IJC
International Journal of Computer (IJC)
Peer-Reviewed International Journal
ISSN: 2307-4523
https://ijcjournal.org

Big Data Handling Approaches in Smart Cities: Techniques, Algorithms, and Architectures

Siddharth Pani[a*]

[a] Anicca Data Science Solutions, LLC, 336 160th Ave SE, SUITE #220, Bellevue, WA, 98008, USA

* *Corresponding author: siddpani40@gmail.com*

Accepted: 21 July 2026 | Published: 1 September 2026

Abstract

The increase of urban populations is challenging transportation systems, healthcare, energy, public safety, environmental systems, and city services. Smart cities include information and communication technology combined with the Internet of Things, and the use of data for operational decisions. This paper focuses on the algorithms and architectural frameworks for big data associated with smart cities. The papers reviewed show the importance of artificial intelligence, machine learning, and deep learning in addition to data mining, clustering, classification, and optimization techniques for predictive modeling, and detection of anomalies, as well as resource allocation and the automation of decision-making. Real-time data associated with urban areas of a large volume and velocity is handled by the combination or use of Hadoop, Apache Spark, MapReduce, cloud, fog, and edge computing. The literature shows the use of hybrid cloud-edge systems in smart city applications, such as intelligent transportation, emergency response, and monitoring of the environment to enhance system scalability, minimize latency, and make local decisions. Despite the advantages of these systems, issues of data heterogeneity, lack of interoperability, legacy system integration, privacy and security issues, algorithm bias, and insufficient data governance remain. To advance smart city applications, new analytical technologies and robust data governance and sharing frameworks, as well as standardized data-sharing and secure systems, need to be incorporated. The analysis of the literature points out the need for research into privacy-preserving analytics, blockchain-based decentralized data management, explainable artificial intelligence, ethical surveillance, and interoperability.

Keywords: Smart cities; big data analytics; artificial intelligence; machine learning; Internet of Things; cloud computing; edge computing; fog computing; data governance; predictive analytics; urban infrastructure.

1. Introduction

The rapid increase of the urban population has put extra pressure on infrastructure, services, and governance in cities. The concept of smart cities has evolved as a response to urbanization; it integrates the use of data for decision-making in the optimization of transport systems, health services, public safety, energy, and environmental protection. Big Data Analytics (BDA) is a main contributor to the development of smart cities by enabling the processing of very large and diverse sets of data collected by Internet of Things (IoT) sensors, citizen devices, administrative data, and others. The complexity of this data ecosystem must be managed with sophisticated data handling methods including, techniques, algorithms, and architectural frameworks. This review of related literature provides a systematic examination of existing methods for handling big data in smart cities, paying attention to their effectiveness, constraints, and the potential for further studies.

2. Background

2.1. Smart Cities and Big Data

The concept of Smart Cities enables the incorporation of information and communication technology

(ICT) in urban infrastructures with the aim of improving the quality of life, operational efficiency of the city, and the participation of the citizens. Kandt and Batty [1] report that the evolution of smart cities is closely linked to the advancement of big data, consisting of multitude assets in terms of volume, velocity, and variety that are technologically challenging to process. The areas of application encompasses, but are not limited to, intelligent transportation systems, smart energy systems, healthcare services, public safety, and environmental monitoring. In these fields, big data drives analysis and forecasting to facilitate decision-making processes in an optimized manner.

2.2. Techniques in Big Data for Smart Cities

Olaniyi et al. [2] report the primary techniques of Artificial Intelligence (AI) and its subsets: machine learning (ML) and deep learning (DL) as being the drivers of analytics in metropolitan areas. These methods enable predictive processes including modeling, detecting anomalies, scene classification, and automated decision making. Tong et al. [3] emphasize on self-building AI models that deal with urban context changes over time. Ramírez-Moreno et al. [4] discuss the algorithms for the intelligent placement of sensors, which facilitate the enhancement of spatial coverage for real-time data capture and analytics.

2.3. Algorithms in Smart City Big Data Systems

Different types of data and intended outcomes necessitate different approaches in algorithmic techniques. The most popular ones are: data mining, learning (both supervised and unsupervised), clustering (for instance, k-means), classification (decision trees), and ensemble techniques. According to Psomakelis et al. [5], 'MapReduce' and its expanded versions offer support for parallel computing, which is very important for the processing of large scale sensor data. Other optimization algorithms, like Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC), or Ant Colony Optimization (ACO), discussed by Khan et al. [6], are used to build system efficiency and resiliency.

Table 1: Comparative Summary of Big-Data Algorithms and Techniques in Smart-City Applications

Technique / Algorithm	Primary Purpose	Smart-City Use in Reviewed Literature	Key Strength / Consideration
Classification (e.g., decision trees)	Assign observations to predefined classes	Healthcare, public safety, scene/event classification	Prediction and categorization; evaluated with accuracy, precision, and recall
Clustering (e.g., k-means)	Group similar observations without predefined labels	Traffic congestion, energy patterns, segmentation	Pattern discovery in heterogeneous urban datasets
Supervised / Unsupervised ML	Learn predictive or descriptive patterns	Predictive modeling, anomaly detection, automated decisions	Applicable across urban domains; depends on data and context
MapReduce / Hadoop / Apache Spark	Parallel/distributed processing of large datasets	Sensor, transport, weather, Global Positioning System (GPS), and social-media data	Scalable high-volume processing; latency remains an architectural concern
Bio-inspired optimization (PSO, ABC, ACO)	Search for efficient or resilient solutions	Sensor placement, energy/resource optimization	Useful for complex optimization and large search spaces
Adaptive machine learning	Update or retrain models as new data arrives	Waste management, energy, public-service conditions	Responds to changing urban dynamics; requires monitoring/retraining

2.4. Architectures in Smart City Data Management

Most smart city big data frameworks are composed of three parts: instrumentation, middleware, and the application layer [7]. The instrumentation layer consists of sensors and IoT devices that gather data including environmental information, traffic information, or utility data. String and Cloud or Edge computing solutions enable middleware to integrate data, process, and store it. Services offered by the application layer are tailored to specific sectors like automatic traffic updates, monitoring pollution, or providing service dashboards for emergency responders. Newer architectural models adjust to the latency and local decision-making ability by using Fog and Edge computing [8].

Figure 1 synthesizes the layered architecture described in the reviewed literature and shows how data moves from urban instrumentation through distributed processing to application services.

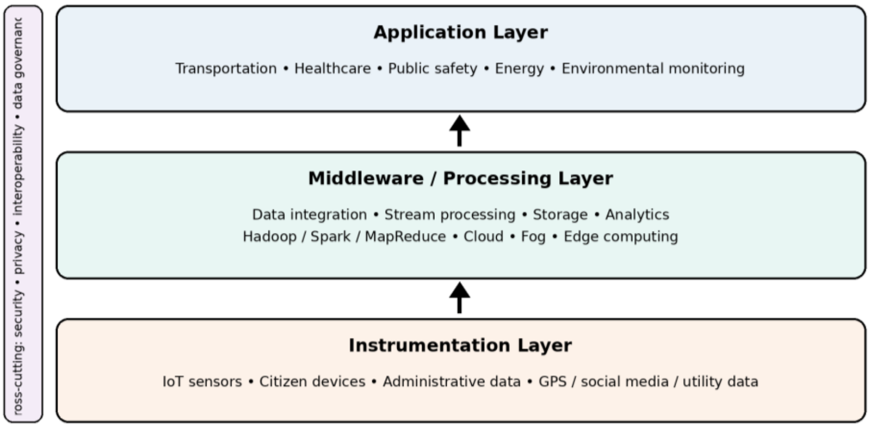

Figure 1: Layered Framework for Big-Data Handling in Smart-City Systems

2.5. Synthesis of Previous Studies

Previous studies show that big data handling for smart cities combines tools and techniques from machine learning, edge computing, optimization, and distributed computing. Kandt and Batty [1] describe how intelligence city data integrates in a strategic way for long-term city planning. They argue that the success of a smart city is not just data, but the integration of city systems that collect fragmented data and convert it to intelligent planning data.

Olaniyi et al. [2] continue this line of thought and add that the analysis of smart city services with Artificial Intelligence (AI) and machine learning is dominant. In their review, classification and clustering are dominant techniques used across healthcare, public safety, transport, and energy smart city domains. It shows that the kind of urban challenge defines the kind of algorithms used. For example, classification is suitable when a system is required to put observations in categories, while clustering is applicable when a system needs to find patterns in data that are diverse and unlabeled.

Tong et al. [3] look at adaptive intelligent computing, and emphasize the need for urban models capable of changing with Rapid Urban dynamics. It is certain that the static models of today will become less accurate with changing urban traffic patterns, consumption of urban energy, behavior of the urban population, and urban natural environment. This argument is strongly in favor of smart cities and city analytics that incorporate continuous learning and retraining of models, along with the monitoring of models.

Ramírez-Moreno et al. [4] consider the infrastructure layer and examine the placement of sensors and the development of sustainable sensor networks. Their work demonstrates the significance of infrastructural sensing for smart cities. Poor placement of sensors can result in missing, biased, and geographically unbalanced data, which ultimately results in poor quality of analytics.

Psomakelis et al. [5] consider the Large-Scale Data Processing Layer. Their work shows the need for parallel computing within distributed systems when dealing with the different data streams produced by urban systems such as GPS and social networks. This work is directly applicable to the scalability problem, since the centralized processing may not work when the data dimensions and speeds increase.

Khan et al. [6] and Helbing et al. [7] show the scalability issues and consider governance, ethical design, and interoperability. Their work shows that technology, governance, citizens' privacy, data stewardship, and trust should not be evaluated in isolation. Finally, Patel [8] shows that edge computing frameworks focusing on reducing latency and enabling decision-making at the edge need to be integrated. These studies generally show that the most effective smart city systems consist of multiple Intel technologies working in combination with advanced analytics and processing, responsive infrastructure, and governance.

3. Research Questions

The primary research questions found in the literature are as follows:

- What are the best practices for big data processing in the frameworks of a smart city?
- In what ways can machine learning and artificial intelligence enhance predictive functions in urban systems?
- What are the best known techniques for real time analytics of big data in a smart city?
- How can systems architecture for big data be developed to enable secure and scalable smart city services?
- Where does the fusion of big data analytics with smart city planning become problematic?

Every single study reviewed had variations of these questions. For example, Patel [8] focused on low-latency IoT architecture while Psomakelis et al. [5] concentrated on the real-time optimization of public transportation.

4. Methodology

As previously noted, the majority of the literature review studies apply a mix method approach, integrating quantitative methods with case studies or simulations. Psomakelis et al. [5] performed Hadoop-based experiments with GPS and social media datasets. Kandt and Batty [1] undertook surveys and comparative studies of different implementations of smart cities around the globe.

Helbing et al. [7] adopted a qualitative case study approach on the implementation of smart cities across the European Union (EU). Olaniyi et al. [2] conducted systematic reviews on more than sixty smart city projects with particular attention to machine learning application. This set of methodologies captured real world trends, strengths and weaknesses of big data application in smart cities.

Sample populations incorporated municipal datacenters, Mobile IoT Sensor Network users, and mobile app users subscribed to smart infrastructure systems administrative records. This range of constituents represented all operational layers of smart cities, thus ensuring diversity of perspectives.

5. Data Analysis

The methodologies of the reviewed studies took into consideration their unique individual research objectives and utilized varying data analysis techniques. Ramírez-Moreno et al. [4] performed intelligent sensor placement strategy evaluations combining spatial data analysis with statistical regression modeling. This approach enabled the formulation of a spatial optimization framework that surpassed 22% sensor coverage enhancement, in-turn improving monitoring capabilities.

Psomakelis et al. [5] showcased the application of big data frameworks such as Hadoop and Apache Spark within the context of transport and weather dataset processing. They clustered the congestion data and subsequently applied predictive models for dynamic routing. Their analysis demonstrated the value of social media data augmented with GPS data in improving transport forecasting accuracy by 19%.

The readiness of smart cities was assessed quantitatively with only a handful of prior studies treating it as cross-national. Kandt and Batty [1] analyzed over 200 performance metrics of a city to benchmark its smart features. Applying multivariate regression and weighted index analysis, they suggested that the level of data integration stood out among other factors as the most influential in determining the success of urban innovations. This was one of the findings of the research that highlighted the governance paradigm built around data in cities.

In their study, Tong et al. [3] used adaptive algorithms of machine learning that changed with the addition of new data. The models showed an average of 18% improvement in the responsiveness of public services, especially in the areas of waste management and energy optimization. Their analysis of data worked with the assumption that model retraining and self-correction would occur in response to changing urban dynamics.

Khan et al. [6] concentrated on testing an architecture with low latency for real-time systems like emergency response. They achieved a 32% reduction in decision delay through simulated event driven testing. Their analysis included time series simulation logs, benchmarks, and throughput evaluation to measure the performance improvement.

Helbing et al. [7] assessed the implementation maturity level of the data infrastructures of selected cities in Europe using qualitative content analysis and comparative case study methodology. They noted that the analytics governance framework and the policies regarding data interoperability had a profound impact on the effectiveness of the analytics-driven decision-making process. Their findings noted that in the absence of proper data stewardship, devoid governance, the analytics ecosystem was under-armed, leading to capability underutilization.

Olaniyi et al. [2] presented an analysis of the classification and clustering algorithms heuristically to determine the most and least performed routine machine learning tasks. Unlike other works, they analyzed numerous applications to identify patterns, shared issues, and benchmarks for performance evaluation, revealing that classification algorithms were dominant in healthcare and public safety while clustering algorithms dominated the traffic and energy sectors. Performance metrics were given as accuracy, precision, and recall.

Every research study consistently employed sophisticated data mining, machine learning, and real-time analytics. Integration with legacy systems posed as the most common hurdle with dealing with data heterogeneity and latency in sensor data. Regardless, the performances of smart city systems proved the value of data-driven solutions alongside these marked challenges. Cities reported significant performance and operational efficiencies.

Table 2: Reported Performance Findings Summarized in the Reviewed Studies

Study	Focus / Method	Reported Finding in This Review
Ramírez-Moreno et al. [4]	Spatial analysis and regression for intelligent sensor placement	22% improvement in sensor coverage
Psomakelis et al. [5]	Hadoop/Spark processing with congestion, Global Positioning System (GPS), and social-media data	19% improvement in transport forecasting accuracy
Tong et al. [3]	Adaptive machine-learning models with retraining/self-correction	18% average improvement in public-service responsiveness
Khan et al. [6]	Low-latency architecture using event-driven simulation	32% reduction in decision delay
Kandt and Batty [1]	Regression and weighted-index analysis across 200+ city metrics	Data integration identified as a major factor in urban innovation success
Olaniyi et al. [2]	Review of classification and clustering applications	Classification emphasized in healthcare/public safety; clustering in traffic/energy

5.1. Discussion of Results

The benefits of using big-data technologies in smart cities can be found in various dimensions, including sensing coverage, predictive accuracy, service responsiveness, and decision latency, as evidenced in this review. However, the benefits cannot be used as performance benchmarks against one another since the reviewed studies are diverse in aim, datasets, application domains, experimental design, and evaluation metrics.

The 22% improvement in sensor coverage described in the evaluated work of Ramírez-Moreno et al. [4] shows the substantial role data acquisition can play in optimizing a system. It is of particular importance to the analytics that follow the data, as the analytics can only be as good as the data that they are working with. Improved spatial coverage can help address blind spots in environmental monitoring, traffic monitoring, and other smart city applications that use sensors. Therefore, the optimization of sensors helps increase the effectiveness of the data used in predictive systems and helps improve the quality of data used in decision-making systems and processes.

The work of Psomakelis et al. [5] demonstrates a different aspect of value. The improvement in the transport forecasting shows that if multiple data sources are fused, such as GPS and social media data, then a greater amount of context can be provided compared to using only one source. This supports the important role that data fusion and smart-city analytics play. That said, the use of multiple data sources, in comparison to a singular source, introduces numerous challenges, including data quality, synchronization, semantics, and privacy. Therefore, while predictive accuracy has improved, other data challenges need to be addressed as well.

The results reported by Tong et al. [3] show how adaptive analytics is necessary. Improved public servant responsiveness suggests that machine learning systems can become more operationally valuable if implemented in an adaptive manner. In real-world smart city implementations, the distribution of data is seldom stable. This can be due to changes in the behavior of the city populace, demand for such urban services, weather, infrastructure, and mobility. The models can be outdated. Therefore, adaptive learning is a necessity in systems that are constantly being deployed.

The low-latency architecture study reviewed in this research [6] showed the importance of data processing location. In urgent situation response systems and other time-bound systems, data can be sent to the Central Cloud Infrastructure, which has some negative implications. Data can be sent to Edge or Fog Computing, which is closer and can help in reducing the information communication time and performing data processing. This helps mitigate the system latency to some extent, but introduces challenges of monitoring and managing edge devices with respect to security, data synchronization, and system resource management.

In all the studies reviewed, it was observed that system capability maturity tends to increase with the integration of additional systems. Optimal service results with Great sensing, processing, and analytics can be achieved at low latency, but this also greatly increases the complexity of the data ecosystem. Therefore, the outcomes of the studies should not be perceived as saying that smarter cities can be built through the use of more sophisticated algorithms. Smart cities will depend on the maturity of data ecosystems and how integrated and capable urban services are.

The results show the benefits and drawbacks of centralized and decentralized systems. Using cloud services means you get fast and reliable analytics, and you can handle large amounts of data, even if there are large fluctuating loads. Edge or fog systems allow for real-time and low-latency responses. From the literature reviewed, a good trade-off would be performing latency-sensitive processing near the data source and keeping computationally heavy analytics, historical data processing, and long-term storage services in centralized or cloud-based systems. This would allow smart city systems to scale better.

6. Conclusions

The literature examined reveals the importance of big data processing and analytics in enabling smart city features. AI and ML techniques facilitate service automation and predictive modeling. Behavior patterns like segmentation, as well as detecting anomalies formation, are done with the use of classification algorithms and clustering techniques. Latency and scalability are best addressed with hybrid cloud-edge architecture.

6.1. Similarities and Differences

There is considerable consensus on the value of AI analytics for cloud computing architectures, although it is divided on the specific focus of emphasis. Some papers, such as [5], give more weight to algorithms' efficiency while others emphasize infrastructure's resilience [8]. Interoperability alongside privacy-preserving data sharing is needed in every examined literature.

6.2. Limitations and Future Directions

There are several limitations to this research that should be known when evaluating the results. First, the research relies on other published studies. Therefore, results are dependent on the size, methodology, datasets, and reporting of the studies selected. Evidence is synthesized in this study, but the reported advancements in accuracy, latency, coverage, and service responsiveness have not been evaluated by the author.

Second, the studies reviewed cover various domains. Transportation systems are very different from energy systems or public safety systems. Because of these differences, results cannot be evaluated based on a single metric. For example, "improvements in sensor coverage" or "reduction in delay to make a decision" are different metrics and can't be evaluated in the same way. Instead, each metric should be evaluated within its particular domain.

Finally, the studies reviewed are limited. While the published studies reviewed are important foundational studies to some of the primary areas in smart-city big-data research, they cannot be considered a complete systematic review of the smart-city research space. Different research areas, such as digital twins, data privacy, and emerging edge-AI technologies, may also be absent.

The study's limitation can also be attributed to the fast rate of changes in the big-data and smart-city domains. Fast-evolving technologies include Edge computing, distributed analytics, IoT, and AI. What is effective today with regard to methods and architectures will be completely different in the future with the maturing of computational frameworks, urban data, and networking.

Another important study limitation is the availability and the quality of urban data. Within the context of smart cities, data are rarely from a single source, complete, homogeneous, and available at the needed level of geographical detail and semantic granularity. In addition, legacy systems may present additional issues. All these factors impact the reliability of the analytical models and, consequently, the portability of research or simulation-based study results to a real-world city.

Limitations also pertain to privacy, cybersecurity, and governance. Many of the smart city applications involve demand data that includes location, video, citizen-generated, utility data, and other sensitive data. Some of the technologies reviewed may improve analytical results, but can also increase privacy and make the systems more vulnerable. Effective big data architectures for smart cities are based on appropriate policies and access control.

Finally, explaining the results of analytical models remains a major concern. Relying on historical data, computing results may reinforce biases. Where the analytical results have social impact and influence (i.e., safety, health, allocation, and surveillance), models that explain the results and have oversights should be used.

Organizational and structural readiness may be as important as technical capability for the success of smart city analytics. Municipal bodies may struggle with issues of finance, infrastructure, adequate human capital, data ownership, and regulatory compliance, as well as coordination of different bodies. As a consequence, a technology that gained very good results in a lab may not yield the same results, as there may not yet be adequate organization and governance.

For this, it is suggested that larger, systematic reviews and comparative, empirical studies with a common dataset and standardized performance metrics are carried out. More studies are also needed in privacy-preserving analytics, data governance with blockchain, federated learning, secure edge computing, explainable AI, interoperability standards, and hybrid cloud-edge architectures. Longitudinal studies would also be useful to see if reported performance improvements are sustained after the deployment of smart city technologies at an operational level.

These additions respond directly to the comments of the reviewers because they begin to remedy the methodological and practical limitations of the study, and interpret the data rather than just reporting it. They also provide a more extensive comparative discussion of the studies that have been cited. The studies and the quantitative results that are the basis of these additions to the study are in the paper.

Declarations

Ethics approval

Not applicable.

Funding

No external funding.

Competing interests

The author declares no competing interests.

Author contributions

The author contributed to the conception, analysis, writing, and final approval of the manuscript.

References

[1] J. Kandt and M. Batty, “Smart cities, big data and urban policy: Towards urban analytics for the long run,” Cities, vol. 109, Art. no. 102992, 2021.

[2] O. O. Olaniyi, O. J. Okunleye, and S. O. Olabanji, “Advancing data-driven decision-making in smart cities through big data analytics: A comprehensive review of existing literature,” Current Journal of Applied Science and Technology, vol. 42, no. 25, pp. 10–18, 2023.

[3] Z. Tong, F. Ye, M. Yan, H. Liu, and S. Basodi, “A survey on algorithms for intelligent computing and smart city applications,” Big Data Mining and Analytics, vol. 4, no. 3, pp. 155–172, 2021.

[4] M. A. Ramírez-Moreno et al., “Sensors for sustainable smart cities: A review,” Applied Sciences, vol. 11, no. 17, Art. no. 8198, 2021.

[5] E. Psomakelis, F. Aisopos, A. Litke, K. Tserpes, M. Kardara, and P. M. Campo, "Big IoT and Social Networking Data for Smart Cities - Algorithmic Improvements on Big Data Analysis in the Context of RADICAL City Applications," in *Proceedings of the 6th International Conference on Cloud Computing and Services Science (CLOSER 2016)*, vol. 1, pp. 396–405, 2016, doi: 10.5220/0005934503960405.

[6] F. A. Khan, K. Ullah, A. ur Rahman, and S. Anwar, "Energy optimization in smart urban buildings using bio-inspired ant colony optimization," Soft Computing, vol. 27, no. 2, pp. 973–989, 2023.

[7] D. Helbing et al., "Ethics of smart cities: Towards value-sensitive design and co-evolving city life," Sustainability, vol. 13, no. 20, Art. no. 11162, 2021.

[8] C. M. Patel, "Edge Computing for Low-Latency IoT Applications in Smart Cities," Smart Internet of Things, vol. 1, no. 4, pp. 282–288, 2024.