2018 International Conference on Information Technology Systems and Innovation (ICITSI)
Bandung - Padang. October 22-25, 2018
ISBN: 978-1-5386-5692-1

Big Data Analytics Algorithm, Data Type and Tools in Smart City: A Systematic Literature Review

Hafid Yoza Putra
Faculty of Information Technology
Universitas Andalas
Padang, Indonesia
hafidyozap@fti.unand.ac.id

Hasdi Putra
Faculty of Information Technology
Universitas Andalas
Padang, Indonesia
hasdiputra@fti.unand.ac.id

Novianto Budi Kurniawan
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
noviantobudik@s.itb.ac.id

***Abstract*—The smart city generated rapidly huge of data. Data can analyze with big data analytics technology to give solution from past data in smart city problem and help better solution in decision making. This paper summarizes the existing condition of big data analytics in the smart city in term of the algorithm, data type and tools were built using systematic literature review (SLR) as the standard methodology used to solve any problem by tracing the result from the previous research. The problem in SLR called as research question (RQ). To achieve that goal, we define some RQs related to that scope and clarify each question by tracing previous research. The research paper from reputable journal databases such as IEEE Xplore, Scopus, ScienceDirect, and Springerlink. After synthesizing 15 articles, the results are: algorithm data mining like ANN, Markov, graph mining, etc. needs to improve. That algorithm not enough to handle high data volume, high variety and high velocity to store and processes data; main data type have big chance to give the solution in the smart city is social media. That data has the potential to help in decision making in the smart city problem; Hadoop is the top tool to store and analyze data with high-performance, stable, reliable computing for the different type of data. Combination Hadoop with spark give less overhead to handle the high velocity and volume of data.**

***Keywords*—Big Data Analytics, Smart City, SLR, Prisma**

I. Introduction

Big data have to change how we work and live today [1]. The implication of implementing big data for work, like in Smart City, has impacted organizations operate activity. The organizational change or transformation is needed to support and take advantages of the big data opportunities. The organizational should be perfecting old roles, and introducing new roles, creating opportunities and anxiety for individuals and organizations. Research about big data became popular today, combined with Smart City make that research can be the solution in providing Smart City service solution.

Big data in Smart City generated from any sensor that allowing the regular collection of huge amounts of data. Six pillars of the Smart City like "Smart Economy," "Smart People," "Smart Environment," "Smart Mobility," "Smart Living," and "Smart Governance"[2]. Big data required technologies to capture, storage, management, and analysis the huge amount of data with more complexity. It characterized by its size and variety[3]. Many cities in the world that share a huge amount of raw open data in the electronic format that can read, share and reused online. A city can be said to be smart, if the city using information technology for improves the quality of services to the public, data communication, reduction in consumption of energy sources, involves interactive applications and active services with people. Domains that are considered smart city technologies contains traffic, health care, government services, water and waste, transportation management, and energy [4].

Systematic Literature Review (SLR) used to comprehend the concept and implementation of big data analytics and smart city. SLR became the standard methodology used to find solutions by tracing the result of previous research. The problem encountered in the SLR is called research question (RQ). Some of the reason for use SLR are summarizing existing research results, identify gaps in current research or provide frameworks for specific research area. To get a comprehensive result, we explored many kinds of literature published in the popular journals database, i.e., IEEExplore, ScienceDirect, Scopus, and SpringerLink from 2013 to 2018.

This paper is arranged as follows based on the research[14]. Chapter 2 describes the methodology SLR in general. Chapter 3 describes the stage used in SLR. Chapter 4 presents the outcome of the research question and discussion. Chapter 5 summary of SLR result and future work.

II. Methodology

Prisma is a part of SLR methodology. SLR is a literature review of several questions that have been formulated using explicit methods to identify, choose, and critically assess several studies then collect and analyze data from the research belonging in the review [19]. Prisma consists of four-phase to examine relevant research that is identification, screening, eligibility and included[19]. The Prisma help authors to improve reporting of systematic reviews and meta-analyses[20]. Prisma flow diagram can help to generate a flow diagram to identification, screening, eligibility, and included articles. Prisma focuses on ensuring the transparent and complete reporting of a systematic review.

The problem in Prisma is prepared and more specific inquisitive statement of the topic under study with Research Question (RQ). RQ is a clear statement about an area that is the focus of attention or questions that exist in scientific literature, theory, or practice that shows the need for understanding and investigation [21]. The result of Prisma is the answer of each Research Question (RQ) and explanation of it. The result statistically used for meta-analyses.

III. PROPOSE METHOD

A systematic literature review as the process of identifying, assessing and interpreting all research results to provide answers to research question consists of several activities, namely: specifying the research questions, selecting studies, extracting required data, synthesizing data and describing the result. The research question defined in this study is shown in Table I.

TABLE I. RESEARCH QUESTION

ID	Research Question	Motivation
RQ1	What algorithms are used in big data analytics of the smart city?	Identify the variety of algorithm used big data analytics in the smart city
RQ2	What type of data big data analytics of the smart city?	Identify the type of data commonly used big data analytics in the smart city
RQ3	What tools to store and process data in big data analytics of the smart city?	Identify tool that helps to store data in storage and process the data in big data analytics in a smart city.

To answer each research question, we tracked published research results in several popular journals databases using specific search string. The search string used in finding the appropriate literature is ("big data analytics") AND ("smart city") and result in proper study findings according to the string as shown in Table II.

TABLE II. RELATED STUDY FINDING RESULT

Database Journal	Article founds
IEEE Xplore	39
Scopus	87
ScienceDirect	50
SpringerLink	39
Total	215

Then we applied the inclusion and exclusion criteria to select the appropriate candidate for the article will be explored further. Prisma flow diagram for this research identified in Figure 1.

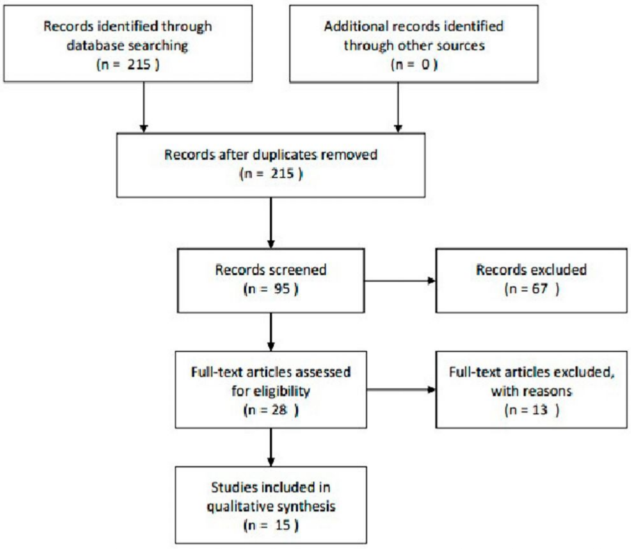
Records identified through database searching (n = 215)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 215)
Records screened (n = 95)
Records excluded (n = 67)
Full-text articles assessed for eligibility (n = 28)
Full-text articles excluded, with reasons (n = 13)
Studies included in qualitative synthesis (n = 15)

Figure 1 Big Data Analytics in Smart City Prisma flow diagram [20]

Total of the article from the search result in the database journal is 215. Then the articles are analyzed based on Figure 1 to identification, screening, eligibility, and included criteria. There are some of the article contexts about "Big data analytics" only. Another context about "Smart city" only. Prisma result 28 articles eligibility about "Big data analytics" and "Smart city."

TABLE III. INCLUSION AND EXCLUSION CRITERIA

Criteria	
Inclusion criteria	- Articles in peer reviewed papers - The article discusses "big data analytics" in "the smart city." - Articles were written in English - The article is open access
Exclusion criteria	- Books, book title and theses - Non-peer-reviewed research articles and white paper - Editorial, abstract or short paper (less than four pages) - Articles use the smart city as a case study

After applying inclusion and exclusion criteria in Table III, we get 15 relevant articles that become the primary reference in completing the SLR. The last activities conducted in SLR are synthesizing and describing the result as defined in chapter 4.

IV. RESEARCH RESULT AND DISCUSSION

This chapter details or describes every research question defined in chapter 3, namely: algorithm of big data analytics in smart city, the data type of big data analytics in smart city, and tools of big data analytics in smart city.

A. The algorithm of Big Data Analytics in Smart City

Big data analytics is considered to bet he ideal first step towards a smarter city. It assures flexible and real-time data processing followed by quick decision procedures [5]. Big data analytics require the algorithm to process data. The algorithm has been developed to merge a big number and type of sensor and link to certain datasets to do deeper computational and analytical solutions.

Learning algorithm consists of supervised and unsupervised. The common processes usings supervised learning algorithms for context identification include various steps [6]:

1. To obtains sensor data about urban systems that representative involved context features including related labeled annotations.
2. To assign the input data features and the representation.
3. To assembly data from many data source and turn them into the application-dependent.
4. To split the data into two part: a training set and a test set.
5. To train the identification algorithm on the training set.
6. To verify the performance of classification trained algorithm on the test set.
7. To apply the best performance algorithm in the context recognition.

The common process of the unsupervised learning algorithm for context identification includes several steps [6] :

1. To get unlabeled data from the sensor.
2. To assemble and change the sensor data into several features.
3. The data is then modeled using density estimation or clustering method.

To identify the algorithm of big data analytics in smart city, some researchers [7], [8], [9], [10], [11], [22] have expressed their opinions. The algorithm of big data analytics in smart city are :

1. Markov models to integrate big data with transport sharing to get better efficiencies to meet the demand for future city services[7].
2. Block-level background modeling (BBM) algorithm and Surveillance Rate–Distortion Optimization (SRDO) algorithm are developed to improve video coding performance. BBM and SRDO algorithm can significantly improve the performance of the compression, which could support a variety of video applications in the smart cities[8].
3. Artificial Neural Network (ANN) and Fuzzy algorithm were used widely for forecastings like water demand, quality monitoring and anomaly detection[9]. Fuzzy also used to predict the health of a machine like components of diesel engine [22].
4. Combination graph theory algorithm with spanning tree based greed algorithm provides a robust method with feature like only applied on two-way roads, guaranteed to traverse each road segment only twice, and runtime on the order of seconds. An evolutionary algorithm is a good option to accomplish the problem of combinatorial optimization in logistics with more optimal solutions in a comparatively short time[10].
5. Data distribution algorithm has result fog nodes are resistant to problems of back-haul connectivity and could transmit data to the cloud location event with problems of severe connectivity [11].
6. Water demand forecasting was done with several forms of the algorithm including EPC, GP, EKF, DGBN, SARIMA, ARIMA, HW, BN, TLBO, LSSVM, EMD and Adaptive[9].

The algorithm of data mining not feasible to handle large data because it is designed to deal with well-defined and limited datasets. Datasets of big data have characteristic high-variety, high-volume, and high-velocity. Existing data mining algorithm needs to be fixed in how to perform huge volumes, various type of data and time constraints in data processing[13].

B. Data Type of Big Data Analytics in Smart City

The smart city has a wide variety of data type. To recognize the data type of big data analytics in smart city, some researchers have expressed their opinions. The data type of big data analytics in smart city presented on Table IV.

Big data and intelligence technologies can store, manage, processes and analyze huge amount data in social media like Twitter. The result is obtained third top tweets words about that is being talked about and the location of each tweet in an area[12]. Social media to be an essential data source in smart cities. It makes we know communication between governments, citizens, and businesses.

TABLE IV. DATA TYPE IN BIG DATA ANALYTICS OF SMART CITY

No.	Data Type	References
1	Social media	[12],[13], [23]
2	Health-related terms	[2]
3	Transport and distribution system	[7]
4	Video	[8]
5	Water	[9]
6	Spatial logistic	[10]
7	Traffic logs	[11]
8	Urban environment	[15]

Health-related terms in google trends data used to know measure public interest in health. Smart health as a part of smart cities needs predicting and analyze the public interest and behavior. Data from google trends as datasets combined with big data technology, the result of analyzed have been guaranteed effective in the past so that it can help predict and nowcasting[2].

Data about transport and distribution system used to help decision making to capacity sharing. The result shows a high request for transport. Big data analytics make it is efficient increasingly than individual health care transportation schemes[7]. Video data in the smart city generate rapidly and need more storage. The video needs to be compressed and has high performance to compressed. The results require less storage, strong coding efficiency and high transmission performance to store data[8].

The demand of the public to consume water can measure with big data analytics. Data water collected then analyzed, and the result used forecasting of the water demand in smart water management [9].

Data of spatial logistic used to provide an efficient solution to logistic problems[10]. The data get from routes that covering in an area. Data traffic logs get from Floating Car Data in one week. Data analyzed with learning method based on result data distribution of process paradigm fog computing in the fog nodes are resistance and can transmit data to cloud event in few connectivity issues[11].

Data selected urban environment indicators have analyzed. The data get from Bristol Open data. The existing data is used to measure the spread signs that have occurred over the years to assess positive and negative trends[15].

C. Tools of Big Data Analytics in Smart City

To identify tools of big data analytics in smart city, some researchers have expressed their opinion. The tools of big data analytics in smart city presented in Table V.

Among the leading platforms for big data storage, processing, and management include IBM infosphere, Hadoop map-reduce, spark, stratosphere and NoSQL database system management [11], [12], [13], [15], [17], and [18]. Based on Table V, Hadoop is the most widely used tools on big data analytics. Hadoop ma- reduce approach for distributing modeling and processing algorithms to deploy traffic analytics[11]. Hadoop used as storage space to store twitter data streaming [12],[18]. Standalone using Hadoop to store data, Hadoop has significant overhead when the works are submitted to the cluster[15]. Hadoop as reliable software with stable, high-performance, and reliable computing capabilities for the various type of data [13].

TABLE V. TOOLS IN BIG DATA ANALYTICS OF SMART CITY

No.	Tools	References
1	Hadoop	[13],[12],[11],[15],[18]
2	Spark	[12],[15]
3	Tableau	[12]
4	Google Trends	[2]
5	Mobile Edge Computing Framework	[4]
6	HPCC	[13]
7	Stratosphere	[13]
8	IBM Infosphere streams	[13]
9	Rapid Miner	[15]
10	R	[15]
11	Apache Mahout	[15]
12	NoSQL	[17]
13	MDM	[16]

Comparative Hadoop and Spark, Spark is faster and give rise to much less over-head[15]. Spark can use to analyze the data in collaboration with geocoding API to produce spatial data from geotagged Twitter[12]. Spark is an efficient step to calculate memory that has better speed, replication for fault tolerance and data parallelization allow reliability when processing large data and require a long time[13].

A tableau is a visualization tool, but the Tableau could be used for analysis as well. Tableau is only used for interpretation to show the distribution of location in the form of the maps[12]. The table also provides several charts to interpretation something interest. Combination Hadoop and RapidMiner to process data in a hybrid cloud. R tool also can be combined with Hadoop to process data statistically. R used customized python language, and it is little different with common python language.

Mobile edge computing framework providing an application that needs weak latency to mobile end users. Mobile devices with limited capabilities can gain access to resources that are stronger than devices around without incurring costs to reach the supply at its core[4].

Tools like Stratosphere, Hadoop MapReduce, NoSQL, IBM Infosphere Streams, and Spark run great on cluster system to fulfill the necessary of huge data applications in smart cities[17]. Big data need tools, that can handle all data type for structured and unstructured data with huge volume. Hadoop map-reduce has become a major primary data to storage and processing system that is given the scalability, simplicity, and smooth grain fault tolerance. NoSQL like Cassandra and Mongo DB have also become an option to store, to sort unstructured data and to cluster them with greater efficiency and scalability. MDM is a framework designed to help build the model of the smart city, which it can improve transparency, accuracy, governance and semantic of the transportation and building master[16].

Based on the result of answer research question about the algorithm of big data analytics in smart city, we can show that in the pie chart in Figure 2.

The fuzzy algorithm is the most used based on answer RQ 1. Fuzzy has to function to predict something that can implement in any sector. If we need to predict something from the data get from the smart city, we can use fuzzy. Algorithms that are suitable for spatial data are Graph and Spanning Tree. It could help to solve the logistic problem.

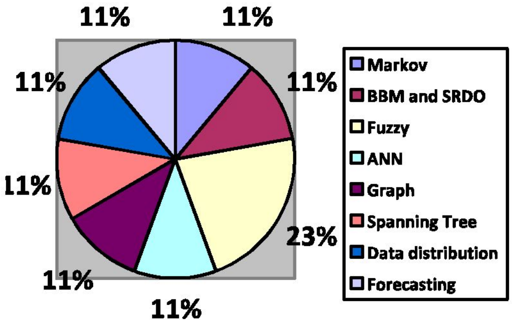

Figure 2 Statistical Algorithm based on RQ 1

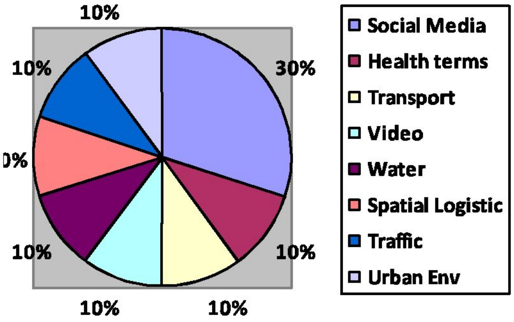

Figure 3 Statistical Data Type based on RQ 2

Answer of RQ 2 about data type shown in figure 3. Social media data became most data type used. Social media data generated with high volume in a short time. That may have spatial or nonspatial data like in twitter. Analyze social media provide insight about communication between people with the government about reality or hot issues in some period and location around the world.

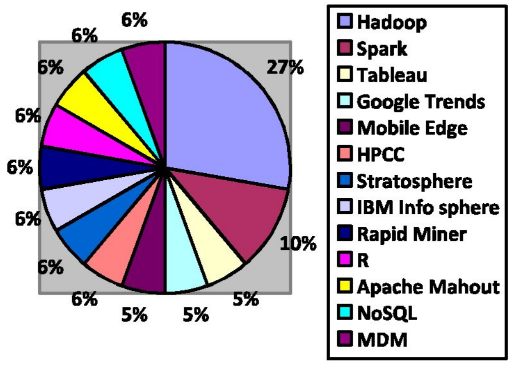

Figure 4. Statistical Tools based on RQ 2

Answer of RQ 3 about tools shown in figure 4. Hadoop tools most used to store, manage and process large data. Then spark is best combination with Hadoop to handle characteristic of big data including variety, volume, and velocity. The other tool to store data in big data is NoSQL but used often. Mobile edge computing, HPCC, Stratosphere, and IBM Infosphere also used to process big data, but the user of this tool still a little bit. Rapid Miner and R have the same function to predict something using an algorithm available in tools or extension.

V. Conclusion and Future Work

Research on the field of big data analytic in the smart city from 15 researchers have identified. Big data analytics technology to process datasets with large volume, wide variety and structured/unstructured from data that generate rapidly from the smart city can give a better solution in decision making. The smart city generates data from transactional data, GPS, cell phone, sensor, and video. The right process and method needed to get a reliable result.

The algorithm used in big data analytics like algorithm Markov, block-level background modeling (BBM), Artificial Neural Network (ANN) and fuzzy systems, graph theory algorithm, spanning tree based greed algorithm, data distribution, and forecasting algorithm. The most used is the fuzzy algorithm. That all algorithm not enough effective and fast to store and processes real-time data with huge volume. The algorithm needs to expand for better performance in big data analytics.

Data type most used based on this paper is social media. It has significant potential to analyzed which have a better solution in the smart city today. That dataset accessible to collected and can collect in high volume in a short time. Besides that, the powerful tool to storage, stable, high-performance, and reliable computing for various type of data. Hadoop map-reduce the best one to handle it. Combination Hadoop with Spark can give better performance and faster time to store data and less overhead.

Our future research to implement big data analytics with data type social media like Twitter or other. Hadoop will be the primary tool to store datasets and combined with Spark to get better performance in big data characteristic. Processed data visualized using Tableau to provide management insight about what happened. The fuzzy algorithm used to predict/forecasting what will happen.

Acknowledgment

This paper is supported by the Faculty of Information Technology, University of Andalas, Padang-Indonesia for publication.

References

[1] Okwechime, E., Duncan, P., & Edgar, D. (2017). Big data and smart cities: a public sector organizational learning perspective. *Information Systems and E-Business Management*, 1–25. https://doi.org/10.1007/s10257-017-0344-0

[2] Sampri, A., Mavragani, A., & Tsagarakis, K. P. (2016). Evaluating Google Trends as a Tool for Integrating the "Smart Health" Concept in the Smart Cities' Governance in the USA. *Procedia Engineering*, *162*, 585–592. https://doi.org/10.1016/j.proeng.2016.11.104

[3] Dwivedi, Y. K., Janssen, M., Slade, E. L., Rana, N. P., Weerakkody, V., Millard, J., … Snijders, D. (2017). Driving innovation through big open linked data (BOLD): Exploring antecedents using interpretive structural modeling. *Information Systems Frontiers*, *19*(2), 197–212. https://doi.org/10.1007/s10796-016-9675-5

[4] Quwaider, M., Al-Alyyoub, M., & Jararweh, Y. (2016). Cloud Support Data Management Infrastructure for Upcoming Smart Cities. *Procedia Computer Science*, *83*, 1232–1237. https://doi.org/10.1016/j.procs.2016.04.257

[5] Silva, B. N., Khan, M., & Han, K. (2017). Big data analytics embedded smart city architecture for performance enhancement through real-time data processing and decision-making. *Wireless Communications and Mobile Computing*, *2017*. https://doi.org/10.1155/2017/9429676

[6] Chen L, Nugent C. Ontology-based activity recognition in intelligent pervasive environment, Int J Web Inf Syst.2009;5(4):410-30

[7] Mehmood, R., & Graham, G. (2015). Big Data Logistics: A health-care Transport Capacity Sharing Model. *Procedia Computer Science*, *64*, 1107–1114. https://doi.org/10.1016/j.procs.2015.08.566

[8] Tian, L., Wang, H., Zhou, Y., & Peng, C. (2018). Video big data in smart city: Background construction and optimization for surveillance video processing. *Future Generation Computer Systems*, *86*, 1371–1382. https://doi.org/10.1016/j.future.2017.12.065

[9] Vijai, P., & Sivakumar, P. B. (2016). Design of IoT Systems and Analytics in the Context of Smart City Initiatives in India. *Procedia Computer Science*, *92*, 583–588. https://doi.org/10.1016/j.procs.2016.07.386

[10] Gutierrez, J. M., Jensen, M., & Riaz, T. (2016). Applied Graph Theory to Real Smart City Logistic Problems. *Procedia Computer Science*, *95*, 40–47. https://doi.org/10.1016/j.procs.2016.09.291

[11] Pérez, J. L., Gutierrez-Torre, A., Berral, J. L., & Carrera, D. (2018). A resilient and distributed near real-time traffic forecasting application for Fog computing environments. *Future Generation Computer Systems*, *87*, 198–212. https://doi.org/10.1016/j.future.2018.05.013

[12] Suma, S., Mehmood, R., Albugami, N., Katib, I., & Albeshri, A. (2017). Enabling Next Generation Logistics and Planning for Smarter Societies. *Procedia Computer Science*, *109*, 1122–1127. https://doi.org/10.1016/j.procs.2017.05.440

[13] Al Nuaimi, E., Al Neyadi, H., Mohamed, N., & Al-Jaroodi, J. (2015). Applications of big data to smart cities. *Journal of Internet Services and Applications*, *6*(1), 1–15. https://doi.org/10.1186/s13174-015-0041-5

[14] Muhamad, W., Kurniawan, N. B., Suhardi, S., & Yazid, S. (2018). Smart campus features, technologies, and applications: A systematic literature review. In *2017 International Conference on Information Technology Systems and Innovation, ICITSI 2017 - Proceedings* (Vol. 2018–Janua, pp. 384–391). https://doi.org/10.1109/ICITSI.2017.8267975

[15] Khan, Z., Anjum, A., Soomro, K., & Tahir, M. A. (2015). Towards cloud-based big data analytics for smart future cities. *Journal of Cloud Computing*, *4*(1). https://doi.org/10.1186/s13677-015-0026-8

[16] Ng, S. T., Xu, F. J., Yang, Y., & Lu, M. (2017). A Master Data Management Solution to Unlock the Value of Big Infrastructure Data for Smart, Sustainable and Resilient City Planning. *Procedia Engineering*, *196*(June), 939–947. https://doi.org/10.1016/j.proeng.2017.08.034

[17] Bibri, S. E., & Krogstie, J. (2017). The core enabling technologies of big data analytics and context-aware computing for smart sustainable cities: a review and synthesis. *Journal of Big Data*, *4*(1). https://doi.org/10.1186/s40537-017-0091-6

[18] Silva, B. N., Khan, M., & Han, K. (2017). Big data analytics embedded smart city architecture for performance enhancement through real-time data processing and decision-making. *Wireless Communications and Mobile Computing*, *2017*. https://doi.org/10.1155/2017/9429676

[19] Green, S., & Higgins, J. (2005). Cochrane handbook for systematic reviews of interventions

[20] Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

[21] Bryman, Alan. (2007). The Research Question in Social Research: What is its Role ?. International Journal of Social Research Methodology 10: 5-20

[22] Zhong, R. Y., Xu, X., Klotz, E., & Newman, S. T. (2017). Intelligent Manufacturing in the Context of Industry 4.0: A Review. *Engineering*, *3*(5), 616–630. https://doi.org/10.1016/J.ENG.2017.05.015

[23] Encalada, L., Boavida-Portugal, I., Ferreira, C. C., & Rocha, J. (2017). Identifying tourist places of interest based on digital imprints: Towards a sustainable smart City. *Sustainability (Switzerland)*, *9*(12). https://doi.org/10.3390/su9122317