Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Design and Implementation of Smart City Big Data Processing Platform Using Big Data Analytics for Decision Management System

[1]J Chandra Sekhar, Professor, Department of Computer Science and Engineering, NRI Institute of Technology, Visadala, Guntur, A.P, India.

[2]V Krishna Pratap, Associate Professor, Department of Computer Science and Engineering, NRI Institute of Technology, Visadala, Guntur, A.P, India.

Email ID's: jcsekhar9@gmail.com, pratapv9@gmail.com

Article Info
Page Number: 617-627
Publication Issue:
Vol. 69 No. 1 (2020)

Abstract: Interest in Smart Cities (SCs) and Big Data Analytics (BDA) has increased in recent years, revealing the bond between the two fields. An SC is characterized as a complex system of systems involving various stakeholders, from planners to citizens. Within the context of SCs, BDA offers potential as a data-driven decision-making enabler. Although there are abundant articles in the literature addressing BDA as a decision-making enabler in SCs, mainstream research addressing BDA and SCs focuses on either the technical aspects or smartening specific SC domains. A small fraction of these articles addresses the proposition of developing domain-independent BDA frameworks. Various components, i.e., smart transportation, smart community, smart healthcare, smart grid, etc. which are integrated within smart city architecture aims to enrich the Quality Of Life (QoL) of urban citizens. However, real-time processing requirements and exponential data growth withhold smart city realization. Therefore, herein we propose a Big Data Analytics (BDA)-embedded experimental architecture for smart cities. For the problem about the low shared efficiency and the poor integrating degree of the massive multivariate data, in this analysis, they described design scheme based architecture. And the real-time analysis system of real-time data and statistical analysis system of offline data are developed. Finally, the test results show the performance of Big Data processing platform. By analyzing authenticated datasets, they obtained the values required for city operation management. Throughput and processing time analysis performed with regard to existing works guarantee the performance superiority of the described work.

Article History
Article Received: 25 October 2020
Revised: 30 November 2020
Accepted: 15 December 2020

Keywords: Big Data Analytics, Big Data Smart City, Smart City Planning

I. INTRODUCTION

The evolution of the Internet of Things (IoT) infrastructure and the falling costs of technology are causing an impressive increase in the number of electronic devices with sensing capabilities pervaded in the urban environment, allowing to monitor temperature, traffic, air quality, flooding, among others. The increasing use of the Internet of Things (IoT) is making the implementation of smart cities becomes increasingly popular [13]. A city with sensors can get a variety of data and generate a huge amount of information that can be used both by inhabitants and administrators for decision making. Moreover, social networks can be combined with mobile devices and sensor networks in order to get contextual information

from users, supporting applications' awareness of their location, preferences and relationships. The processing and analysis of these data are fundamental to the implementation of smart city initiatives, since they provide a better understanding of what happens in cities. They make it possible to identify problems and their probable causes, supporting decision making. This impacts citizens quality of life.

Big Data processing is done in batches or in real-time. In batch processing, previously collected and stored data are processed, which may take hours if batches are large [9]. In real-time processing, the data are processed as it arrives at the application, generating results with low latency. Some applications require the combined use of batch and realtime processing. This can be useful, for example, in a traffic monitoring system: to identify and report real-time crashes quickly, as well as to predict the most dangerous areas and to avoid new situations of risk by querying historical information.

Furthermore, the interaction between Information and Communications Technology (ICT) devices and human can contribute to the vision of smart city, for example, intelligent transport system [14]. This large quantity and variety of data generated by smart city creates a big data.

In addition, data of utilities can be provided to integrate big data because their use can help in the resolution of urban problems. However, data provided by different sources may present a variety of format. To resolve this, it is necessary data integration tools. In other words, tools that transform the different data in a single format in order to allow storage in a single database.

Academics and practitioners tend to model SCs from the perspective of the smart domains constituting the city. However, the interrelationship between the underlying city domains is a crucial concept to realize city smartness. The interrelationship between the inter-domain systems is also an intuitively essential consideration. This integrated view of an SC implies cross-domain sharing of information. From this perspective, SC is viewed as a whole body of integrated systems or a system of systems. Whereas information and communication technology (ICT) and IS are the key enabling technologies in SCs [2], this integral and holistic view of SCs reflects its complexities of the corresponding ICT and IS implementations.

The direct consequence of the intensive diffusion of ICT in SC domains (e.g., internet of things [IoT], mobile applications, smart meters) is the generation of vast volumes of a mixture of structured, semi-structured, and unstructured digital data via machines, organizations, and people. This type of data is known as big data (BD). The analysis of these accumulated data (BDA) to extract latent insights and unknown information encourages academics and practitioners to support decision making in SCs. The connection between SCs and BDA has been covered extensively in the literature. Although there is an abundance of academic articles addressing the use of BDA in SCs, most research efforts focus on either the technological aspects, such as IoT implementations, or reaching *smartness* in specific SC domains (e.g., energy, environment, transportation). Only a small fraction of these articles addresses the proposition of domain-independent BDA frameworks that can support a broad

range of analytics in a multi-stakeholder environment, such as SCs. Taking into consideration the level of details and complexity of SC projects (either establishing a new city from scratch, such as Masdar in the UAE, or changing a legacy city into an SC, such as in Barcelona, Spain) in addition to the diversity of the stakeholders involved in these projects (from highly strategic decision makers to citizens and visitors), we can comprehend the need for resilient and efficient BDA mechanisms that can serve the requirements of various decision makers in SCs and at the same time allow the interchange of extracted analytics.

A city's service provision efficiency is highly correlated with data collection followed by data manipulation and decision generation. Expansion of Wireless Sensor Networks (WSN) increase the amount of urban data circulating across the network, which generates urban Big Data (UBD). In fact, the efficacy of smart city services is empowered by real-time processing of UBD. Conventional data processing mechanisms are unable to meet real-time processing demands with accelerating data growth. Conventional data processing mechanisms use statistical methods to analyze large datasets, in order to extract useful hidden data. Predictive analytics is a commonly used conventional data analyzing mechanism that is based on statistical methods. Predictive analytics can be further categorized with respect to the outcome variable [12]. However, the direct application of predictive analytics has been questioned by the domain experts due to statistical insignificance in large datasets, challenges in efficient computing, and distinctive Big Data (BD) characteristics, i.e., heterogeneity, noise accumulation, false positive correlation, and incidental endogeneity.

BDA-embedded smart city architecture that ensures reliable and efficient data management to derive real-time intelligent decisions. The architecture consolidates data aggregation, data manipulation, and service management tasks. The data management component is considered to be the brain of the architecture as it performs data filtration, analysis, processing, and storing of valuable data. Filtration mechanisms are introduced to the architecture, to further expedite processing and analysis. MapReduce on Hadoop is used for offline data processing and Apache Spark is used for online data processing. Intelligent agents and brokers of the architecture generate and transfer intelligent decisions to the service management layer, to formulate desired urban services. Experimentation results revealed that the occupied filtration technique has significantly improved the processing performance in terms of processing time and data throughput.

Moreover, performance superiority of selected processing mechanisms, namely Apache Spark for online streaming data processing and Hadoop MapReduce for offline batch processing, has been proven by the results. Furthermore, we compared processing time and throughput performance of the work with two previously reported BD analytical systems tested on smart city environments. Owing to performance improvement gained through integrated normalizing and filtering techniques.

II.LITERATURE SURVEY

Osman, A.M.S et.al [3] introduced a domain-independent BDA framework called SCDAP. SCDAP is a three-layer domain-independent end-to-end BDA framework for SCs. The logical design of SCDAP is based on six design principles: a layered design approach,

standardized data acquisition and access, realtime and historical data analytics, iterative and sequential data processing, extracted model management, and aggregation. The last two principles enable maintaining the models extracted from the data analysis. This feature has two benefits. First, creating a repository for the extracted models can be considered a form of knowledge memory for the SC. Second, this feature enables stakeholders and decision makers to share the analysis and knowledge results between them. This feature meets the goal of maintaining the cross-domain interrelationship of SC domains and, intuitively, inter-domain system interrelationship.

Psomakelis, E.; Aisopos, F.; Litke, A.; Tserpes, K.; Kardara, M.; Campo, P.M et.al [10] presented a generic four-tier BDA framework comprising the sensing hub, the storage hub, the processing hub, and the application. The work is a typical BDA framework identified some key challenges in leveraging BDA in SCs, they did not present how the proposed framework is different in handling these challenges. Furthermore, no real realization of the framework or use case was presented.

Iqbal, R.; Doctor, F.; More, B.; Mahmud, S.; Yousuf, U. et.al [1] presented a proposal for a six-layer BDA framework. The proposed framework's core data analysis engine utilizes two computational techniques: deep learning neural networks and fuzzy logic in data analytics. The utility of the proposed framework is demonstrated using the taxi demand prediction case study discussed. The classification results provided by the proposed framework contribute to optimizing the realtime distribution of taxis based on the predicted demand. This optimization directly affects other aspects, such as improving taxi availability, reducing waiting and journey times, and minimizing CO2 emissions.

Shah, S.A.; Seker, D.Z.; Rathore, M.M.; Hameed, S.; Ben Yahia, S.; Draheim, D et.al [4] introduced a general architecture based on IoT and BDA for disaster- resilient smart cities (DRSCs). The main feature of DRSC is providing disaster management with early warnings through collection, integration, and analysis of realtime and offline data from heterogeneous city data resources. Additionally, DRSC enables the prediction and monitoring of disaster situations. DRSC is implemented using a combination of Hadoop and Spark engines. It is evaluated in terms of processing time and throughput with simulated data generated from a) a fire dynamic simulator, b) gas sensors for pollution monitoring, c) road traffic simulators, and d) Twitter data (one-month crowdsourced data about disasters).

Balduini, M.; Brambilla, M.; Della Valle, E.; Marazzi, C.; Arabghalizi, T.; Rahdari, B.; Vescovi, M et.al [5] presented a domain-independent conceptual framework built on their previously proposed framework called Frame, Pixel, Place, and Event (FrAPPE). The idea behind FrAPPE is built on the digital image processing metaphor. Events are recorded in a time series of frames. Each frame includes information about the event location, in which the location is logically represented as hierarchical levels of grids and pixels. The main feature of FrAPPE is its ability to easily track and analyze a sequence of events in terms of time and location and then predict what potentially happens. The framework is implemented using the Hadoop ecosystem (e.g., HDFS, YARN, HIVE, P.I.G., Spark SQL, Spark Streaming, and SparkR). Its feasibility, generality, and effectiveness are demonstrated through multiple use

cases and examples taken from real-world requirements collected in various cities (e.g., Milano Design Week, Milano Fashion Week, and Milan Expo 2015).

D. Dissanayake and K. Jayasena et.al [6] described a platform which combines batch and real-time processing, being capable of analyzing large volumes of data from Internet of Things. The technology used for batch distribution is Hadoop Distributed File System (HDFS). For the real-time processing layer, the system uses Apache Storm, not benefiting from features of other frameworks.

H. Cho, H. Shiokawa, and H. Kitagawa et.al [11] framework integrates real-time and batch processing into a single system that abstracts the data and provides a programming model that uses a JSON-based dataflow algebra. For this, it extends Jaql, a language which provides several processing methods (e.g. *filter*, *join*, *sort*, and *group by*) converted to Hadoop and Spark. JSFlow includes operators for real-time processing in Jaql (e.g., *window*, *tostream* and *append*). A prototype was built to evaluate the framework using Apache Spark. However, the solution is still a prototype and was not evaluated with other frameworks besides Spark.

C. Misale, M. Drocco, M. Aldinucci, and G. Tremblay et al. [7] characterized the dataflow model used in Big Data frameworks from a more theoretical perspective. They also used the model to analyze some frameworks, such as Spark and Storm, and their user APIs. The basic building blocks in this interface are *activities*, *streams*, *data batches*, and *channels* for data transfers (between activities, from data sources to activities or from activities to sinks). An activity is a processing unit which receives input data and generates output data. However, they did not addressed the implementation of the theoretical model.

D. Crankshaw, X. Wang, G. Zhou, M. J. Franklin, J. E. Gonzalez, and I. Stoica, et al. [8] presented CLIPPER, a modular architecture that isolates user applications from the diversity of machine learning frameworks, offering a common interface (API) to access them. In CLIPPER, new frameworks can be added to the abstraction without changing the final user applications.

III. A Frame Work Of Design And Implementation Of Smart City Big Data Processing Platform Using Big Data Analytics For Decision Management System

In this section a frame work of design and implementation of smart city big data processing using big data analytics for decision management system is observed.

Raw data (sometimes called source data, atomic data or primary data) is data that has not been processed for use. A distinction is sometimes made between data and information to the effect that information is the end product of data processing.

As unstructured data are voluminous compared to the semi-structured data, they are processed and analyzed extensively prior to storing in HBase. To maintain processing performance of HBase, voluminous unstructured data processing is not performed in HBase. Instead, real-time data processing and unstructured data processing are handled by Spark, which derives real-time decisions from streaming data. External processing at Spark facilitates flexibility, while increasing the processing efficiency. External Spark processing

creates semi-structured data that are storable in HBase. The intelligent agents derive decisions corresponding to processed data considering rules defined in the rules engine. The rules are created from archived data processing.

Data Acquisition is SC data analysis is characterized by the necessity to connect the analysis with the data's spatial dimension. It is meaningless to analyze data without linking them to the spatial dimension. Of course, the level of spatial analysis changes according to business needs; it may require analysis at the level of the whole city, the level of districts, or even at the street level. The other natural dimension is the time dimension. Collectively, these two dimensions are called spatiotemporal coordinates.

Finding the appropriate datasets for SC analytics is one of the challenges in harnessing BDA in SCs. Conducting analyses that include the spatial and temporal dimensions enables integrating many analyses from several domains. The spatiotemporal coordinates of the domains (d), services (s), events (e) and so on, which are meant to be tracked and analyzed, are therefore recorded with the datasets identified for analysis. This will support establishing a comprehensive reference spatiotemporal repository for the deduced analytical results. The granularity level of the spatiotemporal coordinates depends on the nature of the domain and the required analytics. Considering these characteristics in the data candidate for analysis in SCs, we added one more design principle to SCDAP design principles, which is the principle of spatiotemporal data.

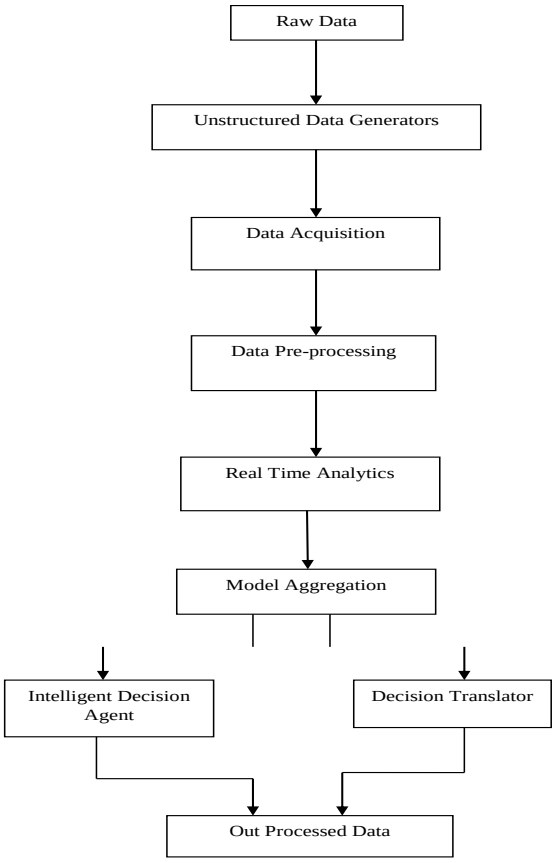

Fig.1: A Frame Work Of Design And Implementation Of Smart City Big Data Processing Platform Using Big Data Analytics For Decision Management System

Data preprocessing functionalities are implemented using Python scripts running through the RapidMiner Python operator. These functionalities include tasks to remove anomalies from raw data in preparation for the analysis stage. Data preprocessing includes the following:

- Removing special characters from username, business name, and user reviews.
- Data cleansing: Filling in missing values and smooth noisy data, identifying or removing outliers, and resolving inconsistencies.
- Data integration: Integration of multiple databases, data cubes, or files.
- Data reduction: Dimensionality reduction, numerosity reduction, and compression.
- Data transformation and discretization

Real-time analytics solutions based on the innovative streaming database support complex query and analysis operations. You can query materialized views with simple SQL statements to gain real-time data insights, leverage data value, and make instant business decisions.

Batch data analytics is batch processing is when the processing and analysis happens on a set of data that have already been stored over a period of time. An example is payroll and billing systems that have to be processed weekly or monthly. Streaming data processing happens as the data flows through a system.

Model aggregation is viewed as a sub-function of the model management functionality, and there is no need to treat it separately. Regarding the definition of these two functionalities, it agree on these two observations, as the word "model" does not reflect the required meaning of the word properly; this word refers to the resultant data model or simply the resultant model. Therefore, it is more appropriate to rename the model management function as resultant model management. As for the second comment, model aggregation is meant to deal with (manage) various resultant data models. In that sense, we agree to include the aggregation functionality in the resultant model management.

Intelligent decision agents classify valuable data and derive intelligent decisions. Since decisions are the key factors for performance improvement the proposed scheme aims to derive intelligent decisions from offline data processing as well as online data processing. The service management tier is responsible for formulating and generating smart city service events with respect to derived intelligent decisions. After the intelligent decision the data transfer to the decision translator.

The intelligent agents are deriving decisions corresponding to out processed data considering rules in the rules engine. The rules are created from archived data processing. Further, threshold values and constraints for each dataset are defined in the rules engine. Subsequently, the intelligent broker represents derived decisions adhering to a pre-defined vocabulary shared throughout the city framework. Shared vocabulary simplifies decision translation throughout the city architecture, while avoiding ambiguities.

IV. RESULT ANALYSIS

In this section, the smart city architecture is designed to perform real-time data processing, which improves autonomous decision generation. Analysis of large datasets belonging to

various fields influence the performance improvement of city operations. Authentic datasets obtained to evaluate data processing performance.

Accuracy, Precision, Recall and F1 score are different parameters used in this analysis for performance evaluation.

Accuracy means the ratio of exactly estimates for complete detections. Accuracy can be mentioned as capability to accurately detect result of a situation.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \dots (1)$$

Precision:

Positive detected value or precision is number of exact positive scores divided by number of true scores detected by the categorization algorithm represented in below equation (2)

$$Precision = \frac{TP}{TP + FP} \dots (2)$$

Where,

1) True Negative (TN) – A percentage of non-instances which are exactly grouped.

2) True Positive (TP) – A percentage of instances which are exactly grouped.

3) False Positive (FP) - The percentage of instances which are not exactly grouped.

4) False Negative (FN) - The percentage of non-instances which are not exactly grouped.

Table 1: Comparative Performance Analysis Of Different Frameworks

Parameters	**Decision Management System**	**FrAPPE**	**DRSC**
Accuracy	97	91	85
Precision	98.2	96.5	92.1
Recall	89.7	78.6	74.9
F1 Score	91.6	90.5	92.7

Graphical representation of accuracy for the Decision Management System**,** Frame, Pixel, Place, and Event (FrAPPE) and Disaster- Resilient Smart Cities (DRSCs). Accuracy of Decision Management System (DMS) shows higher.

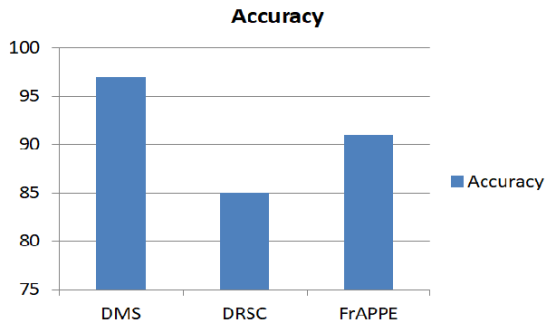

Fig.2 Accuracy Comparison Graph

In fig.3 precision comparision graph is seen. Comparision is inbetween Decision Management System (DMS), FrAPPE and DRSCs.

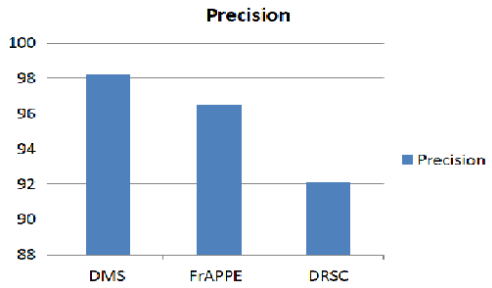

Fig.3 Precision Comparison Graph

Recall comparision is done between frameworks of DMS, FrAPPE and DRSC. Higher recall is observed in DMS.

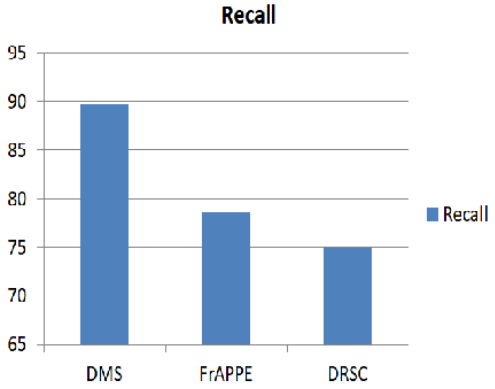

Fig.4 Recall Comparison Graph

In fig.5 F1 score graphical representation is observed in different frameworks of Decision Management System**,** Frame, Pixel, Place, and Event (FrAPPE) and Disaster- Resilient Smart Cities (DRSCs).

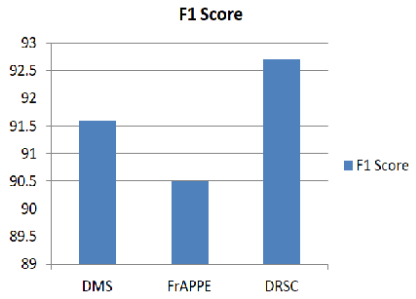

Fig.5 F1 Score Comparison Graph

V. CONCLUSION

Smart cities have improved the community with the aid of sophisticated innovations in healthcare, transportation, utility management, and much more, since a smart city is a consolidation of various smart components. The realization of an intelligent smart city relies on seamless interoperation and coherent integration of all fundamental smart components. In the modern world, smart cities generate data at a very high speed. Due to exponential data growth, real-time data processing and analysis have become tedious and challenging for smart cities. Therefore, herein they have described a BDA-embedded experimental architecture for smart cities. Moreover, data translator and intelligent decision agent techniques are integrated to the described analysis, to further enhance offline and online data processing tasks. Hence, the described frameworks shows higher accuracy, precision, recall and F1Score for Decision management system for Smart city development using big data analytics.

VI. REFERENCES

1. Iqbal, R.; Doctor, F.; More, B.; Mahmud, S.; Yousuf, U. Big data analytics: Computational intelligence techniques and application areas. *Technol. Forecast. Soc. Chang.* **2020**, *153*, 119253, doi:10.1016/j.techfore.2018.03.024.
2. Alexopoulos, C.; Pereira, G.V.; Charalabidis, Y.; Madrid, L. A taxonomy of smart cities initiatives. In Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance, Melbourne, Australia, 3–5 April 2019.
3. Osman, A.M.S. A novel big data analytics framework for smart cities. *Future Gener. Comput. Syst.* **2019**, *91*, 620–633.
4. Shah, S.A.; Seker, D.Z.; Rathore, M.M.; Hameed, S.; Ben Yahia, S.; Draheim, D. Towards Disaster Resilient Smart Cities: Can Internet of Things and Big Data Analytics Be the Game Changers? *IEEE Access* **2019**, *7*, 91885–91903, doi:10.1109/access.2019.2928233.
5. Balduini, M.; Brambilla, M.; Della Valle, E.; Marazzi, C.; Arabghalizi, T.; Rahdari, B.; Vescovi, M. Models and Practices in Urban Data Science at Scale. *Big Data Res.* **2019**, *17*, 66–84, doi:10.1016/j.bdr.2018.04.003.
6. D. Dissanayake and K. Jayasena, "A cloud platform for big IoT data analytics by combining batch and stream processing technologies," in *The 2017 National Information Technology Conference*, 2017, pp. 40–45.

7. C. Misale, M. Drocco, M. Aldinucci, and G. Tremblay, "A comparison of big data frameworks on a layered dataflow model," *Parallel Processing Letters*, vol. 27, no. 01, p. 1740003, 2017.
8. D. Crankshaw, X. Wang, G. Zhou, M. J. Franklin, J. E. Gonzalez, and I. Stoica, "Clipper: A low-latency online prediction serving system." in *The 14th USENIX Symposium on Networked Systems Design and Implementation*, 2017, pp. 613–627.
9. Y. Zhang, T. Cao, S. Li, X. Tian, L. Yuan, H. Jia, and A. V. Vasilakos, "Parallel processing systems for big data: a survey," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2114–2136, 2016.
10. Psomakelis, E.; Aisopos, F.; Litke, A.; Tserpes, K.; Kardara, M.; Campo, P.M. Big IoT and Social Networking Data for Smart Cities Algorithmic Improvements on Big Data Analysis in the Context of RADICAL City Applications. *arXiv* **2016**, arXiv:1607.00509
11. H. Cho, H. Shiokawa, and H. Kitagawa, "Jsflow: Integration of massive streams and batches via JSON-based dataflow algebra," in *The 19th International Conference on Network-Based Information Systems*, 2016, pp. 188–195.
12. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. Int. J. Inf. Manag. **2015**, 35, 137–144.
13. A. J. Jara, D. Genoud, and Y. Bocchi, "Big data in smart cities: from poisson to human dynamics," in *Advanced Information Networking and Applications Workshops (WAINA), 2014 28th International Conference on*. IEEE, 2014, pp. 785–790.
14. N. Bicocchi, A. Cecaj, D. Fontana, M. Mamei, A. Sassi, and F. Zambonelli, "Collective awareness for human-ict collaboration in smart cities," in *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2013 IEEE 22nd International Workshop on*. IEEE, 2013, pp. 3–8.
15. Arunachalam, R.; Sarkar, S. The New Eye of Government: Citizen Sentiment Analysis in Social Media. In Proceedings of the IJCNLP Workshop on Natural Language Processing for Social Media (SocialNLP), Nagoya, Japan, 14–18 October 2013.
16. Greaves, F.; Ramirez-Cano, D.; Millett, C.; Darzi, A.; Donaldson, L. Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. *J. Med Internet Res.* **2013**, *15*, e239, doi:10.2196/jmir.2721.
17. Ceron, A.; Curini, L.; Iacus, S.M.; Porro, G. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media Soc.* **2013**, *16*,340–358,doi:10.1177/1461444813480466.
18. Dunne, T. Big Data, Analytics, and Energy Consumption; Lavastorm Agle Analytics: Boston, MA, USA, 2012.
19. Naccarati, F.; Hobson, S. IBM Smarter City Solutions on Cloud. In IBM Global Services White Paper-Government Solutions; IBM: Somers, NY, USA, 2011.
20. Yuanming Yuan.Key technologies of Smart City information system.Wuhan:Wuhan University,2012