Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 4123–4131

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

Data Architecture and Big Data Analytics in Smart Cities

El Mehdi Ouafiq[1*], Mourad Raif[1], Abdellah Chehri[2], Rachid Saadane[1]

[1] *SIRC-LaGeS, Hassania School of Public Works, Casablanca, Morocco*
[2] *Department of Applied Sciences, University of Quebec, Chicoutimi, QC, Canada, G7H 2B1*
elmehdiouafiq@gmail.com; mourad.raif@yahoo.fr; achehri@uqac.ca; rachid.saadane@gmail.com

Abstract

The smart city has become a persistent need and is no longer just a concept. The concept of smart cities heavily relies on collecting enormous amounts of data. This paper proposes a data-management-based solution for smart city, which is labeled Smart Systems Oriented Big Data Architecture. Big data technologies have become essential to the functioning of cities. The architecture includes complex components to be implemented based on the architectural requirements. A data migration strategy was proposed to handle the various data sources such as IoT devices, video cameras, and drones. The proposed approach also takes into account data processing and data storage. The technical constraints related to data processing in a big-data environment are also studied. We also consider data modeling from a business intelligence point of view and a data science perspective. Our main goal is to favor the facilitation of the daily life practices in the context of a smart city by providing the city administrators with a solution that helps them maintain their city smartly and effectively.

Keywords: Smart cities, big data analytics, Drone, IoT, Data Management.

1. Introduction

The concept of Smart Cities appeared in 1992. It is used to represent the cities of the future, whose objective is to improve the functioning of these cities both economically, of management, or even the ecological plan, from a sustainable development perspective. The purpose of Smart Cities is also to bequeath to future generations something

* Corresponding author.
Email address: elmehdiouafiq@gmail.com

10.1016/j.procs.2022.09.475

sustainable to allow them to build their future on a good foundation. The intelligent technologies implemented in these smart cities are aimed at residents, and the objective is to improve comfort, their standard of living, or even their safety. As stated by [1] the issues addressed by sustainable urban development strategies are:

• The economic development of the city, including the creation of jobs.

• The creation of an environment conducive to the creation of new businesses.

• Improving the level of education in the city.

• Increasing the quality of life in the city with the improvement and maintenance of green spaces, the quality of medical and social support services, and making life safer on the streets of the city.

• The modernization and expansion of the technical and IT infrastructure of the city.

• Increasing the leisure offer in the city and improving its tourist attractiveness.

As the foresight scenario above demonstrates, smart cities are brittle architectures. From technological, social and governance points of view, they have multiple points of failure with cascading, systemic effects. Several definitions of Smart Cities have been studied [2]-[3], to conclude that the axes around which Smart Cities are developing are (1) industry, (2) academic, and (3) the governmental aspect.

The "Smart" aspect takes several forms in these areas. For the industry, it is a question of designing more innovative products or services and integrated technology for technology. For the academic, it is about improving existing technologies. Finally, for the governmental aspect, which may seem unclear, refers to the concept of smart, sustainable growth to prevent urban sprawl, for example.

Six basic components for making a Smart City have been identified [1]:

• An intelligent governance system with transparent information exchange between residents, the city, municipal services, and emergency services (police, firefighters).

• An intelligent economy, enabling an efficient flow of products, services, and knowledge at the city level and between cities.

• Smart mobility is an interconnected, secure, and efficient system for the management of transport, logistics, parking lots, and public transport.

• The intelligent environment is an innovative resource management system such as devices for energy storage, reducing energy consumption, power supply management, smart lighting system, renewable energy development, or waste management.

• Smart residents have access to education and training through modern telecommunication and information technologies. In addition, the accompaniment of residents in terms of resource creativity of human potential encourages the active participation of residents in the life of the city.

• An intelligent lifestyle, allowing to improve the quality of life, to develop better health services and infrastructures, but also enlarge and diversify the culture and service offered.

In this research paper, we explore the different intelligent cities' challenges. Furthermore, we expose NiFi solution for dataflow between various systems within a smart city. The NiFi was built to automate the flow of data and manage the flow of information between systems. NiFi's fundamental design concepts closely relate to the main ideas of Flow-Based Programming. These challenges can be resumed into two categories [1], as described in Table 1.

Table 1. Smart City Challenges – Global View.

Perspective	Requirement
Business	The cities supposed to be more productive and considers the logistic facilities at the first level.
Public	The Smart-City should remedy to the Global-Problems difficulties, e.g. • Pandemic. • Crime and Social disorder. • Cyber Security Risks. • Traffic and Transportation.

2. Connectivity, Data Analytics in Smart Cities

Technological innovations facilitate the implementation of smart city infrastructure. The lifeblood of a smart city, data and connectivity power enables everything from digital twins, fiber networks, and the Internet of Things (IoT), 4G, 5G, robotics, cloud, edge computing, and artificial intelligence (AI), machine learning [4]. The primary purpose of a Smart City is to collect data from drones, video cameras, and IoT sensors, which the relevant departments of the city will analyze to make "quick or preventive" decisions. Also, data analytics-based integrated command and control centers can be planned mainly for the missions of Smart Cities. Predictive analytics can be developed to track clusters zones, suspected cases, Ambulances, disinfection services, and quarantined people to provide the latest information based on real-time data monitoring. In every pandemic period, such as Covid-19, the city can be mapped based on a geospatial information system (GIS) from which the city administration can monitor the areas affected and create buffer zones.

Privacy concerns and the integrity of personal data are just part of the debate over smart cities and a crucial part of the intersection between technological vulnerabilities and human-centered and societal dynamics [5].

As described in Table 2, the heart of smart city planning is data collection and data analytics which can help us accelerate the digital transformation to get value from our Smart-Cities. As a matter of fact, a smart city project stands on:

1. Strategy;
2. Data collection;
3. Data regulation and data governance to implement compliance measures to consider the rights and the privacy laws which might restrict certain data collection and the use of it, especially when tracking people's behavior;
4. Data analytics which is the source of truth of the smart-city.

Table 2. Smart City Stakes.

Smart City Stakes	Details
Building blocks	- Drones; - Telecommunications networks (4G, 5G); - Smartphones and IoT sensors; - Big data centers; - AI-Computation to analyze the data and find a solution;
Areas	- Streets; - Traffic signals; - Buildings; Airport
Data Collection	- Data migration strategy from the building blocks; - Data architecture; - Technical components.
Data Analytics Benefits	- More energy efficient transport: e.g., traffic management application when an accident is happening; - Upgraded and smart water supply and waste-management facilities: e.g., sewage leak alarms; - Heating and lighting networks; - Secure space: e.g., crime reduction applications; - Flood control; - Responsive and interactive city administration: e.g., E-government.

Smart Street Lights can provide multiple functions like surveillance cameras, weather data collection for further visualization, management of traffic (also sending signals to facilitate driverless cars), providing chargers for e-vehicles, converting 4G and 5G signals of telecommunication using Wi-Fi hotspots [6]. Figure I describe how many Data Analytics based Smart Street lights can conduct Smart Systems.

Drones will play an essential role in urban environments, offering multiple advantages in various sectors such as crisis management, medical transport, security, safety, and law enforcement [7]-[8]. For example, cities like Singapore and Dubai have already implemented Intelligent Police Systems based on massive networks of video cameras and AI-based solutions from a security perspective. As shown in Figure 1, in our research, from a data analytics perspective, we considered that drones could complement these video-camera networks by providing mobile aerial vision and increasing the responsiveness and efficiency of law enforcement agencies.

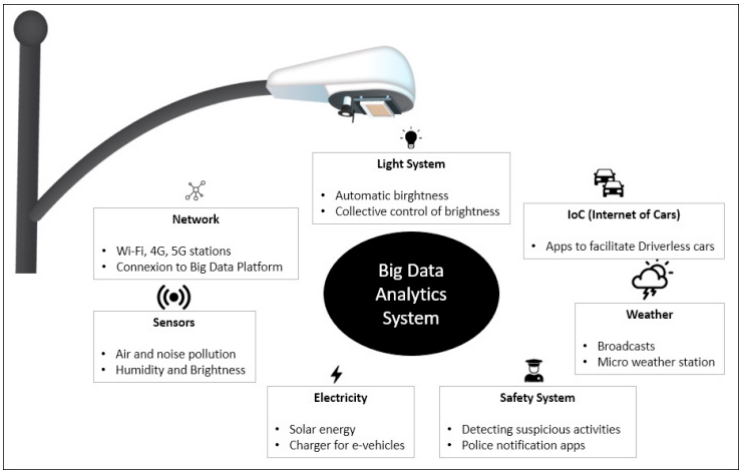


Figure 1. Smart Systems based on Data Analytics

From a Data Analytics perspective, smart cities contain three types of data-nature:

1. DGM data or data generated by machines is mainly collected from Drones, IoT, sensors, smartphones, and other intelligent devices;
2. DHO or data from human origins is primarily generated from social media, smartphones, and other sources;
3. DMP or data mediated by the process mostly in administrations.

As described in Table III, the nature of data should be handled differently in terms of data acquisition, storage, and technical components.

Table 3. Different nature of Smart City's data.

Data Source Nature	Data Source	Data Ingestion Constraints	Data Storage Constraints	Technical Components
DGM	Drones IoT Devices Sensors Video Cameras	Handling micro-batches and Streams. Handling real-time data processing	Satellite Images are stored as separate files. Handling huge amounts of Small-Files in Hadoop, which is made for storing only massive files. Many small files will make the NameNode run out of memory.	NiFi; Kafka Flume Spark Streaming AWS Kinesis Data Streams AWS Kinesis Data Firehose AWS Simple Queue Service
DMP	Relational database management system (RDBMS)	Parallelizing batch data acquisition jobs	A logical and physical Data Model should be defined to store the massive data in a way that can be processed efficiently.	. Sqoop; Spark . Oozie; Data-Factory . AWS Database Migration Service
DHO	Social Media	Handling data	From security perspective this data	. Spark; Python

	and Manual Files	quality	should land on a permanent storage area before pushing into the Data Lake

3. Data Architecture

Under the MapReduce framework of the Hadoop Ecosystem, we can store and process the massive data coming from different sources, guaranteeing more scalability and high availability [9]. Furthermore, as shown in Figure 2, with the Hadoop architecture, the data can be distributed across multiple nodes on a single or multiple on-premise, hybrid, and cloud clusters, based on horizontal scalability.

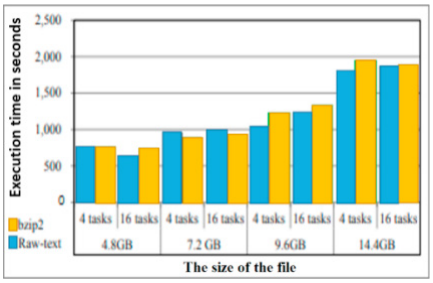
Data will not land permanently on a single server but across multiple ones, which leads to more flexibility in managing the required resources by adding other nodes to the cluster instead of powering one existing node, which is the case of vertical scalability.

In our research, we adapted the best of the existing techniques for Smart Cities and Smart Farming Analytics. As described in Table IV, the proposed architecture, labeled IISOBA, has two vertical layers:

1. Batch-Layer: This is responsible for processing, transforming, and storing the smart city's batch and micro-batch data using Spark, Python, Kyling, HDFS, Hive, and Impala. And then building data workflows using Oozie and airflow in for of actions built as a directed acyclic-graph;

2. Real-Time-Layer: Data pipelines will be built to consume, transform, and store streaming data that will be joined directly with the Batch-Data on the Enrichment-Layer.

Table 4. IISFOBA Architecture

Vertical Layers	Horizontal Layers	Specificity	Issue and Solution
Batch Layer	Shared Area	It is built on top of a network file system in the form of a group of folders structured based on the business logic. Landing-Area for the Ephemeral-Data (DHOs) and the data from less-secure systems (especially DGMs and DHOs) is also considered a gateway of the Hadoop platform.	
	Raw Zone	Where raw data (from DMPs and DGMs) will be stored directly from the sources, and DHOs data will be pushed from the Shared Area into external tables based on an HDFS directory with metadata assigned to it in Hive Metastore to handle the schema on-read.	The bzip2 compression algorithms can be used to save more storage space which saved the storage space 70% times better than a raw text file, with almost a compatible performance capacity. 
	Structured Layer	Data will be cleaned, have a specific structure and data types, and be stored in the form of partitions in Avro format. In this layer, Historical-Data will be	

		stored as per the data vault data model, which proved capable of handling historical data	
	Trusted Zone	The data will be stored based on the business logic defined in a Logical-Data-Model in the form of a Snowflake schema. This layer will be considered the source of truth of the smart-city analytics and the centralized Hadoop Data Warehouse of the Smart City. Since only the needed columns for calculation will be used on the analysis queries, the data files should be stored in Parquet format	
Real-Time & Batch-Layer	Enrichment Layer	It is based on transforming data and enriching it with calculations of the Dashboard KPIs. In this layer, the data will be stored in data marts and accessible via data visualization tools. To enhance the NameNode performance and the memory capacity, after successful execution of the Enrichment-Layer workflow, another workflow is supposed to be executed to delete the raw zone files so that we can lessen the number of metadata handled by the NameNode.	The New Hadoop Archive, which uses a hash table containing the information of indexes and splits it across multiple index files, is highly recommended for this perspective because it outperformed the HAR by 85.47%

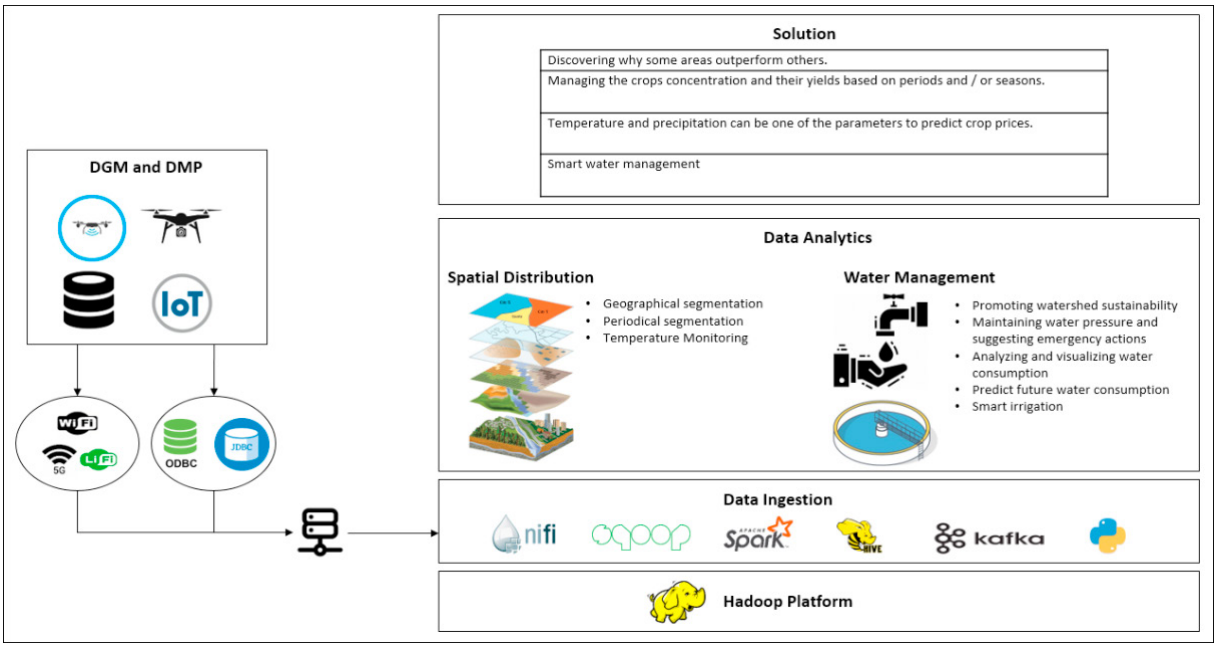


Figure 2. Actionable Smart City based information and services.

There are multiple options to ingest Drones' data based on the location, model, requirements, and networking. The Message Queuing Telemetry Transport (MQTT) can be a better option when the Drone and Data Processing platform exists on the same network / WiFi. The recommended data platform, in this case, should be built on top of a Cloud MQTT server on which UAVs data. Since we include the Data Management aspects in our research, data security

should be considered when sending the data to the protected internal Hadoop Cluster firewall. Thus, on the network file system, which is the kernel of the shared area of the IISFOBA. A cron job can be used by a remote node with an attached sense hat to send the messages of the MQTT to a broker of the cloud-hosted MQTT. As a matter of fact, they can be accessed by offline download, UDP, TCP, and REST.

The primary differences are the scope of complexity, the rate of change necessary to adapt, and that, at scale, the edge case becomes a common occurrence. NiFi is built to help tackle these modern dataflow challenges [10]-[12].

Then, as we suggested in Table 3, NiFi can be used to subscribe to the queue and collect the messages asynchronously. Given the fact that the most flexible packaging file format should be a Json format. Table 5 represents the Data Flow using NiFi [13].

Table 5. Data Flow using NiFi strategy.

Data Flow Phase	Data Ingestion Steps
Drone Image's Data	Directory of Drone images that should be parsed
Data Flow	1. Bring data from the image directory using GetFile; 2. HDFS storage of raw images of Flow-file using PutHDFS; 3. Passing an API Key by adding an attribute using Update Attribute; 4. Json conversion of the body using ReplaceText; 5. Calling the HTTP endpoint using InvokeHTTP; 6. Changing the name of the file ${filename:replace('jpg','json')} using UpdateAttribute; 7. Json file storage in Hadoop using PutHDFS; 8. Collect the values of the attribute using ExtractImageMetadata; 9. Json metadata storage in Hadoop using PutHDFS; 10. Replacing images by Json = attributes conversion into a Json file using AttributesToJson; 11. Changing the name of the file ${filename:replace('jpg','json')} using UpdateAttribute; 12. Json file storage in Hadoop using PutHDFS; 13. Attributes collection from image by Apache Tika using ExtraMediaMetadata; 14. Json file storage in Hadoop using PutHDFS; 15. Resizing the original image to a specific pixel image using ResizeImage; 16. Resized image storage in HDFS using PutHDFS.
Hive DDL	Metadata should be assigned in Hive Metastore to the data that has been put into in HDFS: Create External Table MONITORING([Columns] [Data Types], …) ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe' LOCATION '[The directory of data in HDFS]'
Apache Phoenix Table for Drone Data	Create Table DRONES_INPUT (DATE_ID VARCHAR NOT NULL PRIMARY KEY, [Columns] [Data Types], …);
UPSERT Query	UPSERT INTO DRONES_INPUT (datekey, [VALUES]) values ('${'COLUMN_VALUE'}',...)
Displaying drone data	Data can be displayed by a class Java: package com.dataflowdeveloper; public class DRONE_OUTPUT implements Serializable {}
Image Data Analytics	Machine learning algorithms can then be used to some Data Analytics

e.g. classifying images.

4. IISFOBA Abstraction Layer

An abstraction layer can be built on top of the IISFOBA architecture to handle the schema-on-read problems on the hive data warehouse and data quality simultaneously since we deal with a different source of data. Most of it comes with other data structures.

As described in Figure 3, the kernel of the abstraction layer is a configuration table named ABSCT with parquet file format since it is columnar based. Each column will be solicited for specific usage.

The ABSCT will be responsible for creating DDL (deep data locality) for each file in HDFS (Hadoop Distributed File System) at the RAW-Zone's level, based on the inferred schema from the file and then affect the Metadata to it in Hive Metastore. In addition, the table will be connected to data quality configuration tables to launch technical controls and business rules verifications before ingestion of data from the Raw-Zone to the structured layer.

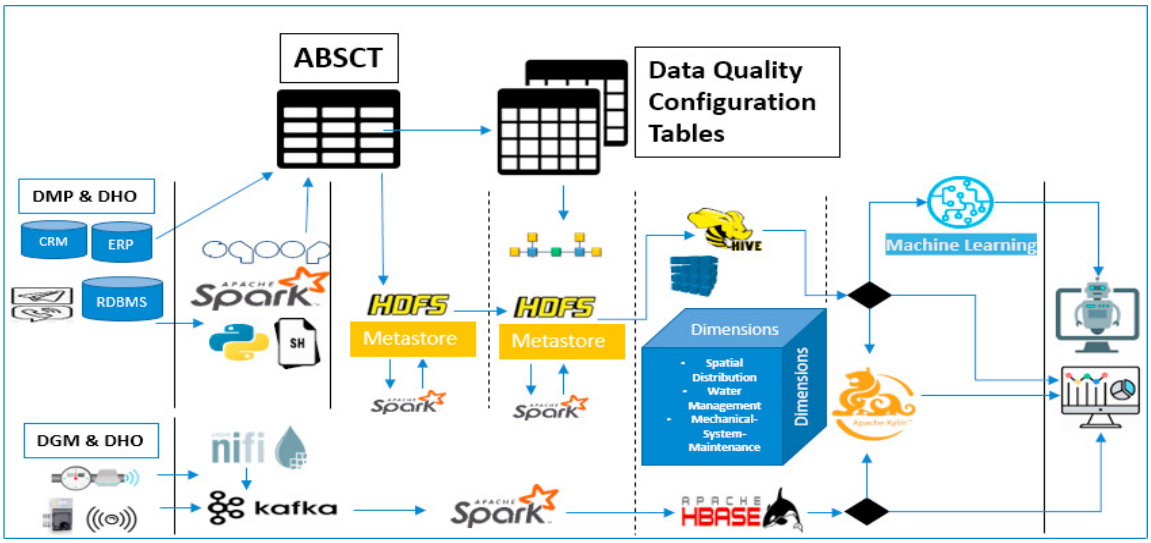


Figure 3. The scalability in Hadoop Architecture

5. Conclusion

Smart cities can be defined as those that effectively integrate physical, digital, and human systems in urban environments to deliver sustainable, prosperous, and inclusive outcomes for their citizens.

In this paper, we have designed what a smart city as a platform can look like, which combines drones, networks, and data analytics performances to deliver digital and intelligent services to make life better in the smart city by engaging stakeholders and citizens straightly.

At the same time, we defined the data migration strategy and the data architecture to process and ingest Smart City's data coming from different sources, especially from drones and IoT devices. The proposed architecture handles the real-time process and should be combined with Batch processing workflows to build the pipelines where data flows from the source to the smart city dashboards and intelligent machine algorithms. A NiFi was built to automate the flow of data between systems. A core philosophy of NiFi has been that even at a very large scale, guaranteed delivery is a must. It is also necessary that the architecture should be a built-in layer to handle data quality, data compression, data security, and data modeling on top of Hadoop's Data Lake. Therefore, from this perspective and in our future research will be continued to optimize the setup of the abstraction layer of the proposed architucture.

References

[1] Barbara Kos. (2019). "Intelligent Transport Systems (ITS) in Smart City," Springer Proceedings in Business and Economics, in: Michał Suchanek (ed.), Challenges of Urban Mobility, Transport Companies and Systems, pages 115-126, Springer.

[2] Mosannenzadeh, Farnaz & Vettorato, Daniele. (2014). "Defining Smart City, A Conceptual Framework Based on Keyword Analysis". TeMA - Journal of Land Use, Mobility and Environment. 683-694. 10.6092/1970-9870/2523.

[3] Chehri, A., Sharma, T., Debaque, B., Duclos, N., Fortier, P. (2022). Transport Systems for Smarter Cities, a Practical Case Applied to Traffic Management in the City of Montreal. In: Littlewood, J.R., Howlett, R.J., Jain, L.C. (eds) Sustainability in Energy and Buildings 2021 . Smart Innovation, Systems and Technologies, vol 263. Springer, Singapore. https://doi.org/10.1007/978-981-16-6269-0_22

[4] Ahmed, Imran & Jeon, Gwanggil & Chehri, Abdellah & Hassan, Mohammad. (2021). Adapting Gaussian YOLOv3 with transfer learning for overhead view human detection in smart cities and societies. Sustainable Cities and Society. 70. 102908. 10.1016/j.scs.2021.102908.

[5] Joe Burton, & Simona Soare. (2020). Smart Cities, Cyber Warfare and Social Disorder. Zenodo. https://doi.org/10.5281/zenodo.6143265

[6] A. Chehri and H. T. Mouftah, "New MMSE Downlink Channel Estimation for Sub-6 GHz Non-Line-of-Sight Backhaul," 2018 IEEE Globecom Workshops (GC Wkshps), 2018, pp. 1-7, doi: 10.1109/GLOCOMW.2018.8644436.

[7] A. Chehri, G. Jeon, I. Fofana, A. Imran and R. Saadane, "Accelerating Power Grid Monitoring with Flying Robots and Artificial Intelligence," in IEEE Communications Standards Magazine, vol. 5, no. 4, pp. 48-54, December 2021, doi: 10.1109/MCOMSTD.0001.2000080.

[8] S. H. Alsamhi, O. Ma, M. S. Ansari and F. A. Almalki, "Survey on Collaborative Smart Drones and Internet of Things for Improving Smartness of Smart Cities," in IEEE Access, vol. 7, pp. 128125-128152, 2019, doi: 10.1109/ACCESS.2019.2934998.

[9] Diaconita, V., Bologa, A. R., & Bologa, R. (2018). Hadoop Oriented Smart Cities Architecture. Sensors (Basel, Switzerland), 18(4), 1181. https://doi.org/10.3390/s18041181

[10] H. Isah and F. Zulkernine, "A Scalable and Robust Framework for Data Stream Ingestion," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 2900-2905, doi: 10.1109/BigData.2018.8622360.

[11] Ouafiq, E.M.; Saadane, R.; Chehri, A. Data Management and Integration of Low Power Consumption Embedded Devices IoT for Transforming Smart Agriculture into Actionable Knowledge. Agriculture 2022, 12, 329. https://doi.org/10.3390/agriculture12030329

[12] El Mehdi Ouafiq, Rachid Saadane, Abdellah Chehri, Seunggil Jeon, AI-based modeling and data-driven evaluation for smart farming-oriented big data architecture using IoT with energy harvesting capabilities, Sustainable Energy Technologies and Assessments, Volume 52, Part A, 2022, 102093, ISSN 2213-1388, https://doi.org/10.1016/j.seta.2022.102093.

[13] A. Pandya et al., "Privacy preserving sentiment analysis on multiple edge data streams with Apache NiFi," 2019 European Intelligence and Security Informatics Conference (EISIC), 2019, pp. 130-133, doi: 10.1109/EISIC49498.2019.9108851.