Volume 1, 2025, Pages 33-42 *Journal of Computer and Communication Networks*
| Regular Article | Open Access

Real Time Data and Decisions in Aarhus and Surrey Smart

Nandhini Priya T
Department of Electrical and Electronics Engineering, Excel Engineering College, Komarapalayam, Namakkal, India. nandhinipriyat.eec@excelcolleges.com

Article Info
Journal of Computer and Communication Networks
https://www.ansispublications.com/journals/jccn/jccn.html

© The Author(s), 2025.
https://doi.org/10.64026/JCCN/2025004

Received 25 December 2024
Revised from 29 January 2025
Accepted 12 February 2025
Available online 28 February 2025
Published by Ansis Publications

Corresponding author(s):
Nandhini Priya T, Department of Electrical and Electronics Engineering, Excel Engineering College, Komarapalayam, Namakkal, India.
Email: nandhinipriyat.eec@excelcolleges.com

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/ licenses/by/4.0/).

Abstract – This study investigates the performance of real-time data processing systems in smart cities, with a focus on the implementation of autonomous decision-making processes for urban management. The research collects data sets from Aarhus (Denmark) and Surrey (Canada) to analyze data processing performance using traffic, parking, pollution and water consumption datasets. The data processing timeframe and data passage rates benefit from implementing the min-max normalization technique together with other filtration methods. The BDA-based smart city architecture shows a positive effect on real-time decision efficiency which benefits traffic congestion control and parking space tracking and water consumption assessment. Through independent event-driven notifications the system helps urban management as well as improves service quality delivered to city dwellers. Application of data filtration improves Hadoop and Spark-based framework performance for large dataset processing according to the study which demonstrates better processing time and throughput results.

Keywords – Real-Time Data Processing, Smart Cities, Autonomous Decision-Making, Urban Data Analysis, Traffic Management, Pollution Monitoring, Data Normalization, Hadoop, Spark, Smart Parking Systems.

I. INTRODUCTION

In September 2008, Nature released a special edition entitled “Big Data”. In February 2011, Science released a special edition entitled “Dealing with Data” [1]. These releases indicate that the epoch of big data has commenced. In March 2012, the Obama government formally initiated and announced the “Big Data Research and Development Initiative.” This program regards big data as the "engine" of the future. This initiative's significance is analogous to the 20th century's data highway strategy. Researchers and policymakers have acknowledged that big data will serve as a valuable resource for extracting knowledge and information. As smart cities are progressively developed and implemented, humans and diverse sensors will generate increasing amounts of data. Furthermore, numerous initiatives have been undertaken by industrial and academic specialists to actualize the concept of smart urban environments. Islam et al. [2] address various particular topics of interest, including water management, waste management, and parking management, among others. Consequently, a comprehensive and robust smart city framework has emerged as an essential requirement, as a deficiency in integrity undermines its feasibility. Furthermore, it must enable automated behaviors, real-time decision-making, and real-time data processing, and intelligent energy usage and customization. Consequently, the analysis and processing of big data becomes essential. Consequently, the smart IoT incorporates big data analytics (BDA) for the actualization of the intelligent urban environments. An intelligent meter in a suburban building gathers meter readings, which are then compared to a predetermined power consumption threshold; based on this comparison, the present energy requirement is communicated to the intelligent grid.

Concurrently, consumers are informed of the present energy consumption level, enabling them to manage energy usage effectively. The aforementioned scenario produces a substantial quantity of data for each one residence. Furthermore, data processing and decision-making must occur promptly. Nonetheless, several household and public facilities in the city provide a substantial volume of information pertaining to the aforementioned duty. Consequently, the integration of data sources and BDA is regarded as an effective method to enable real-time functionality within the intelligent city. Data noise constitutes a prevalent issue in metropolitan settings. It influences human behavior, health, and children's cognitive development. The

European Commission acknowledged this significant issue and enacted a rule mandating that big cities (with populations over 100,000) collect empirical data on noise exposure to develop local action plans and create precise mappings of noise pollution levels [3]. The conventional method of performing noise measurements involves the manual collection of noise samples; however, this approach has numerous disadvantages. On one hand, only localized and sparse measurements are obtained. Conversely, it incurs high expenses attributable to the expenditures of measurement equipment and personnel. The European Commission advocates for increased granularity of noise data in both spatial and temporal dimensions. In this context, Wireless Sensor Networks (WSNs) serve as a viable solution to mitigate the limitations of the existing noise data collection methodology. These wireless sensor networks consist of diminutive autonomous nodes equipped with sensing capabilities. Every node possesses an independent power source, computing unit, and memory. Processing delays are a significant challenge for urban environments. Several initiatives have examined this issue within the framework of cloud computing. Rusitschka, Eger, and Gerdes [4] examined the prerequisites for implementing smart grid applications on cloud servers. They discovered that numerous viable energy management applications necessitate the scale that only cloud computing can offer. Nonetheless, these applications also entail supplementary needs, including the provision of scalable real-time services, the assurance of consistent and fault-tolerant services, and the safeguarding of privacy. Cloud computing presently fails to meet these requirements. The scholars presented a cloud-based architecture for the management of smart grid data to satisfy the near real-time information retrieval requirements of diverse energy market participants. The distributed data management and parallel processing methodologies are specifically tailored for time series, the predominant data type produced by metropolitan environments.

A significant problem inherent in the intelligent city project is the architecture of data flow, from the initial data acquisition by a sensor to its ultimate application. According to Mahbub [5], a conventional IoT environment is structured into 3 layers: (i) a foundational tier comprising objects and devices, (ii) an upper tier featuring cloud system nodes, and (optional) (iii) a central tier with edge nodes and intelligent gateways. The processing capabilities, scalability, and adaptability of every layer will dictate the efficacy of an intelligent urban ecosystem. In the context of intelligent cities, particularly to facilitate smartphones, it is imperative to establish an intermediary tier that offers intelligent gateways for managing links and processing data. Nodes rely on these gateways for connectivity to the fixed web. An effective infrastructure for end-users must facilitate the migration of connections among gateways to ensure link quality between gateways and devices, as well as to maintain load balance among the gateways. Conversely, it is imperative to process, filter, and aggregate information as near to the source as feasible to minimize superfluous network traffic and decrease reaction times. This research aims to assess how effective BDA systems integrated into smart city systems function as they support real-time decision-making abilities. The study uses genuine Aarhus and Surrey data to illustrate how data filtration and normalization techniques enhance both urban data processing time and system throughput which enables efficient management of traffic systems and parking functions and water usage control. This study dedicates efforts to determine the best threshold values that emerge from analyzing gathered city operation data. The rest of the paper have been systematized as follows: Section II reviews existing literature on autonomous decision-making and real-time data processing in smart urban environments. Section III describes the dataset information, data preprocessing, and data processing approach. A detailed discussion of the findings, which include dataset information, simulation scenario for data analyzing, and data processing performance evaluation, has been provided in Section IV. Lastly, Section V concludes the study and demonstrates enhanced decision-making capabilities that amount to data processing speed in smart cities.

II. LITERATURE REVIEW

Bukhari, Alshibani, and Ali [6] assert that the primary aims of smart cities are to enhance human well-being, foster economic progress, and uphold sustainability. Intelligent urban environments can improve various services, such as agriculture, transportation, education, and health, among others. Intelligent cities are founded on the ICT models, which encompasses IoT technologies. These technologies generate substantial quantities of heterogeneous information, generally termed big data. Nevertheless, they assert that this evidence is devoid of significance in isolation. New methodologies must be established to analyze the vast quantities of data collected, and one potential solution is the implementation of BDA tools. Big data may be analyzed and modeled using analytical approaches to obtain improved insights and enhance the functionality of smart cities. In [7], four advanced big data analytics methodologies are introduced. The authors address the applications of BDA across 5 sectors of these cities and provide a comprehensive background of the security problems associated with BDA and big data in smart cities. As the globe transitions toward the era of smart cities, which are heavily reliant on web and IoT applications. Urban environments are more popular due to their beneficial effects on a nation's economy.

According to Silva, Khan, and Han [8], smart and swift decision-making are essential requirements of an advanced smart city system. This system simultaneously produces several files referred to as big data, characterized by the 3 V’s, resulting in the acknowledgment of a significant issue. Innovative concepts, tactics, and models must be implemented to effectively address the challenges of scheduling and managing big data Their research presents a comprehensive analysis of research conducted on scheduling approaches within the Spark and Hadoop frameworks. Active scheduling is essential for attaining optimal performance in large-scale data processing. The challenges associated with big data integrate data diversity, volume, velocity, privacy and security, connectivity, data sharing, and cost. The review indicates that the baseline is sufficient for processing when the data is static and batch processing may be deferred until completion.

According to Rathore et al. [9], Spark possesses a distinct advantage in real-time data processing through its parallelism capabilities. Extensive research is still required to determine that Spark is the exclusive answer for evaluating real-time streaming data. Furthermore, the study revealed that Spark may assess data rapidly. Spark is a premier memory analysis tool that facilitates real-time streaming information processing on extensive datasets. In comparison to Hadoop, Apache Spark is far more advanced. It accommodates many requirements, such as real-time process, streaming, and batch. Future advancements may enable Hadoop schedule optimization. For Spark, this may be achieved by adjusting default parameter configurations, implementing novel scheduling methodologies, and utilizing hybrid artificial intelligence scheduling.

In [10], sensor networks and the Internet of Things (IoT) represent a contemporary communication framework in the digital realm, poised for significant growth in the near future. In this framework, everyday objects (electronic devices) will be equipped with sensors, actuators, microcontrollers, and artificial intelligence, facilitating digital communication through various networking protocols. This integration will enable these objects to interact with one another and with users, thereby becoming essential components of digitization. The notion of WSN seeks to enhance the Internet's immersiveness and accessibility. Consequently, facilitating seamless interaction and access with an extensive array of devices, including automobiles, mobile phones, vehicle trackers, forest fire detection sensors, humidity sensors, surveillance cameras, and home appliances, will result in the integration of numerous surrounding objects into networks in various forms. Radio Frequency Identification (RFID), sensor technology, and other intelligent technologies will be integrated into diverse applications. The Internet of Things (IoT) will oversee these devices via sensors to facilitate the development of numerous applications that leverage the substantial volume and diversity of data produced by such objects, thereby offering innovative smart services to individuals, businesses, and public administrations.

Anon [11] examined various real-time data analysis systems that have developed to satisfy the increasing requirements of high-performance applications, facilitating prompt decision-making and perspectives in various real-life contexts. They examined the historical evolution, fundamental attributes, and obstacles of these architectures, emphasizing computing models such as Apache Kafka and Flink. The incorporation of innovative technologies, including AI, IoT, and edge computing, is emphasized for augmenting system features. Critical issues such as latency enhancement, error resilience, and expandability are examined by Ma et al. [12], along with suggested solutions including hybrid architectures and event-driven models. They emphasize the necessity for novel models, which integrate minimal delay and high data rate to tackle the difficulties of contemporary applications.

III. METHODOLOGY

This section explains the methodology for evaluating the proposed smart city architecture through data evaluation about processing large datasets in real-time. Available open-data portals provide the study with several datasets which examine different urban management areas including traffic and parking zones and water usage and pollution control.

Dataset Information

This research utilizes genuine public datasets collected from cities Aarhus in Denmark and Surrey in Canada which can be accessed to the public see **Fig. 1**. These datasets cover a range of parameters critical to smart city operations. **Table 1** below presents detailed requirements regarding the performance datasets.

Table 1. Data Related to Analyzed Datasets

City	***Dataset***	***Size***	***Sensor/Source Details***	***Collection Methods***	***Analysis/Use Case***
Aarhus, Denmark	Traffic	3.04 GB	Vehicle tracking systems put together with roadside sensors gather the data.	Sensors measure vehicular density as well as speed and flow records data points after each triplet of 122 seconds	The device helps authorities check traffic congestion while detecting busy hours then generates instant route suggestions for different paths.
	Parking Spaces	0.20 MB	The urban parking facilities use parking sensors for space availability monitoring functions.	The wireless IoT sensors shout parking empty and full data instantly to central command centers.	Helps citizens find empty parking spots which lowers search time while decreasing city traffic congestion.
	Pollution	77.25 MB	The pollution sensors used for air quality monitoring track down pollutants consisting of Ozone (O_3), Nitrogen Dioxide (NO_2), and Carbon Monoxide (CO).	Air-quality sensors periodically transmit data across multiple distribution points from urban to suburban areas.	Analyzes pollution levels then evaluates resulting health risks before creating public warning messages to inform citizens about air conditions.
Surrey, Canada	Water Usage	4 MB	The 61,263 homes participate in the program by using smart meters to track their water consumption.	Regular monitoring of flow and consumption data takes place at every residential property through smart meters.	Sends threshold alerts for overused water supplies which helps authorities create water conservation strategies and policies.

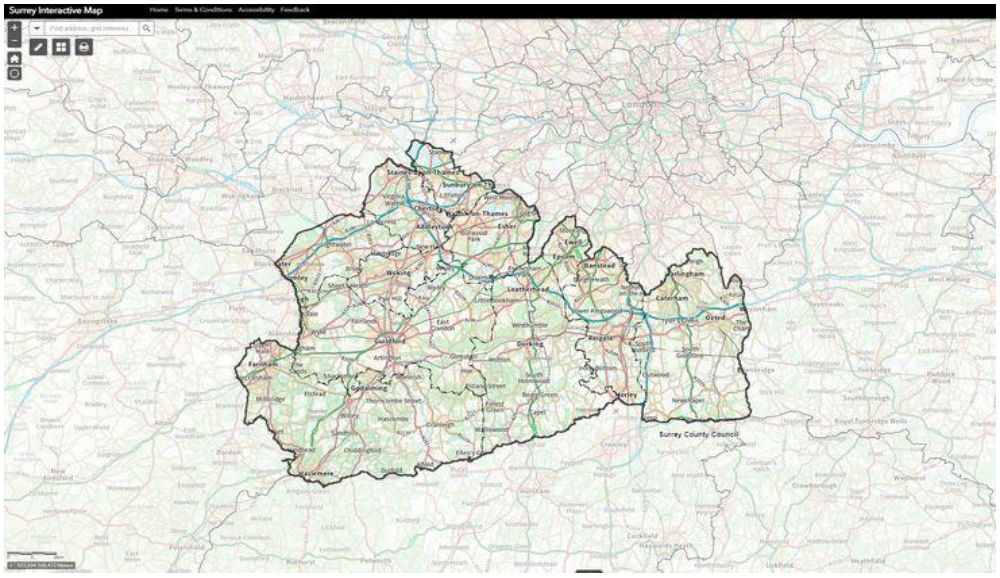

Fig 1. Surrey City Map

Multiple urban components contain embedded sensors which provide data for these datasets. The water and energy consumption data from Surrey contains information derived from 61,263 smart homes. Through Aarhus City sensors accumulate data regarding road traffic and parking with pollution measurements see **Fig. 2**. The available datasets contain information about toxic gas concentration levels of Ozone (O_3) Nitrogen Dioxide (NO_2) and Carbon Monoxide (CO) [13].

Data Preprocessing

The collected real-time data suffered from data noise which frequently occurs during point-of-origin collection tasks. The processing performance received improvement through the implementation of data filtration along with min-max normalization techniques. The preliminary data processing steps remove inaccurate and unclear data to enhance both system accuracy and processing speed [14]. The filtration method removes excessive data while normalization enables data scales to match each other for analytical purposes.

Simulation Setup and Performance Analysis

The Hadoop cluster processing the datasets operates on an Ubuntu 16.04 LTS system that has an Intel Core i5 processor paired with 8 GB RAM. The proposed architecture depends heavily on the Hadoop framework because it facilitates the processing of extensive data volumes. The Wireshark library executed real-time traffic generation through its ability to build traffic from capture files that transmitted toward the Hadoop cluster. An analysis of the pcap-formatted data took place through Hadoop-pcap and Hadoop-pcap-scr-de libraries. The network packet processing functionality produces Sequence Files from which MapReduce and Spark can extract offline and real-time analysis benefits. The experiment operated with three various buffers assigned to reception whose sizes were identified as minimum (4096 bytes), maximum (4,001,344 bytes), and default (87,370 bytes). The data transmission process became sequential because it depended on buffer availability.

Event generation and threshold limits

Event generation performance together with processing times and event generation times and user notification durations were measured across different datasets. **Table 2** shows threshold values for different datasets that determine when event generation should start.

The threshold values in **Table 2** indicate the exact points in time or conditions for event generation to occur. The variable θ represents the complete duration that spans data processing alongside event generation and user notification. The processing duration displays proportional growth according to dataset size but streaming data real-time processing remains unaffected by this increase. The system needs to maintain a fast capability for handling the quick production of data streams.

Data Processing and Analysis

The assessment of each dataset sought to determine its influence on smart city operational effectiveness and decision-making capabilities. Information from Aarhus Road traffic sources in Denmark served to identify congestion patterns. Every 122 seconds periodic measurements exist in the dataset for determining vehicular congestion threshold values. The system transforms analysis results into autonomous operations at the same time it delivers traffic congestion updates coupled with alternative routing suggestions to drivers through real-time functionality. Through its capabilities the system identifies empty parking positions in Aarhus which automatically notifies residents about vacant spots while eliminating the requirement for human observation.

Urban water management received assessment through analysis of water consumption data from Surrey Canada. Each household in Surrey consumes an average of 60,000 liters every month. The established threshold points enabled authorities to run alerts for excessive water usage thus creating opportunities for municipalities to establish conservation strategies. The pollution analysis compiled data from Aarhus about Ozone (O_3), Nitrogen Dioxide (NO_2) and Sulphur Dioxide (SO_2) to monitor environmental contamination and its resulting health consequences. NO_2 concentrations emerged as the highest pollutant because of vehicular emissions which negatively affect health conditions. Public members received real-time pollution reports through notifications in order to enhance health decision making and promote public awareness.

Table 2. Event Generation Efficiency and Threshold Limits

Dataset	***Size***	***Threshold (θ)***	***Threshold Context and Criteria***	***Practical Applications***
Water Usage	4 MB	80 Cubic Liters	The system will activate an alert if any household consumes more than 80 cubic liters through their daily water usage.	Municipalities utilize this system to both start conservation practices and detect overconsumption and waste water problems.
Traffic	3.04 GB	Varies with time	The system creates events through traffic density threshold changes at precise times or spatial positions.	Users receive instant updates about traffic jams followed by proposed routes that enhance overall traffic performance.
Pollution	77.25 MB	80%	When air quality reaches 80% beyond the suggested limit for safety the system will activate warning alerts.	Enables warnings about public health as well as sustainable urban plan development and pollution mitigation decisions.
Parking Spaces	0.20 MB	<10/parking lot	Users receive alerts when any parking facility has less than ten empty parking spaces.	Helps drivers find empty parking spots while cutting down search duration and lowering city traffic congestion.

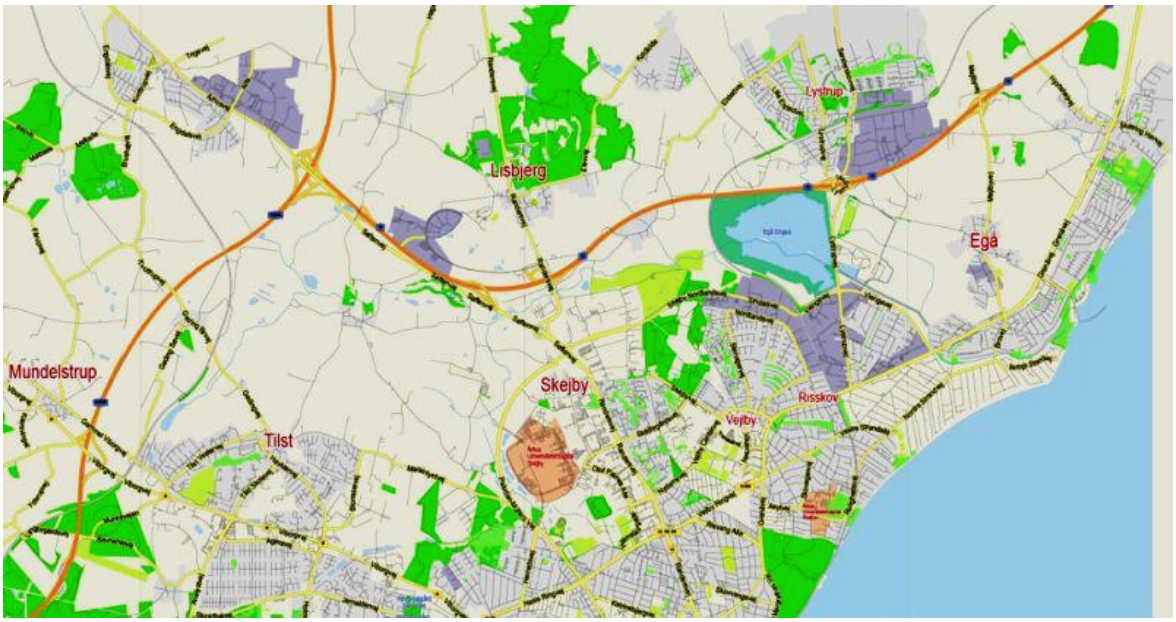

Fig 2. Aarhus City Map

Comparison of Processing Time and Throughput

The evaluation of our proposed data processing system involved seeking performance comparisons between its time efficiency and data handling capacity with existing standard data processing approaches. The data processing system achieved better processing times along with improved throughputs by implementing data filtering and normalization approaches. The processing data reveals throughput along with processing time for Spark and MapReduce systems after adding filtration methods as well as filtration techniques. The proposed method outperforms existing works by delivering better processing time and throughput for large datasets. Throughput measures demonstrate that the proposed approach excels beyond other methods when analyzing large datasets despite general throughput growth across all systems.

IV. RESULTS AND ANALYSIS

This study presents a smart city infrastructure engineered for real-time data processing, enhancing autonomous decision-making aptitudes. The analysis of extensive datasets from several fields enhances the efficiency of urban operations. The

genuine datasets referenced in **Table 1** were retrieved to evaluate data computation efficacy. The first datasets were unfiltered and checked data. Consequently, range scaling and data cleansing approaches were implemented to increase the processing efficiency of huge datasets.

Dataset Information

Genuine statistics for energy usage, water usage, parking availability, pollution metrics, and vehicular traffic, were sourced from publicly accessible databases [15]. Intelligent meters from 61,263 residences in Surrey, Canada, have gathered data on water and energy usage. Sensors installed in Aarhus, Denmark, collected data on pollution levels, parking facilities, and urban traffic. Pre-established sensor sets were utilized to assess road congestion between these two locations. Sensors installed in designated parking spots of Aarhus city collected data on occupied parking spots and accessibility. Concentrations of toxic gases, namely Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), and Ozone (O_3) are documented in the pollution quantities dataset for urban suburbs. All these datasets utilized in the evaluation of data analysis are verified and publicly accessible. The Survey's Open Governmental License, encompasses energy and water datasets. The datasets for City Pulse EU-FP7 development, including pollution data, parking spot data, and traffic data, are meaningfully published and tagged under the International License Attribution 4.0 of the Creative Commons.

Model Scenario for Data Analysis

A machine running Ubuntu 16.04 LTS, equipped with a Core i5 CPU and 8 GB of RAM, is utilized to manage the cluster of Hadoop, which evaluates the mentioned datasets. Wireshark libraries generated real-time traffic data. We utilized Wireshark's traffic data generator features to produce data from capture files. Consequently, the produced data was redirected to the established Hadoop clusters. Data in pcap set-up was evaluated with the Hadoop-pcap-scr-de and Hadoop-pcap packages. Hadoop-pcap-scr-de and Hadoop-pcap handles system packets and produce Sequence Files. The production of Sequence Files is beneficial due to its compatibility with offline evaluation with real-time analysis and MapReduce through Spark. About 3 buffer dimensions were established for receipt end: minimum (4,096 bytes), maximum (4,001,344 bytes), and default (87,370 bytes). Information transmission occurred sequentially and was contingent upon the presence of the receipt buffer. **Fig. 3** depicts the arrangement situation for the preliminary dataset evaluation and system efficiency assessment.

The assessment of data processing validated that the suggested BDA-integrated urban framework possesses. The examination of data analysis established that the suggested BDA-integrated urban architecture has enhanced information data rate and execution time, while identifying the threshold figures for every city information parameter. The results indicate that θ grows with the amount of the dataset. Nonetheless, real-time analysis focuses on streaming information that is minimally affected by the size of the dataset. Nonetheless, a model with elevated efficiency is crucial for managing swift data generation.

Data Assessment and Processing Efficiency Evaluation

The implementation of an intelligent city is heavily reliant on timely decision-making, facilitated by real-time big data computation [16]. Autonomous smart decision-making aids municipal authorities in delivering superior services to smart city populace. This section discusses dataset evaluation, dataset computation efficiency, and the significance of dataset evaluation in urban organization. Efficiency of data processing and the significance of data evaluation in urban organization. Traffic dataset is essential for effective smart transportation management. Traffic flow data collected from Aarhus; Denmark has been studied to determine threshold levels for vehicle traffic.

Dataset encompasses road congestion data for a distance of 3655 meters between the observation points of Aarhus and Hinnerup. The collection comprises data collected at regular intervals of 122 seconds. Traffic data are essential for effective smart transportation management. Traffic congestion levels vary throughout the day. The findings of the vehicular data evaluation are portrayed in **Fig. 4** and **5**. illustrates fluctuations in vehicular mass for each observation.

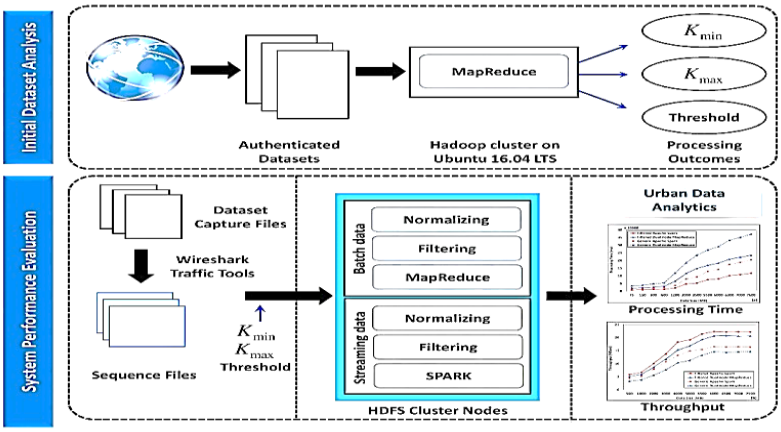

Fig 3. Experimental Setup for Preliminary Data Processing and System Evaluation

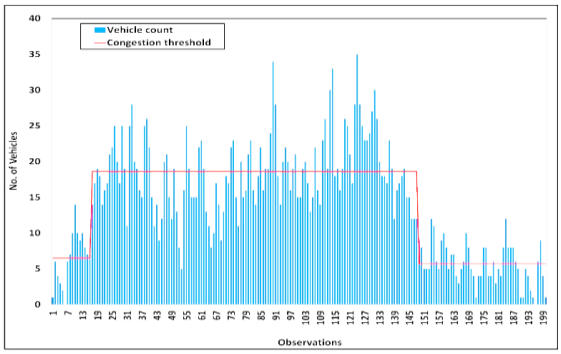

Fig 4. Vehicular Mass Evaluation of Aarhus, Denmark.

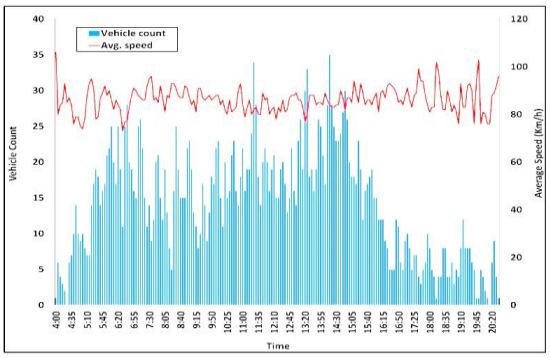

Fig 5. Vehicular Velocity Analysis of Aarhus, Denmark.

The smart city architecture utilizes proposed data analysis technologies to produce real-time decisions, including traffic congestion data for vehicles. The threshold for congestion levels fluctuates over time to facilitate more realistic decision-making. Upon detecting a vehicle amount that exceeds the limit of congestion for that period, the SCSC informs potential drivers appropriately. In accordance with the analytical findings, SCSC can accurately ascertain road traffic congestion, thus facilitating automated event production. The generated events are sent individually from intelligent transportation elements to their respective service layer elements. Consequently, the road traffic control service disseminates traffic alerts to all prospective vehicles and recommends alternative pathways to circumvent crowded path segments. On one side, parking management is a laborious duty for municipal authorities. Conversely, locating adequate parking places in metropolitan settings poses significant challenges for city dwellers. The smart parking feature integrated within SCSC manages parking lot dataset in real-time and alerts city inhabitants to present parking lots. As a result of the advantages of smart parking systems, citizens will efficiently locate required parking lots without longer searches.

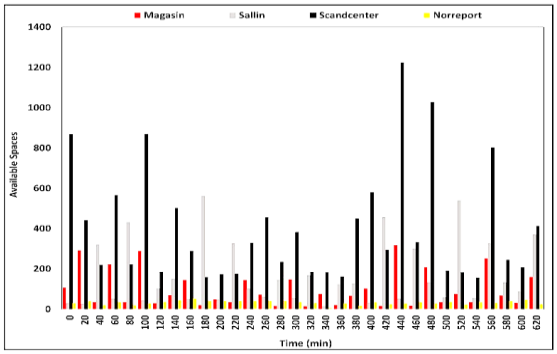

Fig 6. Accessibility of Parking Spots of Aarhus

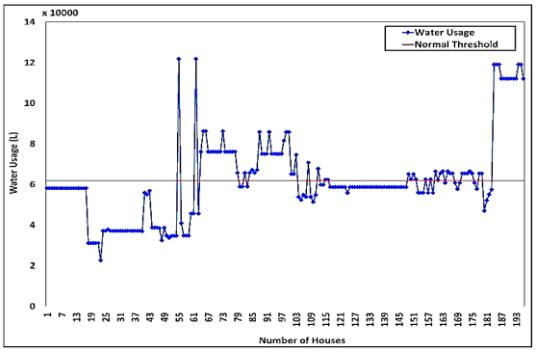

Fig 7. Evaluation of Water Usage of Surrey, Canada

The system evaluated the parking information of Aarhus and revised the parking data on Hadoop through the transportation element at SCSC. Inhabitants can identify empty parking lots by using the parking organization sub-element of the intelligent transportation system. The information collecting layer changes the availability of the parking space instantaneously upon occupation and release. **Fig. 6** illustrates the parking availability across several parking facilities in Aarhus. Parking spots have significantly diminished at midday. Alongside a parking spot searching service, parking control may be enhanced to facilitate parking reservations.

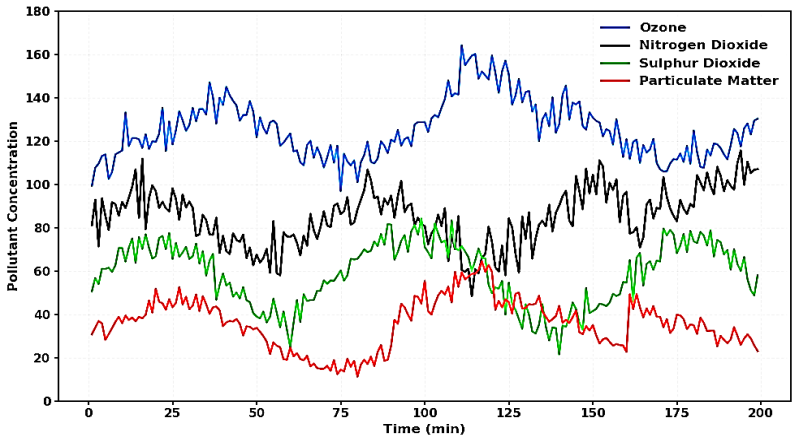

Fig 8. Pollution Dimensions Assessment of Aarhus, Denmark

Nevertheless, time limitations should be implemented for the parking booking service to prevent unoccupied parking lots. City water usage has surged significantly, attributable to human activities stemming from rapid industrialization and urbanization [17]. Water usage data from Surrey was evaluated to identify possible remedies for future city water control. **Fig. 7** illustrates that the average monthly water use in a typical household is approximately 60,000 liters. The water consumption threshold was derived from an analysis of the water usage data from Surrey. The smart brokers issue alerts to inform consumers when water use over the established threshold figure. Beyond regulating domestic water usage, municipalities may implement innovative methods for water reuse to satisfy the requirements of the urban populace in the future.

Rapid urbanization leads to heightened waste output, noise pollution, and air pollution, which have profound implications for human health and environmental sustainability. To ascertain dangerous pollutant amounts, we examined the pollution measurements data from Aarhus. The investigation of changes in O_3, NO_2, SO_2, and particulate matter concentrations throughout the day is depicted in **Fig. 8**. NO_2, a byproduct of vehicle traffic, exhibits maximum concentration. Prolonged exposures to NO_2 results in significant medical repercussions, including lung cancer. Elevated quantities of O_3 enhance the greenhouse effect, which is associated with global warming. Furthermore, O_3 interactions with skin lipids induce dermal and respiratory inflammations. The planned scheme's real-time data analysis capabilities allow urban residents to ascertain pollution concentrations in urban environments. Pollution alerts are managed by intelligent meteorological and healthcare systems.

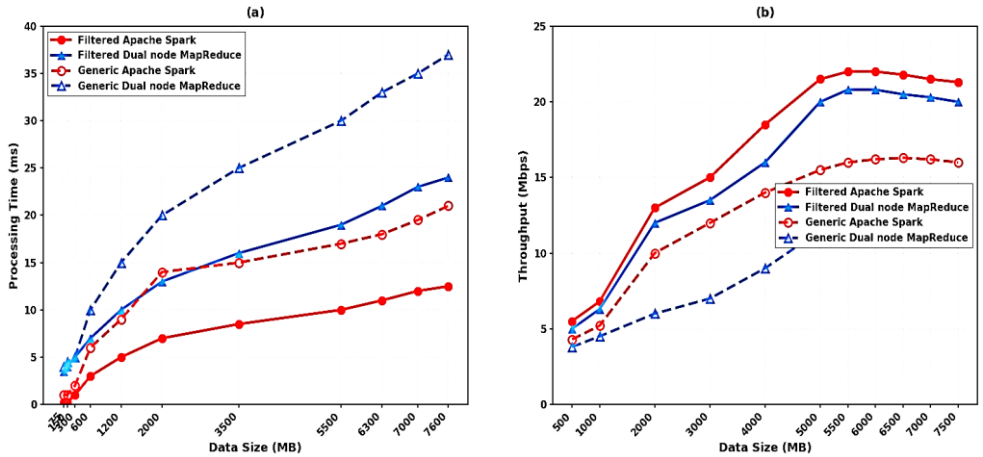

Fig 9. Filtration Integrated Processing Efficiency. (A) Processing Duration, (B) Output Rate

The throughput and processing time of real-time and proposed historic data processing were evaluated against standard data evaluation on dual-node Hadoop MapReduce and Spark. Grouped allocation, data normalization, and data filtration strategies have diminished processing duration while enhancing throughput. The elimination of confusing and erroneous data from the evaluation dataset during normalization and filtration processes has enhanced the efficiency of the suggested strategy. **Fig. 9(a)** illustrates the examination of data processing times for both offline and real-time processing. The implementation of filtration techniques has markedly decreased the processing duration for both MapReduce and Spark. The throughput improvement achieved by the suggested approach is illustrated in **Fig. 9(b)**. The findings indicated that throughput of both MapReduce and Spark augmented with the size of data.

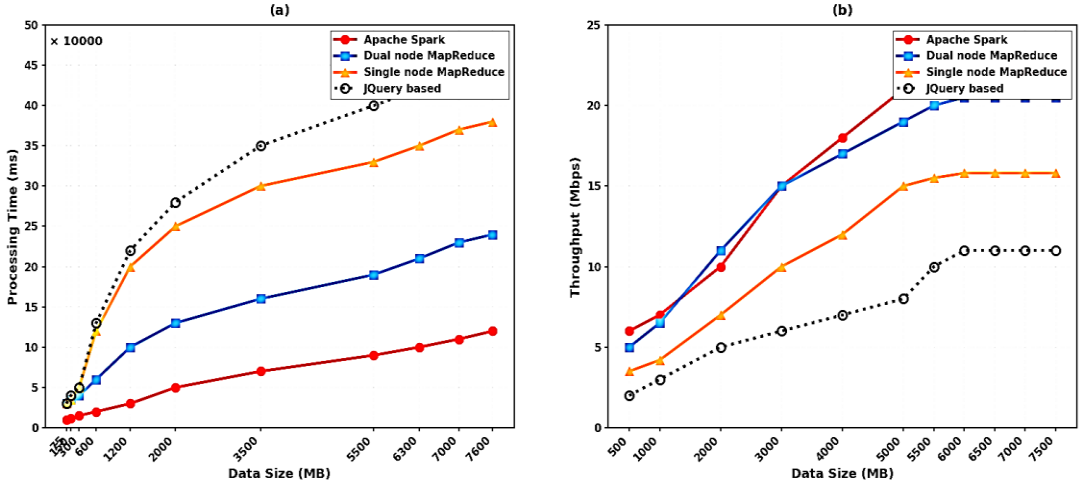

Fig 10. Comparative Analysis of Mapreduce and Spark Performance on A Dual-Node Configurations of Hadoop. (A) Processing Duration, (B) Output Rate

This study utilized MapReduce on Apache Spark and dual-node Hadoop cluster for real-time and offline data analysis, correspondingly. The use of MapReduce on dual-node clusters has enhanced processing efficiency due to effective data dispersion. **Fig. 10(a)** illustrates the data analysis duration for JQuery-based, single-node Hadoop MapReduce, dual-node Hadoop MapReduce, and Apache Spark data computation. The increase in processing time is swift with JQuery-based and single-node Hadoop processing. Prolonged processing duration impedes the real-time functionality of the intelligent city, resulting in reduced decision-making efficacy. The throughput enhancement of the recommended approach is contrasted with MapReduce on a JQuery-based and single-node Hadoop system, with the findings depicted in **Fig. 10(b)**. **Fig. 10(b)** illustrates that quantity has risen in correlation with the size of data. We utilized a 4G Core i5 3.6 GHz CPU featuring 4 cores.

Single-core programs typically exhibit a consistent throughput, while additional cores remain inactive and contribute nothing significant. A single processor maintains the same throughput, as it is unable to allocate tasks and does not facilitate parallelism. Applications with multiple cores such as Hadoop implement parallelism by utilizing several available cores to divide the processing load, hence optimizing the use of available cores. Consequently, parallel distribution enhances throughput in relation to both data size and core usage. With minor datasets, throughput diminishes due to underutilization of the cores. However, at a certain juncture, cores attain the maximum occupation threshold. From that moment forward, throughput became steady. All the 4 approaches provide comparable throughput levels for smaller datasets. The diminished data processing efficiency of JQuery system and single-node Hadoop is evident in throughput evaluation findings.

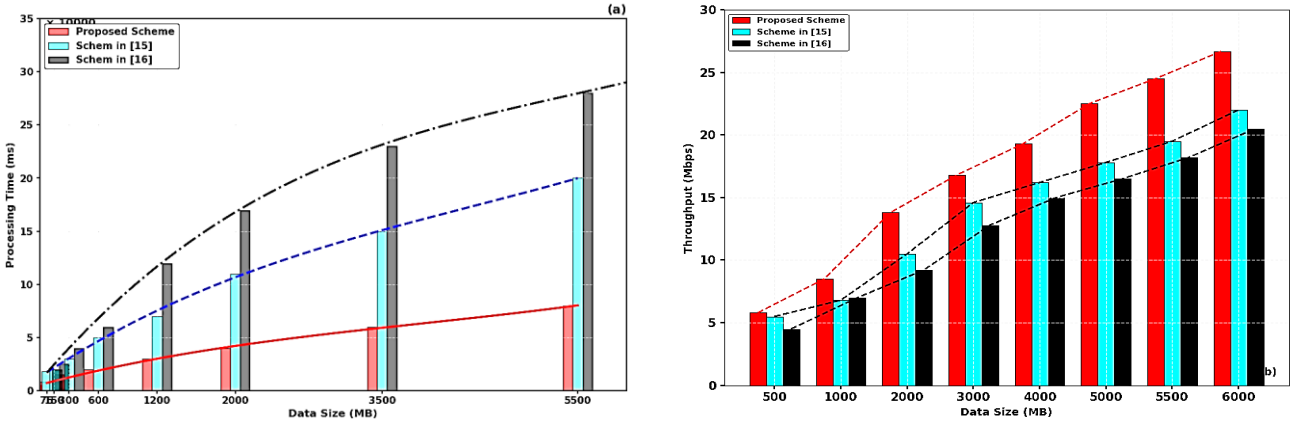

Fig 11. Comparison of The Performance of Filtered Spark Evaluation with Current Methodologies. (A) Analysis Of Processing Duration in Relation to Data Expansion, (B) Analysis of Throughput in Relation to Data Increase

Fig 11 depicts the efficiency evaluation of the suggested system relative to one of our prior studies [18] and a study by Fleckinger [19]. **Fig. 11(a)** illustrates the analysis of processing time across different data sizes, reaching up to 5500 MB. All the 3 methodologies evidently prolong processing duration as data size increases. Nonetheless, the proposed study has achieved the optimum dataset processing rate in comparison to the studies reported in [20]. **Fig. 11(b)** illustrates that the method in [21] has achieved a maximum throughput of 2000 MB. Nonetheless, the anticipated work has advanced with datasets over 2000 MB. The comparison clearly demonstrates that throughput gains from dataset expansion are constrained for other approaches, despite the suggested work achieving a substantial throughput increase.

V. CONCLUSION

We demonstrate enhanced decision-making capabilities that result from data processing speed within smart cities by referencing specific use cases. Big data analytics (BDA) through its role within urban management systems implements traffic system and water management optimization to enhance standardized metropolitan conditions. The proposed architecture utilizes Hadoop and Spark tools to execute big datasets thus enabling the processing of raw and erratic data. The execution of data filtration and normalization produces vital performance strengths which result in elevated processing capability together with shortened durations. Real-time decision-making processes help organizations deal with rapid urban transformations to deliver improved resource management systems while sustaining urban growth. The effective execution of scalable data processing systems is necessary to control present and anticipated urban requirements that improve smart city operations. Proposed research aims at examining advanced merges between machine learning with IoT smart technologies to construct superior evaluation frameworks for the future.

CRediT Author Statement

The author reviewed the results and approved the final version of the manuscript.

Data Availability
The datasets generated during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interests
The authors declare that they have no conflicts of interest regarding the publication of this paper.

Funding
No funding was received for conducting this research.

Competing Interests
The authors declare no competing interests.

References
[1]. R. C. Alvarado, "Data Science from 1963 to 2012," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arxiv.2311.03292.
[2]. M. R. Islam et al., "Smart Parking Management System to Reduce Congestion In Urban Area," 2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), pp. 1–6, Nov. 2020, doi: 10.1109/icecie50279.2020.9309546.
[3]. A. K. Paschalidou, P. Kassomenos, and F. Chonianaki, "Strategic Noise Maps and Action Plans for the reduction of population exposure in a Mediterranean port city," The Science of the Total Environment, vol. 654, pp. 144–153, Nov. 2018, doi: 10.1016/j.scitotenv.2018.11.048.
[4]. S. Rusitschka, K. Eger, and C. Gerdes, "Smart Grid Data Cloud: A Model for Utilizing Cloud Computing in the Smart Grid Domain," 2010 First IEEE International Conference on Smart Grid Communications, Oct. 2010, doi: 10.1109/smartgrid.2010.5622089.
[5]. M. Mahbub, "IoT ecosystem: functioning framework, hierarchy of knowledge, and intelligence," in Internet of things, 2022, pp. 47–76. doi: 10.1007/978-3-030-87059-1_2.
[6]. A. Bukhari, S. M. Alshibani, and M. A. Ali, "Smart City as an Ecosystem to Foster Entrepreneurship and Well-Being: Current state and Future Directions," Sustainability, vol. 16, no. 24, p. 11209, Dec. 2024, doi: 10.3390/su162411209.
[7]. K. Soomro, M. N. M. Bhutta, Z. Khan, and M. A. Tahir, "Smart city big data analytics: An advanced review," Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, vol. 9, no. 5, Jun. 2019, doi: 10.1002/widm.1319.
[8]. B. N. Silva, M. Khan, and K. Han, "Big Data Analytics Embedded Smart City Architecture for Performance Enhancement through Real-Time Data Processing and Decision-Making," Wireless Communications and Mobile Computing, vol. 2017, pp. 1–12, Jan. 2017, doi: 10.1155/2017/9429676.
[9]. M. M. Rathore, H. Son, A. Ahmad, A. Paul, and G. Jeon, "Real-Time Big Data Stream Processing Using GPU with Spark Over Hadoop Ecosystem," International Journal of Parallel Programming, vol. 46, no. 3, pp. 630–646, Jun. 2017, doi: 10.1007/s10766-017-0513-2.
[10]. K. P. Seng, L. M. Ang, and E. Ngharamike, "Artificial intelligence Internet of Things: A new paradigm of distributed sensor networks," International Journal of Distributed Sensor Networks, vol. 18, no. 3, p. 155014772110628, Mar. 2022, doi: 10.1177/15501477211062835.
[11]. S. Anon, "A comprehensive analysis of real-time data processing architectures for high-throughput applications," SSRN Electronic Journal, Jan. 2025, doi: 10.2139/ssrn.5034117.
[12]. Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H. V. Poor, and B. Vucetic, "High-Reliability and Low-Latency Wireless Communication for Internet of Things: challenges, fundamentals, and enabling technologies," IEEE Internet of Things Journal, vol. 6, no. 5, pp. 7946–7970, Mar. 2019, doi: 10.1109/jiot.2019.2907245.
[13]. S. A. Buthelezi and T. C. Davies, "Carbon monoxide (CO), ozone (O3) and nitrogen dioxide (NO2) exposure from vehicular transportation and other industrial activities in the vicinity of Umlazi Township, South of Durban, KwaZulu-Natal Province, South Africa," Transactions of the Royal Society of South Africa, vol. 70, no. 3, pp. 277–283, Jul. 2015, doi: 10.1080/0035919x.2015.1046972.
[14]. C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building Operational data," Frontiers in Energy Research, vol. 9, Mar. 2021, doi: 10.3389/fenrg.2021.652801.
[15]. G. Oh, D. J. Leblanc, and H. Peng, "Vehicle Energy Dataset (VED), a Large-Scale Dataset for Vehicle Energy Consumption research," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 4, pp. 3302–3312, Nov. 2020, doi: 10.1109/tits.2020.3035596.
[16]. R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A Survey," International Journal of Information Management, vol. 45, pp. 289–307, Sep. 2018, doi: 10.1016/j.ijinfomgt.2018.08.006.
[17]. V. Saxena, "Water Quality, Air Pollution, and Climate Change: Investigating the environmental impacts of industrialization and urbanization," Water Air & Soil Pollution, vol. 236, no. 2, Jan. 2025, doi: 10.1007/s11270-024-07702-4.
[18]. C. Chen, J. M. Ricles, T. L. Karavasilis, Y. Chae, and R. Sause, "Evaluation of a real-time hybrid simulation system for performance evaluation of structures with rate dependent devices subjected to seismic loading," Engineering Structures, vol. 35, pp. 71–82, Jan. 2012, doi: 10.1016/j.engstruct.2011.10.006.
[19]. P. Fleckinger, "Correlation and relative performance evaluation," Journal of Economic Theory, vol. 147, no. 1, pp. 93–117, Nov. 2011, doi: 10.1016/j.jet.2011.11.016.
[20]. N. M. Nawi, W. H. Atomi, and M. Z. Rehman, "The effect of data pre-processing on optimized training of artificial neural networks," Procedia Technology, vol. 11, pp. 32–39, Jan. 2013, doi: 10.1016/j.protcy.2013.12.159.
[21]. B. Silva et al., "Urban planning and smart city decision management empowered by Real-Time data processing using big data analytics," Sensors, vol. 18, no. 9, p. 2994, Sep. 2018, doi: 10.3390/s18092994.

Publisher's note: The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. The content is solely the responsibility of the authors and does not necessarily reflect the views of the publisher.