University*

Today, information-analytical systems are used in various fields. This article examines the development of the architecture of the system, the architecture of data storage and application in the field of ecology and health. An analysis of existing information-analytical systems and their capabilities was carried out. Based on the study, the author proposes to create an information-analytical system for the correlation analysis of structured and unstructured Big Data on environmental pollution and health of the population.

Keywords*: information-analytical system, Big Data, correlation analysis, artificial Intelligence, ecology, health care.*

Введение

Основным механизмом принятия решений, направленных на улучшение качества воздуха и снижение риска возникновения патологических состояний у человека, является проведение анализа загрязнения воздуха и здоровья граждан в регионе. Оценка влияния экологических факторов на здоровье человека часто основывается на математическом моделировании причинно-следственных связей между различными показателями окружающей среды и изменением физиологических показателей населения в конкретных временных и территориальных условиях.

В настоящее время проводится большое количество исследований воздействия загрязнений воздуха на здоровье граждан. Изучение ситуации в различных регионах показало корреляцию в развитии патологий органов дыхания с долгосрочным воздействием взвешенных частиц, оксида серы [1–8], твердых частиц [9–11], черного дыма [12] и оксида азота [8]. Кроме того, исследования случаев госпитализации и летального исхода указывают на связь краткосрочного и долгосрочного воздействия загрязнителей воздуха с развитием заболеваний не только органов дыхания, но и сердечно-сосудистой системы [13–20].

Таким образом, важно изучить взаимосвязь между уровнем загрязнения воздуха и здоровьем человека, в частности с таким симптомом, как диспноэ (одышка) – патологическое состояние, при котором наблюдается нарушение частоты и глубины дыхания, а также возникает чувство нехватки воздуха [1–2]. В данной статье приводятся результаты исследования этой проблемы в городе Ницце (Франция). Для обработки и анализа структурированных и неструктурированных данных, учитывавшихся при проведении этого исследования, был разработан программный продукт с использованием подхода Big Data, что позволило повысить эффективность анализа данных и визуализировать его результаты.

Большие данные (Big Data) – совокупность подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объемов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста [20].

Существуют различные системы анализа данных о загрязнении воздуха и медицинских данных. AirPaca – open-source-проект, занимающийся сбором и анализом данных о загрязнении воздуха в городе Ницце; OpenAir – британский

open-source-проект, посвященный сбору и анализу данных о загрязнении воздуха на территории страны; OHDSI, EasyMedStat, REMMIT – проекты, занимающиеся анализом данных о пациентах, страдающих общей патологией; FreeMED – open-source-проект, цель которого – анализом данных о пациентах, страдающих общими заболеваниями. Сравнительная характеристика указанных выше систем сведена в таблицу 1.

Таблица 1

Сравнительная характеристика систем

Проект	*Критерий*				
	Анализ медицинских данных	*Анализ данных о загрязнении*	*Open-source решение*	*Применение технологий Big Data*	*Анализ в реальном времени*
AirPaca		+	+	+	+
OpenAir		+	+	+	
OHDSI	+			+	
EasyMedStats	+			+	
FreeMED	+		+	+	
REMITT	+			+	

Информационно-аналитическая система для корреляционного анализа краткосрочного и долгосрочного воздействия загрязнения воздуха на здоровье населения

При создании информационно-аналитической системы для корреляционного анализа краткосрочного и долгосрочного воздействия загрязнения воздуха на здоровье населения (EcoHealth) были рассмотрены следующие альтернативные виды архитектуры хранения данных:

1. Для хранения неструктурированных данных о загрязнении воздуха изучались три альтернативных варианта:

а) использование Oracle NoSQL DB (рис. 1);

б) использование MongoDB (рис. 2);

в) использование HDFS (рис. 3).

2. Для хранения структурированных данных о пациентах использовались локальные таблицы Oracle SQL Database 12c.

Были проведены испытания трех вариантов БД для хранения неструктурированных данных о загрязнении воздуха в г. Ницце за период с 2014 по 2016 г. Общий объем данных – 90 Мб. Описание данных о пациентах сведено в таблицу 2, описание данных о загрязнении воздуха – в таблицу 3.

Таблица 2

Описание данных о заболеваемости в городе Ницце за период с 2014 по 2016 г.

Поле	*Описание*
Gender	Пол пациента
Age	Возраст пациента
Address	Адрес проживания пациента
Postal code	Почтовый индекс
Ville	Город проживания пациента
Admission	Дата обращения за медицинской помощью
Sortie	Дата выписки из лечебного учреждения
Examen	Дата проведения диагностических и лечебных мероприятий
Categorie de Recours	Группа препаратов, которые были назначены пациенту
Libelle de Recours	Код выписки пациента
Code de Recours	Предварительный диагноз; код заболевания согласно МКБ-10 (Международной классификации болезней 10 пересмотра)
Libelle gravite	Код степени отягощенности заболевания пациента
Libelle CCMU	Код медицинского страхования пациента
Destination Confirmee	Лечебное учреждение, оказавшее медицинскую помощь пациенту
Type de sortie	Код выписки из лечебного учреждения
Diag1 – diag10	Окончательный диагноз; код заболевания согласно МКБ-10

Таблица 3

Описание данных о загрязнении воздуха г. Ниццы за период с 2014 по 2016 г.

Поле	*Описание*
Station	Название станции, где расположен сенсор измерения уровня загрязнения воздуха
Polluant	Химическая формула загрязнителя
Mesure	Полное название загрязнителя
Unité	Единицы измерения (μg/m^3)
Date	Дата измерения объема содержания загрязнителя в воздухе
Value	Объем содержания загрязнителя в воздухе

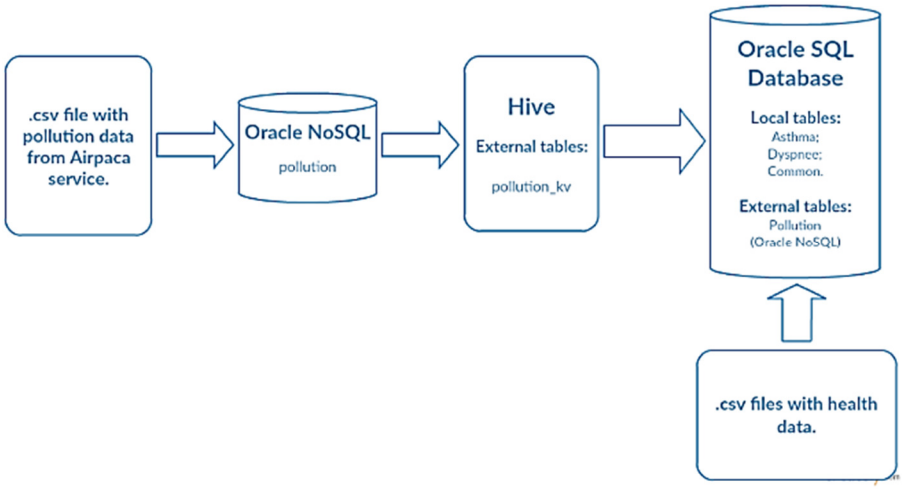

Рис. 1. Использование Oracle NoSQL DB

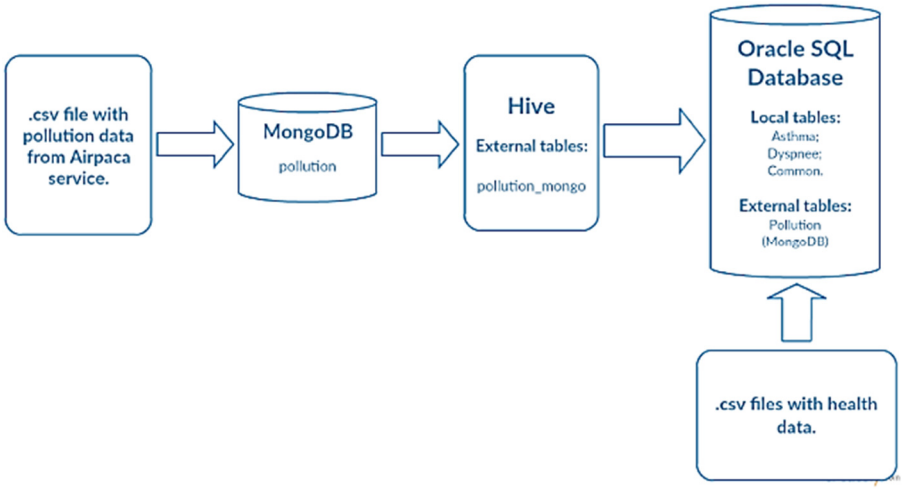

Рис. 2. Использование MongoDB

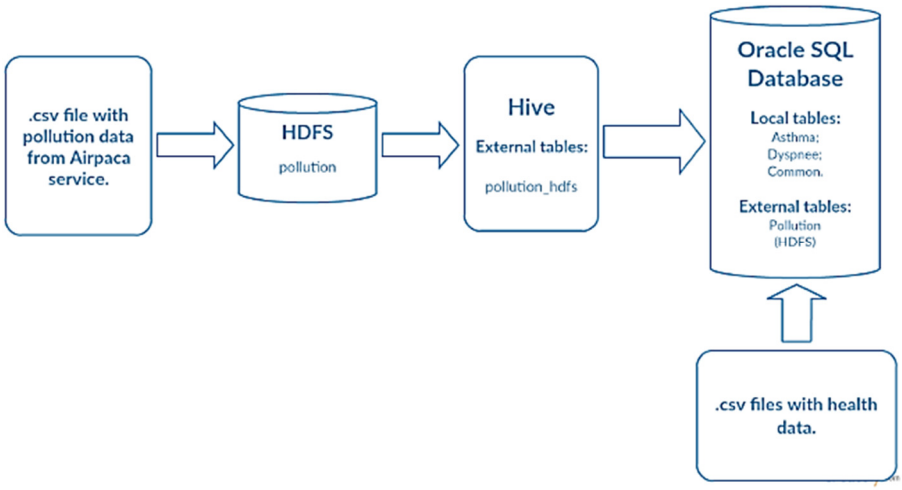

Рис. 3. Использование HDFS

На рис. 4 представлена UML-диаграмма классов. Отношение – один ко многим. С помощью данной схемы классов произведена визуализация результатов анализа данных с использованием Google Maps API (рис. 5) и графиков Prime Faces (рис. 6.)

На карте представлены: название станции, где расположен сенсор сбора данных о загрязнении воздуха; количество пациентов, проживающих на территории станции; название химического вещества и средний объем данного вещества.

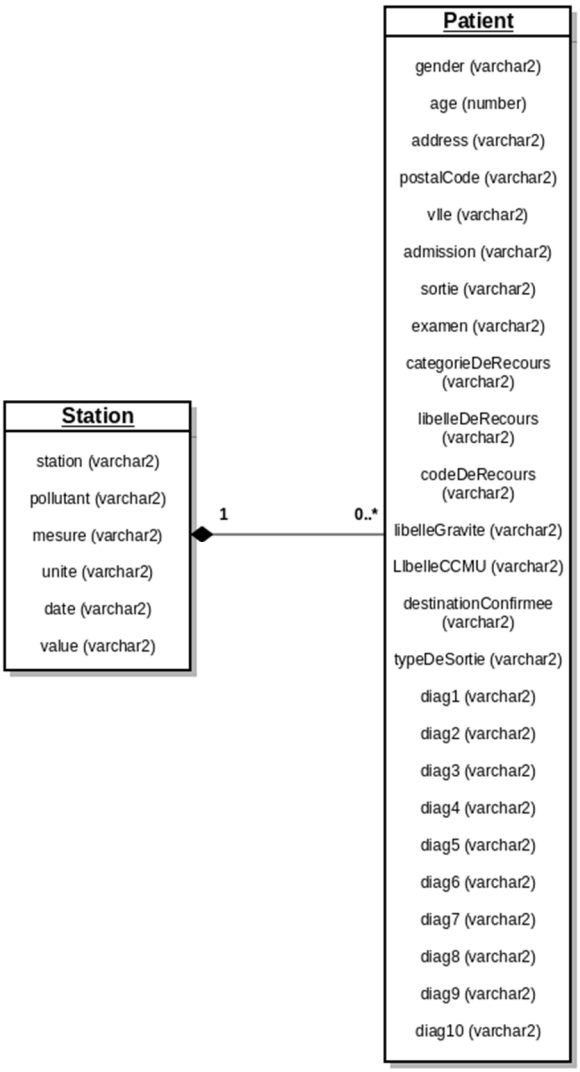

Рис. 4. UML-диаграмма классов

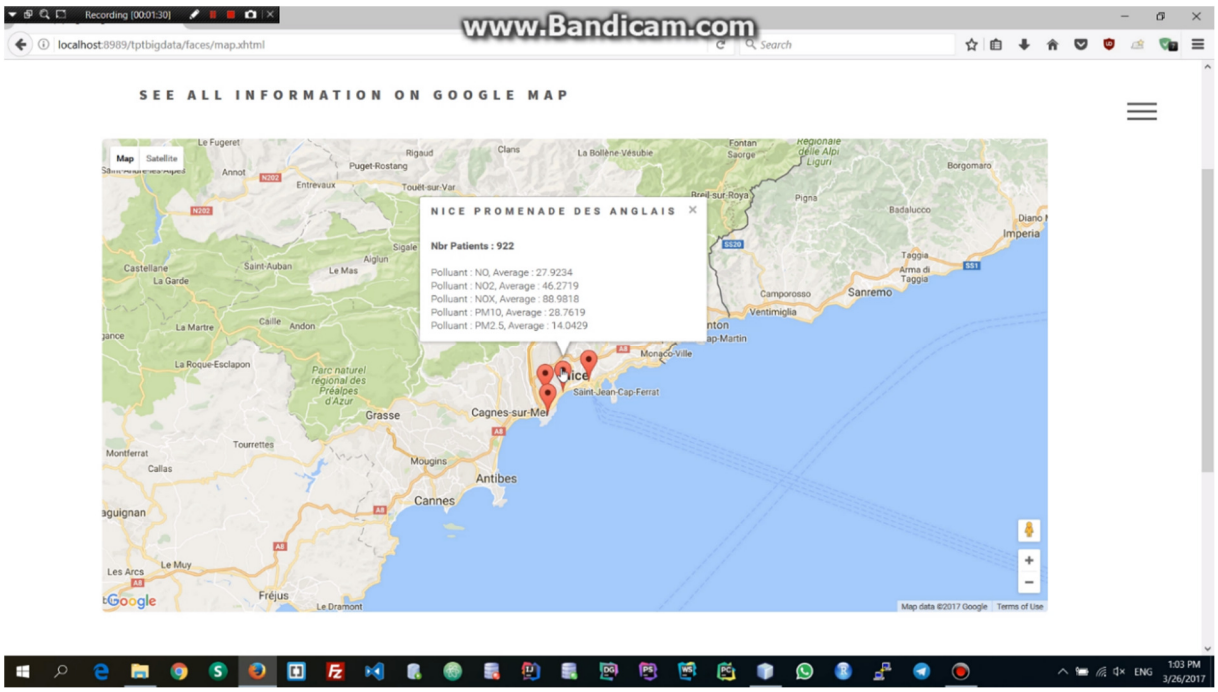

Рис. 5. Визуализация данных с помощью Google Maps API

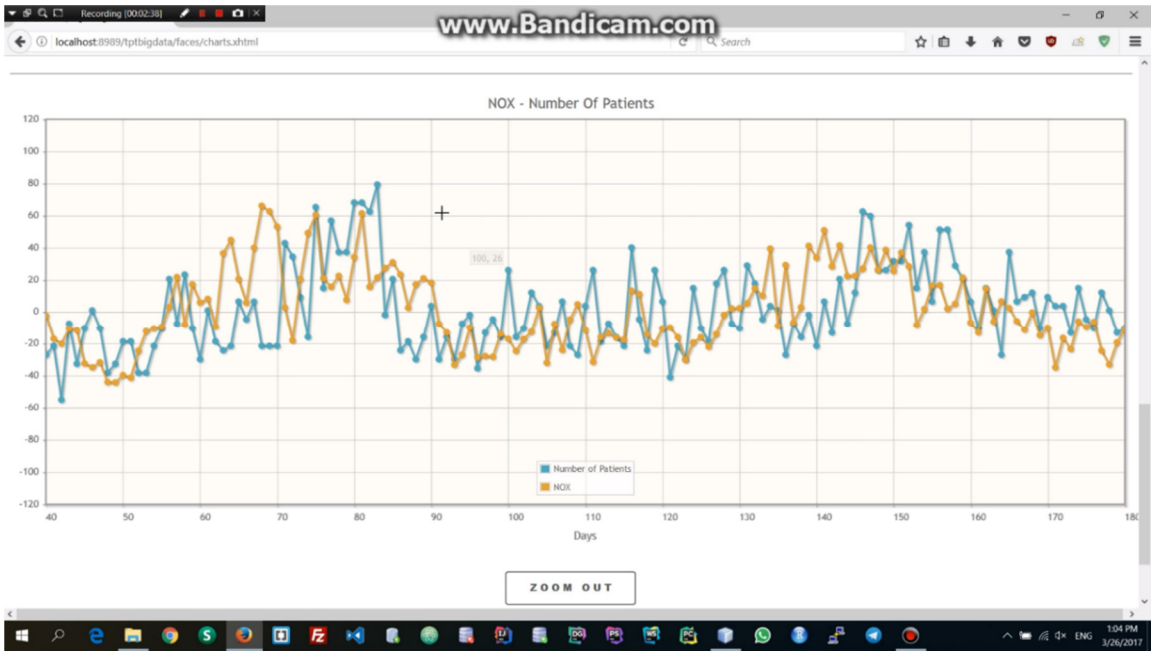

Рис. 6. Визуализация данных с помощью Prime Faces Charts

Указанные выше базы данных были выбраны по следующим причинам:

1. HDFS и Oracle NoSQL DB интегрированы в систему Oracle Big Data Approach.
2. HDFS предоставляет быстрый и удобный импорт файлов.csv с данными.
3. Oracle NoSQL DB, MongoDB и HDFS являются легконастраиваемыми для создания внешних таблиц на стороне Apache Hadoop Hive.
4. Импорт файлов.csv с данными в Oracle NoSQL DB и MongoDB также легок и прост.

Поскольку в результате испытаний БД HDFS показала самое высокое быстродействие выбора данных (табл. 4), было принято решение о ее использовании для хранения неструктурированных данных о загрязнении воздуха.

Таблица 4

Результаты тестирования скорости выполнения запросов

БД	*Oracle NoSQL DB*	*HDFS*	*MongoDB*
Время (сек.)	0,83	0,257	0,57

В результате была предложена архитектура автоматизированной системы для корреляционного анализа краткосрочного и долгосрочного воздействия загрязнения воздуха на здоровье населения, показанная на рис. 7.

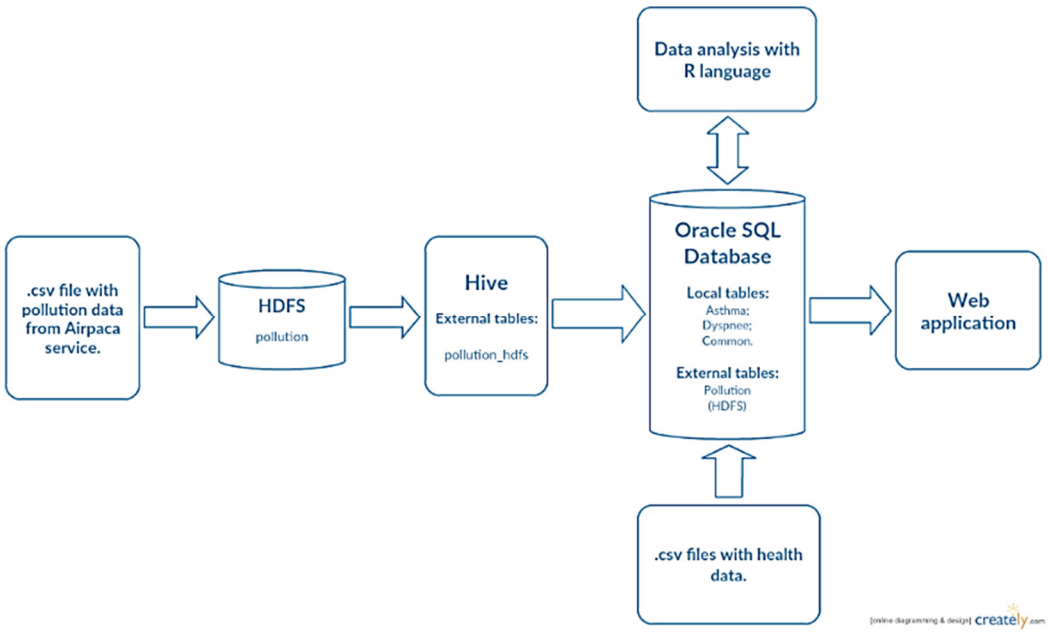

Рис. 7. Архитектура автоматизированной системы

Данную архитектуру можно применять для работы с другими типа данных, поскольку в качестве «универсального моста» между неструктурированными и структурированными данными используется Apache Hadoop Hive (система управления базами данных на основе платформы Hadoop, дающая возможность выполнять запросы, агрегировать и анализировать данные, хранящиеся в Hadoop), позволяющий создавать внешние таблицы с неструктурированными данными на стороне Oracle SQL Database. Эта возможность открывает перспективы использования всей мощи языка SQL для работы с различными типами данных.

Система EcoHealth уже используется организацией IMREDD (г. Ницца, Франция) для анализа данных о загрязнении воздуха и уровня заболеваемости. Полученные результаты не установили наличия корреляции между экологическим состоянием г. Ниццы и проявлением симптомов диспноэ. Однако работа продолжается, но уже с новыми данными.

Список литературы

1. Holland W. W., Reid D. D. The urban factor in chronic bronchitis. // Lancet. 1965, Feb. 27. Vol. 1 (7383). P. 445-448.
2. PAARC: Groupe Cooperative / Lelouche JPollution atmosphérique et affections respiratoires chroniques ou à répétition // Bull. Eur. Physiopathol. Respir. 1982. Vol. 18 (1). P. 87–116.
3. Schenker M. B., Samet J. M., Speizer F. E., Gruhl J., Batterman S. Health effects of air pollution due to coal combustion in the Chestnut Ridge region of Pennsylvania: results of cross-sectional analysis in adults // Arch. Environ. Health. 1983, Nov.-Dec. Vol. 38 (6). P. 325–330.
4. Euler G. L., Abbey D. E., Magie A. R., Hodlkin J. E. Chronic obstructive pulmonary disease symptom effects of long term cumulative exposure to ambient levels of total suspended particulates and sulfur dioxide in California Seventh-Day Adventist residents // Arch. Environ. Health. 1987, Jul.-Aug. Vol. 42 (4). P. 213–222.
5. Portney P., Mullahy J. Urban air quality and respiratory disease // Reg. Sci. Urban Econ. 1990. Vol. 20. P. 407–418.
6. Schwartz J. Particulate air pollution and chronic respiratory disease // Environ. Res. 1993. Vol. 62. P. 7–13.
7. Forsberg B., Stjernberg N., Wall S. Prevalence of respiratory and hyperreactivity symptoms in relation to levels of criteria air pollutants in Sweden // Eur. J. Public Health. 1997, Jun. 7. Vol. 314 (7095). P. 291–296.
8. Abbey D. E., Lebowitz M. D., Mills P. K., Petersen F. F., Beeson W. L., Burchette R. J. Long-term ambient concentrations of particulates and oxidants and development of chronic disease in a cohort of nonsmoking California residents // Inhal. Toxicol. 1995, Jan. Vol. 7 (1). P. 21–34.
9. Abbey D. E., Ostro B. E., Petersen F., Burchette R. J. Chronic respiratory symptoms associated with estimated long-term ambient concentrations of fine particulates less than 2.5 microns in aerodynamic diameter (PM2.5) and other air pollutants // J. Exp. Anal. Environ. Epidemiol. 1995, Apr.-Jun. Vol. 5 (2). P. 137–159.
10. Abbey D. E., Hwang B. L., Burchette R. J. Estimated long term ambient concentrations of PM10 and development of respiratory symptoms in a nonsmoking population // Arch. Environ. Health. 1995. Mar.-Apr. Vol. 50 (2). P. 139–150.
11. Scarlett J. F., Griffiths J. M., Strachan D. P., Anderson H. R. Effect of ambient levels of smoke and sulphur dioxide on the health of a national sample of 23-year-old subjects in 1981 // Thorax. 1995. Vol. 50. P. 764–768.
12. Schwartz J., Dockery D. W. Increased mortality in Philadelphia associated with daily air pollution concentrations // Am. Rev. Respir. Dis. 1992, Mar. Vol. 145 (3). P. 954–960.
13. Spix C., Heinrich J., Dockery D., Schwartz J., Volksch G., Schwinkowski K., Collen C., Wichmann H. E. Air pollution and daily mortality in Erfurt, East Germany, 1980–1989 // Environ. Health Perspect. 1993, Nov. Vol. 101 (6). P. 518–526.
14. Dockery D., Pope A., Xu X., Spengler J. D., Ware J. D., Fay M. E., Ferris B. J., Speizer F. E. An association between air pollution and mortality in six U.S. cities // N. Engl. J. Med. 1993, Dec. 9. Vol. 329 (24). P. 1753–1759.
15. Touloumi G., Pocock S. J., Katsouyanni K., Trichopoulos D. Short-term effects of air pollution on daily mortality in Athens – a time-series analysis // Int. J. Epidemiol. 1994. Vol. 23. P. 957–967.
16. Schwartz J. Air pollution and daily mortality: a review and meta-analysis // Environ. Res. 1994. Vol. 64. P. 36–52.
17. Pope A., Thun M., Namboodiri M., Dockery H. D. W., Evans J. S., Speizer F. E., Heath C. W. Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults // Am. J. Respir. Crit. Care Med. 1995, Mar. Vol. 151 (3, Pt 1). P. 669–674.
18. Schwartz J., Morris R. Air pollution and hospital admissions for cardiovascular disease in Detroit, Michigan // Am. J. Epidemiol. 1995, Jul. 1. Vol. 142 (1). P. 23–25.
19. Burnett R., Dales R., Krewski D., Vincent R., Dann T., Brook J. Associations between ambient particulate sulfate and admissions to Ontario Hospitals for cardiac and respiratory diseases // Am. J. Epidemiol. 1995, Jul. 1. Vol. 142 (1). P. 15–22.
20. Hey T., Tansley S., Tolle K. The Fourth Paradigm – Data-Inetnsive Scientific Discovery // Microsoft Research. 2009. P. 3–99.

Ссылка для цитирования:
Петрова И. Ю., Горянин С. В. Информационно-аналитическая система EcoHealth для хранения и анализа структурированных и неструктурированных больших данных // Инженерно-строительный вестник Прикаспия : научно-технический журнал / Астраханский государственный архитектурно-строительный университет. Астрахань : ГАОУ АО ВО «АГАСУ», 2017. № 3 (21). С. 66–71.