scientific reports

Check for updates

OPEN

AI edge cloud service provisioning for knowledge management smart applications

Antonio Maciá-Lillo✉, Higinio Mora, Antonio Jimeno-Morenilla, Nahuel E. García-D'Urso & Jorge Azorín-López

This paper investigates a serverless edge-cloud architecture to support knowledge management processes within smart cities, which align with the goals of Society 5.0 to create human-centered, data-driven urban environments. The proposed architecture leverages cloud computing for scalability and on-demand resource provisioning, and edge computing for cost-efficiency and data processing closer to data sources, while also supporting serverless computing for simplified application development. Together, these technologies enhance the responsiveness and efficiency of smart city applications, such as traffic management, public safety, and infrastructure governance, by minimizing latency and improving data handling at scale. Experimental analysis demonstrates the benefits of deploying KM processes on this hybrid architecture, particularly in reducing data transmission times and alleviating network congestion, while at the same time providing options for cost-efficient computations. In addition to that, the study also identifies the characteristics, opportunities and limitations of the edge and cloud environment in terms of computation and network communication times. This architecture represents a flexible framework for advancing knowledge-driven services in smart cities, supporting further development of smart city applications in KM processes.

In an era defined by rapid technological progress and societal change, the integration of Knowledge Management (KM) with Society 5.0 offers both vast opportunities and distinct challenges. KM encompasses the processes of creating, sharing, utilizing, and managing organizational knowledge and information[1]. Within organizations, KM aims to improve decision-making, drive innovation, and enhance overall performance, providing a competitive edge in today's dynamic market. While KM's importance is clear, successful implementation requires a supportive organizational culture, dedicated leadership, and continuous improvement of both processes and technologies[2].

As the world transitions into Society 5.0, a paradigm that emphasizes the integration of technology and innovation to enhance societal well-being, the role of KM becomes even more pronounced. Society 5.0 envisions a human-centric society, where technology facilitates seamless integration between physical and virtual realms, fostering creativity, critical thinking, and collaboration[3]. Key principles of Society 5.0 include social responsibility, data democratization, and the convergence of cyber and physical domains. Underpinning this vision is the notion of Industry 5.0, which seeks to optimize collaboration between humans and machines, thus maximizing productivity and efficiency while ensuring sustainability and environmental concerns[4].

The emergence of the Smart City concept has launched a new era in urban management, defined by the seamless integration of technology, people, and institutions. In this framework, KM is essential, enabling more efficient governance, resource allocation, and citizen engagement. This convergence of technology, society, and infrastructure creates a cyber-physical social system, transforming the traditional boundaries of urban management[5]. Knowledge management in the context of the smart city encompasses the efficient utilization of technology-mediated services, data analytics, and citizen participation to optimize urban infrastructure and governance. Through the aggregation and analysis of vast amounts of data generated by various stakeholders, smart cities can derive actionable insights to enhance public services and decision-making processes[6].

At the core of the smart city ecosystem lie technologies such as cloud computing, which provide the necessary computational resources to process and analyse data at scale. Cloud computing offers unparalleled scalability, cost-effectiveness, and accessibility, enabling organizations to harness the power of data without significant upfront investments in hardware infrastructure. By leveraging cloud resources, organizations can focus on innovation and service delivery, rather than infrastructure management[7–9].

Department of Computer Science and Technology (DTIC), University of Alicante, 03690 San Vicente del Raspeig, Spain. ✉email: a.macia@ua.es

The rapid urbanization and digital transformation of modern cities necessitate innovative approaches to managing knowledge efficiently. The evolution of smart city applications has led to an increasing reliance on cloud computing edge computing, and serverless paradigms to enhance service delivery, infrastructure governance, and public engagement[6,10]. Traditional cloud-based architectures, while providing scalability, often struggle with high latency, network congestion, and cost inefficiencies. At the same time, edge computing solutions, though reducing data transmission times, face constraints in computational power.[11]. While previous research has explored these architectures in various contexts, there is a critical need to examine their integration within KM applications to enhance decision-making and data-driven governance in smart cities.

This research is motivated by the need to develop a hybrid edge-cloud architecture that balances these trade-offs, enabling efficient, scalable, and cost-effective KM services. Using serverless computing, this work aims to simplify application deployment, optimize resource utilization, and enhance responsiveness for AI-driven smart city applications. This approach aligns with the goals of Society 5.0, fostering a human-centric, data-driven urban ecosystem that supports innovation, sustainability, and efficient governance.

This study introduces a novel serverless edge-cloud architecture specifically designed to support knowledge management processes within smart cities. Unlike traditional cloud-centric approaches, the proposed hybrid model takes into account the characteristics of knowledge processes to optimally distribute computational resources across the edge and cloud layers, significantly reducing data transmission times, network congestion, and operational costs. Using edge computing's proximity to data sources and serverless computing's scalability, this architecture enhances real-time analytics, intelligent decision-making, and adaptive service provisioning.

There are two main key contributions in this work. First, it explores the intersection of Knowledge Management (KM), cloud computing, edge computing, and serverless computing within the context of smart cities. By examining the technological foundations, applications, and implications of these paradigms, it highlights their transformative role in urban management and governance. Through a comprehensive analysis of existing literature and emerging trends, this study identifies critical challenges, opportunities, and advancements in KM-driven smart city services, particularly in AI-enabled applications. The second contribution is the conceptualization and empirical evaluation of a serverless edge-cloud architecture specifically designed to support KM processes in AI-driven smart city applications. This novel approach integrates low-power edge servers, cloud resources, and Content Delivery Network (CDN) edge infrastructure to optimize data processing, reduce latency, and enhance scalability. The study provides experimental validation demonstrating the benefits of architecture in terms of cost efficiency, network congestion reduction, and computational trade-offs between edge and cloud environments. By offering a flexible and adaptive framework, this work contributes to the development of scalable, cost-effective, AI-driven knowledge management solutions that align with the goals of Society 5.0 and smart urban governance.

This paper is organized as follows: After the introduction, we review the concepts of Knowledge Management (KM) and Society 5.0, exploring their interconnections. Next, we examine the role and importance of cloud and AI technologies as enablers of KM within the smart city paradigm. This section covers various cloud models, from core cloud computing to serverless edge computing, and discusses their implications for KM. Following this, we present our findings. Subsequently, we propose an architecture for AI-driven serverless edge applications in smart cities and perform experiments to analyse the opportunities, challenges, and unique aspects of KM services in this context. The paper concludes with a summary and final remarks.

Background

Knowledge management

Interest in the concept of Knowledge Management sparked in the 1990 where scholars started trying to define the concept. An accepted definition of Knowledge Management is that it is the process of creating, sharing, using and managing the knowledge and information of an organization[1]. From an organization perspective, it is aimed to optimize its ability to create, share, and use knowledge for improved decision-making, innovation, and performance, which can grant them a competitive edge over competitors. It is crucial for firms to understand the fundamental ideas behind knowledge and how to successfully manage their knowledge assets, as they are of vital importance to businesses and organizations due to its power to increase profits[12,13].

Knowledge management cannot be "bought", but instead it is a process that has to be implemented over a period of time. Successful knowledge management requires a supportive organizational culture, leadership commitment, and continuous efforts to refine processes and technologies to meet evolving needs. Managing knowledge involves knowledge gathering, organization and structuring, refinement and distribution[2]. There is a set of resources that facilitate the implementation of knowledge processes. Knowledge enablers are mechanisms that act as structural organizational means to foster knowledge processes[14]. They are categorized from people, organization, process, and system perspectives. Although they are essential in the capability of firms to manage knowledge effectively, they need to be used in knowledge management strategies for effective use, as they determine how to utilize knowledge resources and capabilities[12].

Knowledge creation and acquisition rely on technologies such as data and text mining, machine learning, and the IoT. For storing knowledge, tools like databases, knowledge bases, blockchain, and repositories are used. Sharing knowledge involves visualization tools, simulations, webinars, videoconferences, and social media. To apply knowledge, organizations implement knowledge-based systems, enterprise resource planning (ERP) systems, management information systems (MIS), and cognitive computing systems. Together, these technologies interconnect various processes, and the system's effectiveness depends on their seamless operation to ensure an uninterrupted flow of knowledge throughout the organization[14].

Society 5.0

Society 5.0 is a concept that emphasizes social responsibility and improving the quality of life through innovation and technology. It involves the integration of physical and virtual spaces, and the development of skills such as creativity, critical thinking, communication, and collaboration. In fact, it aims for a human-centered society. Society 5.0 envisions human beings interacting with social robots and artificial intelligence in their daily lives[3]. It also involves the tokenization process, creating tokenized digital twins of assets and access rights, which plays a central role in the future Web3 and its underlying token economy. Additionally, Society 5.0 promotes ethics, decentralization of power, data democratization, connected cyber/physical society, and resiliency by design. It aims to create personalized and purpose-led services, involving ecosystem participants from multiple industries, and decreasing customer acquisition costs[4,15].

Emerging technologies have played a pivotal role in shaping the foundation of Society 5.0. Today, many processes operate by collecting data from the environment, processing it to extract knowledge, and using these insights to drive changes or responses. A defining feature of a Smart Society is the interconnection of these processes, allowing them to work together seamlessly[16]. This concept extends to specific areas, including Smart Agriculture, Smart Industry, Smart Cities, and Smart Businesses, with ongoing research advancing these fields through smart applications.

KM in Society 5.0

Knowledge management is vital in Society 5.0, a knowledge-driven society that merges cyber and physical spaces, addressing challenges from the Fourth Industrial Revolution and Industry 5.0[14]. This integration enhances knowledge sharing and processes using Industry 4.0 technologies like IoT, cloud computing, and blockchain[17], and continues with Industry 5.0 innovations such as Big Data and AI[10,18]. These technologies reduce manpower and time for data analysis, transforming organizations and society by providing high-quality products and services efficiently. Knowledge management supports human-centric goals of sustainability and resilience, recognizing the centrality of knowledge assets in sustainable development by facilitating the acquisition, use, and communication of sustainable practices and coordinated community input[5,14].

KM facilitates the efficient sharing, creation, and utilization of knowledge for the betterment of society. Its integration in the Society 5.0 paradigm can be highlighted in the following key points:

- **Digital transformation:** Society 5.0 emphasizes the integration of digital technologies such as artificial intelligence, big data, IoT, and robotics into various aspects of society. Knowledge management helps in effectively harnessing these technologies by organizing and managing the vast amount of data generated, extracting valuable insights, and facilitating informed decision-making[19].
- **Collaborative innovation:** Knowledge management promotes collaboration and knowledge sharing among individuals, organizations, and communities. In Society 5.0, this collaborative approach is essential for fostering innovation and addressing complex societal issues. Platforms and tools for knowledge sharing and collaboration enable collective intelligence to be leveraged for problem-solving and innovation[20].
- **Lifelong learning:** Society 5.0 emphasizes the importance of lifelong learning to adapt to rapidly changing technological advancements and societal needs. Knowledge management systems support continuous learning by providing access to relevant knowledge resources, personalized learning experiences, and opportunities for skill development[21].
- **Human-centric design:** In Society 5.0, technology is designed to enhance human capabilities and improve quality of life. Knowledge management helps in understanding human needs, preferences, and behaviours, which is essential for designing and implementing human-centric solutions. By incorporating user feedback and insights into the knowledge management process, technology can be tailored to better serve society[22].

At the same time, the integration of Society 5.0 in Knowledge Management represents a paradigm shift towards decentralized, AI-driven, and human-centric knowledge systems. The use of IoT, AI, blockchain, and cloud-edge computing enables real-time knowledge acquisition, improved decision-making, and seamless information exchange in smart cities, industries, and public governance[23,24]. Table 1 summarizes the techniques used and the limitations and challenges of different aspects of the integration of society 5.0 in KM.

Examples of the symbiosis of the concepts of knowledge management and Society 5.0 can be seen in smart agriculture, where collection of information on the farm, field and culture, data analysis and decision-making and implementation of decisions are needed for precision farming[25]. In the smart city, intelligent transport monitoring systems (ITMS) use IoT to gather data, and the cloud to compute AI algorithms to helps authorities to gain knowledge that results in better planing that reduces accidents and traffic[26]. Advanced technologies such

Integration aspect	Techniques used	Limitations and challenges
AI-powered knowledge discovery	Machine Learning, NLP, Decision Support Systems	Algorithmic bias, lack of transparency in AI decision-making
IoT and smart knowledge sharing	Smart sensors, real-time data analytics, predictive insights	Interoperability issues, high data transmission and storage costs
Cloud & edge computing	Cloud-based KM systems, serverless edge computing	Privacy concerns, high infrastructure costs
Blockchain for knowledge governance	Decentralized knowledge storage, tokenized intellectual property	Scalability issues, complex regulatory compliance
Human-centered decision systems	Augmented intelligence, AI-driven personalization	Ethical concerns, risk of AI over-reliance
Digital twin for knowledge simulation	Real-time system modelling, data-driven decision-making	High computational requirements, need for extensive IoT integration

Table 1. Techniques and Limitations of Society 5.0 in Knowledge Management.

as 5G, virtual reality, machine learning, augmented reality, and data analytics are being used in smart factories to improve manufacturing and production processes. These technologies help gain valuable information that optimizes production lines and pushes manufacturing plants towards smart manufacturing[27,28].

Service provisioning models for KM in the smart city

Smart cities enhance the quality of life of its residents, and promote transparent management of public resources, including financial assets, natural resources, and infrastructure. Successful smart city initiatives rely on three core components in knowledge management: technology, people, and institutions. Together, these elements support urban infrastructure and governance, improving public services and citizen engagement. Technology-driven services leverage data to gain insights, crowdsource ideas, and deliver enhanced public services. Additionally, smart applications empower citizens to participate in shaping public policy, enriching both decision-making processes and the value of business operations.

KM in a smart city forms a cyber-physical social system that encourages collaboration among organizations and stakeholders, creating a sophisticated technological network within the urban ecosystem. Key components of this network are AI, IoT and Big Data, which play critical roles in knowledge co-creation, restructuring KM processes, and facilitating the exchange of human and organizational knowledge. Stakeholders contribute essential data, and trust is a crucial factor in fostering agreements between them and the municipality to effectively deploy smart technologies and successfully implement smart initiatives[6,29].

For transformative innovations that reshape how organizations and companies deliver benefits within the ecosystem, a supportive framework must emerge. This includes investments in new systems and skill development by firms, as well as an adaptation to the new approach by the customers, learning to use it to generate value. Over time, customers and citizens themselves become more adept at using information to manage their lives as workers, consumers, and travellers.

Modern platforms depend on the ability of businesses and individuals to create, access, and analyse vast amounts of data across various devices. Digital technologies such as social networks and mobile applications are driving the expansion of platforms into smart applications. These platforms utilize Big Data to collect, store, and manage extensive data sets[30]. A major challenge for the integration of smart applications into knowledge management processes is the workforce skill gap. To avoid the cost of acquiring new talents with the right expertise, Kolding et al.[31] conclude that organizations should plan ahead training programs to update the skills of the employee base in order to meet long-term development goals and other enterprise-wide priorities.

Figure 1 shows examples of smart city applications that involve knowledge management processes. The following sections will explore the technologies that establish platforms for knowledge management in smart cities.

Cloud computing

Cloud computing is a computing paradigm where the resources, be it applications and software, data, frameworks, servers or hardware, are stored on remote servers and accessed over the internet. This makes those resources available from anywhere with an internet connection. "Weak" computing devices send their computations to cloud servers, a practice known as Computing Offloading[32].

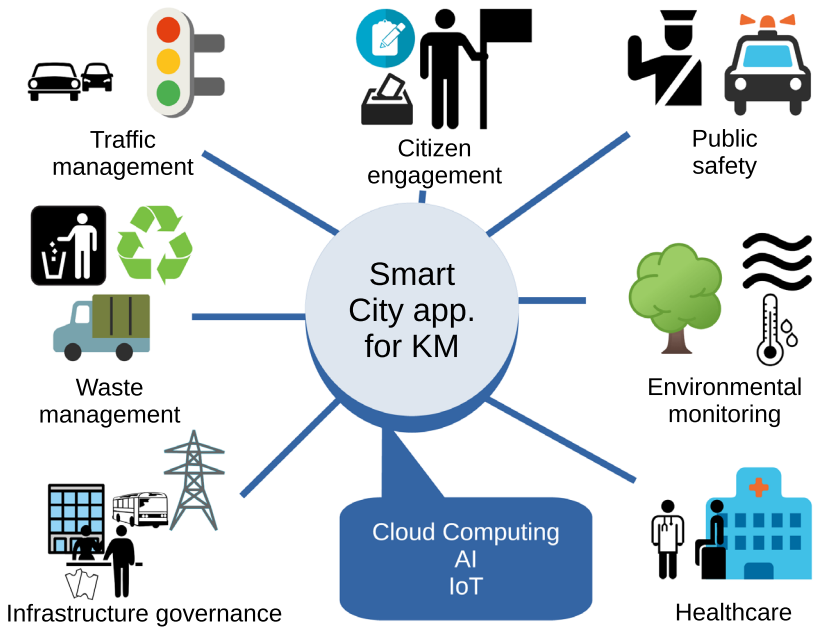

Fig. 1. Examples of smart city applications and technologies where knowledge management is involved.

The cloud brings numerous key benefits. It enables ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction[33]. It is a rapidly evolving field that has gained significant traction in recent years. It encompasses a range of technologies and market players, and is expected to continue growing in the future[34].

Organizations gain a set of advantages by using the cloud, including cost saving, scalability, reliability and flexibility[7]. It also allows organizations to focus on their core competencies instead of managing IT infrastructure. Additionally, cloud computing can help organizations improve their agility, since they no longer need to wait for hardware to be provisioned and deployed. Furthermore, cloud computing enables organizations to access powerful applications and services that they may not have been able to afford on their own[8].

Cloud resources range from data and software to concrete hardware resources. The following list provides an overview of the typical applications of cloud computing.

- **Web-based applications:** Cloud computing provides a platform for hosting web applications, allowing businesses to deploy and scale their applications without the need for significant upfront investment in hardware infrastructure.
- **Data storage and backup:** Cloud storage services offer scalable and reliable data storage solutions, allowing businesses to store and back up their data securely in the cloud.
- **Software as a service (SaaS):** Cloud-based SaaS applications enable users to access software applications over the internet, eliminating the need for local installation and maintenance.
- **Infrastructure as a Service (IaaS):** IaaS providers offer virtualized computing resources, enabling businesses to build and manage their IT infrastructure in the cloud.
- **Platform as a service (PaaS):** PaaS offerings provide a platform for developers to build, deploy, and manage applications without the complexity of managing underlying infrastructure.
- **Big data analytics:** Cloud computing platforms provide the scalability and computing power required for processing and analyzing large volumes of data, making it easier for organizations to derive insights from their data.
- **Internet of things:** Cloud platforms offer services for managing and processing data generated by IoT devices, enabling real-time analytics and decision-making.
- **Artificial intelligence:** and Machine Learning: Cloud-based AI and machine learning services provide access to powerful algorithms and computing resources for training and deploying machine learning models.
- **Content delivery networks (CDNs):** Cloud-based CDNs distribute content such as web pages, images, and videos to users worldwide, reducing latency and improving performance, which in term translates to better user experience.
- **Development and testing environments:** Cloud platforms offer on-demand access to development and testing environments, allowing developers to quickly provision resources and collaborate on projects.
- **Disaster recovery and business continuity:** Cloud-based disaster recovery services provide organizations with the ability to replicate and recover their IT infrastructure and data in the event of a disaster.

To smart city applications, offloading allows them to deploy complex applications on low power mobile devices. An intrinsic problem for an offloading application is resource allocation. Smart City applications need to allocate remote computing and networking resources to satisfy[35–37]. One of the main challenges of the offloading model is complying with the Service Level Agreement (SLA) between the cloud provider and the client. Generally, cloud platforms optimize resource scheduling and latency. These techniques strive to maintain deadlines during task execution due to urgency and resource budged limits. These extra resources needed during peak demand, known as marginal resources, are a challenge due to the variability of demand[38,39]. There are several strategies to this problem, such as decision support models using top-k nearest neighbour algorithm[38] or minority game theory[40]. Other authors incorporate deep learning into their resource allocation schemes[35,41].

Cloud computing helps knowledge management processes. It reduces the barrier to entry, as it eliminates the need to invest in IT infrastructure to implement knowledge management systems. Instead of investing in expensive infrastructure and hardware, organizations can leverage cloud resources on a pay-as-you-go basis, reducing upfront costs and operational expenses. It provides an alternative to the classical approach, as it provides the mechanisms to control, virtualize and externalize the infrastructure[42]. It is specially important SMEs, where the lack of adequate technical capabilities can hinder their implementations of knowledge management strategies[43] A framework for Knowledge Management as a Service (KMaaS) allows the users to access services from anytime, anywhere, and from any devices based on the user subscription for a specific domain. The knowledge management processes are impacted by this technology, specially knowledge sharing, knowledge creation, and knowledge transfer[44]. The implementation of cloud computing technologies impact in the overall organizational agility. Cloud computing gives them the capacity to deploy mass computing technology quickly, responding quickly to changes in the market. As a result, the performance of an organization is positively affected by cloud computing[43]. Practical applications of Cloud Computing in Knowledge Management include[45]:

- **Knowledge storage and sharing:** Cloud platforms facilitate centralized storage, allowing SMEs to store, retrieve, and share knowledge in real time.
- **Collaboration and communication:** Cloud-based tools (e.g., Google Drive, Microsoft Teams) enhance teamwork, remote work, and knowledge exchange.
- **Data security and backup:** Cloud computing provides secure environments with automated backups, minimizing data loss risks.

- **Scalability and cost-efficiency:** SMEs can scale their KM systems without significant infrastructure investments.
- **AI and analytics integration:** Cloud computing enables AI-driven knowledge discovery, pattern recognition, and decision-making.
- **Knowledge process automation:** Automating workflows and document management improves efficiency in KM processes.

Edge computing

Edge computing has emerged as a transformative paradigm in distributed computing, bringing computational and storage resources closer to the point of origin or consumption of data. This paradigm shift addresses the critical limitations of traditional cloud computing by reducing latency, saving bandwidth, and enabling real-time data processing[46]. Unlike centralized cloud architectures, edge computing strategically places resources at the network edge, facilitating ultra-responsive systems and location-aware applications[47].

The importance of edge computing is further underscored by its potential to support latency-sensitive applications in areas such as IoT, intelligent manufacturing, and autonomous systems. By processing data locally, edge nodes alleviate network congestion and enhance system reliability[48]. Moreover, edge computing is uniquely positioned to complement existing cloud infrastructure by acting as a bridge, ensuring seamless data transfer and computational efficiency[49].

This paradigm also introduces new opportunities and challenges in distributed system design, including efficient resource allocation, robust system architectures, and scalable management solutions. Edge computing's adaptability to evolving computational demands and its proximity to users make it a cornerstone of next-generation digital ecosystems[50]. By leveraging its unique capabilities, edge computing enhances existing technologies but also to paves the way for innovative applications across diverse domains.

Edge computing plays a pivotal role in enhancing knowledge management processes by addressing the critical challenges of data accessibility, processing efficiency, and timely decision-making. Its ability to integrate seamlessly with existing technologies and facilitate localized data processing has made it invaluable for knowledge-intensive operations.

One notable application of edge computing in knowledge management is in fostering collaboration across industrial systems. For instance, multi-access edge computing (MEC) frameworks enable the creation of knowledge-sharing environments within smart manufacturing contexts, such as intelligent machine tool swarms in Industry 4.0. This integration supports real-time data exchange and decision-making, critical for efficient knowledge dissemination[51].

Furthermore, edge computing contributes significantly to green supply chain management by facilitating the sharing of critical knowledge across enterprises. It enhances transparency and reduces resource consumption by integrating blockchain technologies to secure data transactions[52]. In open manufacturing ecosystems, edge computing has been employed to establish cross-enterprise knowledge exchange frameworks. By integrating blockchain and edge technologies, these frameworks ensure secure and efficient management of trade secrets and regional constraints[53]. From an enterprise innovation perspective, edge computing-based knowledge bases allow for enhanced data processing and storage at the edge, aligning with organizational goals of speed and reliability[54]. Edge computing also demonstrates its versatility in master data management by processing data at the source. This capability minimizes latency and ensures real-time updates for critical knowledge databases, particularly in dynamic business environments[55]. Finally, edge computing's integration into virtualized communication systems has paved the way for knowledge-centric architectures. Such systems optimize data collection and sharing, ensuring that actionable knowledge reaches stakeholders effectively[56].

There are several examples of the adoption of edge computing in KM processes. For example, Coppino[57] studied Italian SMEs for Industry 4.0 adoption for knowledge management. The research highlights that edge computing, when integrated with knowledge management enables SMEs to scale knowledge-sharing processes while improving real-time decision-making. However, SMEs struggle with insufficient IT infrastructure and lack of expertise, which limits overall adoption. For that reason, the author recommends developing frameworks to reduce technological investments. Stadnika et al.[58] perform a survey of representatives of companies from mainly European countries. Their results show that enterprises use edge data processing to increase data security and reduce latency.

Serverless computing

Serverless computing is a paradigm within the realm of cloud computing where developers can focus solely on writing and deploying applications without concerning themselves with the underlying infrastructure. In a serverless architecture, the cloud provider dynamically manages the allocation and provisioning of servers, allowing developers to create event based applications without having to explicitly manage servers or scaling concerns[59,60].

Instead of traditional server-based models where developers need to provision, scale, and manage servers to run applications, serverless computing abstracts away the infrastructure layer entirely. Developers simply upload their application, define the events that trigger its execution (such as HTTP requests, database changes, file uploads, etc.), and the cloud provider handles the rest, automatically scaling resources up or down as needed[61,62].

This model offers several advantages:

- **Scalability:** Serverless platforms automatically scale resources based on demand. Applications can handle sudden spikes in demand without manual intervention.

- **Cost-effectiveness:** With serverless computing, you only pay for the actual compute resources consumed during the execution. There are no charges for idle time, which can lead to cost savings, especially for applications with sporadic or unpredictable workloads.
- **Simplified operations:** Since there's no need to manage servers, infrastructure provisioning, or scaling, developers can focus more on developing applications and less on system configuration. This can accelerate development cycles and reduce operational overhead.
- **High availability:** Serverless platforms typically offer built-in high availability and fault tolerance features. Cloud providers manage the underlying infrastructure redundancies and ensure that applications remain available even in the event of failures.
- **Faster time to market:** By abstracting away infrastructure concerns and simplifying operations, serverless computing allows developers to deploy applications more quickly, enabling faster iteration and innovation.

Serverless computing is being explored as a solution for smart society applications, due to its ability to automatically execute lightweight functions in response to events. It offers benefits such as lower development and management barriers for service integration and roll-out. Typically, IoT applications use a computing outsourcing architecture with three major components for the processing of knowledge: Sensor Nodes, Networked Devices and Actuators. In the data gathering point, sensors gather data from specific locations or sites and submit it to the cloud service. Later, the analysis of the sensor data is carried out at cloud servers, where the data is processed to get useful knowledge from it. Applications have connection points for the clients to get access to the knowledge in the form of web applications[63,64].

Using the traditional cloud approach, the services should always be active to listen to service requests from clients or cloud users. The implementation of microservices is not a viable solution when considering the green aspect of systems. Holding servers for a longer period consequently increases the cost of the cloud services, as cloud computing is a pay-as-you-go computing model. With the serverless model, when an event is triggered by a request of the application, the needed computing resources to execute the function are provisioned, and released after they are not needed (scale to zero). With this model, applications, processes and platforms benefit from reduced cost of operation, as only useful computing time is paid for[65]. In fact, case studies show that entities that adopt the serverless paradigm achieve a lower cost of operation and faster response times[66].

Serverless edge computing

With the increasing number of IoT devices, the load on cloud servers continues to grow, making it essential to minimize data transfers and computations sent to the cloud[11]. Edge computing addresses this need by relocating computations closer to where data is gathered, utilizing IoT devices or local edge servers to perform processing tasks[67]. This proximity enhances latency, bandwidth efficiency, trust, and system survivability.

Serverless edge computing enables running code at edge locations without managing servers. It introduces a pay-per-use, event-driven model with "scale-to-zero" capability and automatic scaling at the edge. Applications in this model are structured as independent, stateless functions that can run in parallel[67]. Edge networks typically consist of a diverse range of devices, and a serverless framework allows applications to be developed independently of the specific infrastructure[68]. With the infrastructure fully managed by the provider, serverless edge applications are simpler to develop than traditional ones, making them ideal for latency-sensitive use cases.

Numerous studies have examined the challenges and opportunities within this model, proposing various approaches to leverage its strengths. Reduced latency and cost-effective computation are especially valuable in real-time data analytics[69]. Organizations benefit from the scalability of the serverless paradigm, which supports a flexible, expansive data product portfolio[70]. In this context, serverless edge computing enables more affordable data processing and improves user experience by reducing latency in data access and knowledge delivery interfaces.

Today, there are several serverless edge providers offering their services[71]. We studied ten different providers (Akamai Edge, Cloudflare Workers, IBM Edge Functions, AWS, EDJX, Fastly, Azure IoT Edge, Google Distributed Cloud, Stackpath and Vercel Serverless Functions) to understand how they offer their services. These providers can be divided into two main categories. The first category improves traditional CDN functionality by leveraging serverless functions to modify HTTP requests before they are sent to the user. The second group offers serverless edge computing frameworks that integrate edge infrastructure with their cloud platforms, enabling clients to build and incorporate their own private edge infrastructure into the public cloud. Additionally, we identified one provider, EDJX, that employs a unique Peer-to-Peer (P2P) technology to execute its serverless edge functions.

The technologies used by these providers are similar to their traditional serverless cloud counterparts. To implement the applications, high-level programming languages are typically used, like JavaScript, Python, Java. Usually, development is streamlined through the provider framework, which provides all the tools needed.

AI in KM processes

Knowledge management processes benefit significantly from AI through enhanced data processing, analysis, automation, and decision-making capabilities. AI brings sophisticated tools to KM that help organizations capture, organize, share, and apply knowledge more effectively. The following key points were extracted from the literature review:

- **Knowledge discovery and extraction.** AI tools, particularly natural language processing (NLP) and machine learning, can extract insights from vast amounts of unstructured data (e.g., documents, emails, reports, and social media). NLP enables AI to parse and understand text, identifying valuable patterns, relationships, and topics within data sources. For instance, AI can automatically categorize and tag documents, identify key insights, and detect trends that are relevant to the organization. In scientific research or industry, AI-driven

tools can mine research papers, patents, and technical documents to highlight emerging technologies and innovations[72,73].

- **Organizing and structuring knowledge.** AI-driven categorization, clustering, and tagging help organize knowledge into structured formats for easier retrieval. Using machine learning algorithms, AI can automatically classify information based on its content and relevance, enabling employees to find information more quickly. Semantic analysis, powered by AI, also groups related documents and concepts, creating a connected network of information that mirrors human knowledge organization[74].
- **Knowledge sharing and recommendation systems.** AI-powered recommendation systems suggest relevant knowledge resources to users based on their roles, recent activities, or queries. It is most common used to recommend content in digital platforms such as YouTube, Netflix, or Amazon. By analysing usage patterns and preferences, AI-driven KM systems can suggest documents, experts, or solutions that match immediate needs. This tailored approach to knowledge sharing ensures that clients and users receive the most relevant information without needing to sift through large repositories[75].
- **Automating knowledge capture.** AI technologies, such as robotic process automation (RPA) and machine learning, facilitate automatic knowledge capture by monitoring and recording daily activities and processes within an organization. For example, AI-powered chatbots can log interactions with customers or employees, storing valuable insights from these interactions in a knowledge base. AI also captures information from emails, meetings, or customer support calls, automatically adding relevant details to knowledge repositories[76].
- **Enhancing knowledge retrieval with search and NLP.** AI improves knowledge retrieval by enabling more sophisticated search mechanisms. With NLP, AI systems understand user queries in natural language, refining search results based on the intent behind queries rather than just matching keywords. Advanced AI-driven search engines in KM systems also employ semantic search to understand contextual relationships between terms, improving the accuracy and relevance of results[77,78].
- **Contextualizing and personalizing knowledge.** AI can personalize the KM experience by tailoring knowledge delivery based on an employee's role, department, or project involvement. Using machine learning, KM platforms analyse patterns in user behaviour to predict and deliver information that aligns with individual needs. This contextualization makes knowledge sharing more effective, ensuring that the right knowledge reaches the right person at the right time[79].
- **Augmenting decision-making and expertise.** AI supports decision-making by providing analytical insights drawn from historical data, documents, and external sources. In KM, AI-driven predictive analytics and machine learning models assess past data to offer insights, identify risks, and make informed predictions. Expert systems can also use AI to simulate the decision-making processes of human experts, providing guidance on complex tasks[80].
- **Developing virtual assistants for knowledge management.** AI-based virtual assistants, like chatbots, facilitate KM by answering employee questions, providing document links, or assisting with common tasks. NLP-powered chatbots in KM systems help employees access the knowledge they need by interacting through natural language queries. These assistants can handle frequently asked questions, provide guided instructions, and retrieve information, making KM accessible and interactive, which ends up increasing overall productivity[81].
- **Supporting knowledge creation through insights and innovation.** AI-driven analytics can highlight trends, patterns, and gaps in an organization's knowledge, encouraging innovation and knowledge creation. By identifying emerging trends, AI helps organizations remain competitive and proactive in knowledge development. AI tools can also support research and development by generating insights from internal and external data, suggesting new ideas or directions for innovation[82].

Applications and platforms for knowledge management in the smart city

Research on service provisioning models for KM in smart cities reveals a variety of approaches, emphasizing the integration of technological and organizational strategies to optimize services.

Several studies highlight the critical role of structured frameworks and platforms to facilitate efficient service delivery. For instance, Prasetyo et al.[83] propose a service platform that aligns with smart city architecture, promoting digital service introduction in dynamic urban environments. Similarly, Yoon et al.[84] describe HERMES, a platform that uses GS1 standards to streamline service sharing and discovery, enabling citizens to efficiently engage with services in a geographically and linguistically optimized manner.

Smart city service models also emphasize interoperability and tailored service offerings for diverse urban needs. The study by Kim et al.[85] discusses adapting service provisioning models according to urban types, underscoring the need for knowledge services tailored to the unique characteristics of each city. Additionally, Weber & Zarko[86] argue for regulatory frameworks to support interoperability, ensuring that services can be consistently deployed across various smart city contexts.

Technological advancements, including AI, IoT and big data are also foundational to KM service provisioning in smart cities. Sadhukhan[87] develops a framework that integrates IoT for data collection and processing, addressing the challenges posed by heterogeneous technologies in smart city infrastructures. In the same vein, 5G technology is viewed as a transformative tool for enhancing service efficiency, especially in traffic management, healthcare, and public safety domains[88].

The convergence of these platforms and frameworks signals a broader move towards knowledge-driven, citizen-centric service models. Efforts like Caputo et al.[89] model illustrate the importance of stakeholder engagement in building sustainable, effective service chains within smart cities. Ultimately, these models aim to transform urban living by making smart city services accessible, efficient, and responsive to the needs of the citizens.

Findings

Knowledge management has driven organizations and companies into investing heavily to harness, store and share information in knowledge networks. Through the smart city context, it has meant the sprawl of innovative platforms and applications. They are based on technologies such as AI, big data and IoT. To process the vast amount of data, they need a supporting computing architecture. In this regard, cloud computing comes as a great way to provide the necessary resources, due to its flexibility and pay per use model.

Serverless edge computing has gained significant attention recently for its potential to reduce costs and simplify development. Numerous technologies have emerged to enhance performance and minimize latency, reflecting strong industry interest. This trend is evident in the growing number of providers now offering serverless edge services. For knowledge management, it enables faster data processing, reduced latency, improved scalability, and enhanced reliability. The serverless edge paradigm also reduces the costs of developing and maintaining new knowledge management systems, which is specially for SMEs.

KM in smart cities requires structured frameworks and platforms to enable efficient service delivery. Frameworks aligned with smart city architecture support the smooth integration of digital services, while platforms that prioritize interoperability and adaptability ensure that services meet diverse urban needs. Foundational technologies such as AI, IoT, and big data play a critical role in facilitating KM, as they streamline data collection, processing, and sharing. These technologies, combined with 5G connectivity, support essential services like traffic management, healthcare, and public safety. The convergence of these tools indicates a shift toward knowledge-driven, citizen-centric models that prioritize accessibility and responsiveness, with stakeholder engagement emerging as a vital component for building effective, sustainable smart city services.

Particularly, AI significantly enhances KM by improving data processing, analysis, automation, and decision-making. Key aspects include knowledge discovery, where AI tools extract insights from unstructured data, and organization, where AI-driven categorization and clustering make knowledge more accessible. AI also aids in knowledge sharing by recommending relevant resources to users based on their roles and activities. Automation supports the capture of knowledge from daily activities, while enhanced retrieval techniques allow AI systems to understand and respond accurately to user queries. Personalization and contextualization further tailor the KM experience, ensuring that knowledge reaches the right individuals. Additionally, AI augments decision-making by providing predictive insights and supports innovation by identifying trends, gaps, and opportunities for knowledge creation.

Providers have leveraged existing infrastructure to deliver their services, with most vendors basing their offerings on established CDNs. These CDNs provide a global network of strategically located computing nodes, enabling reduced latency compared to traditional cloud servers. Many vendors also offer frameworks that facilitate easy integration of clients' own edge infrastructure. However, despite their improved latency, CDN servers remain centralized in specific geographic locations[90].

Related work

The integration of artificial intelligence and edge cloud computing has made advances in various applications. Duan et al.[91] provide a comprehensive survey on distributed AI that uses edge cloud computing, highlighting its use in various AIoT applications and enabling technologies. Similarly, Walia et al.[92] focus on resource management challenges and opportunities in Distributed IoT (DIoT) applications.

The synergy of cloud and edge computing to optimize service provisioning is well-documented. Wu et al.[93] discuss cloud-edge orchestration for IoT applications, emphasizing real-time AI-powered data processing. Hossain et al.[94] highlight the integration of AI with edge computing for real-time decision-making in smart cities, focusing on intelligent traffic systems and other data-driven applications. Kumar et al.[95] presents a deadline-aware, cost-effective and energy-efficient resource allocation approach for mobile edge computing, which outperforms existing methods in reducing processing time, cost, and energy consumption.

The taxonomy and systematic reviews by Gill et al.[96] provide a detailed overview of AI on the edge, emphasizing applications, challenges, and future directions. Their research underscores the potential of AI-driven edge-cloud frameworks for improved scalability and resource management.

The synergy between edge and cloud computing has been applied to improve predictive maintenance systems. For example, Sathupadi et al.[97] highlight how real-time analysis of sensor networks can enhance outcomes through AI integration. Additionally, Campolo et al.[98] explore how distributing AI across edge nodes can effectively support intelligent IoT applications.

Iftikhar et al.[99] contribute to the discussion on AI-based systems in fog and edge computing, presenting a taxonomy for task offloading, resource provisioning, and application placement. Gu et al.[100] propose a collaborative computing architecture for smart grids, blending cloud-edge-terminal layers for enhanced network efficiency. Jazayeri et al.[101] propose a latency-aware and energy-efficient computation offloading approach for mobile fog computing using a Hidden Markov Model-based Auto-scaling Offloading (HMAO) method, which optimally distributes computation tasks between mobile devices, fog nodes, and the cloud to balance execution time and energy consumption.

Lastly, Ji et al.[102] delve into AI-powered mobile edge computing for vehicle systems, emphasizing distributed architecture and IT-cloud provisions. These studies collectively underscore the transformative impact of AI-integrated edge-cloud frameworks across domains, particularly in knowledge management and smart applications. The novelty of our work lies in the conceptualization of a general model architecture for these use cases, where resources can be more efficiently used across applications.

Table 2 provides a summary of studies on AI-based edge-cloud architectures. While this topic has already been discussed extensively, the aim of this paper is to integrate this architecture with knowledge management based AI applications.

Study	Evaluation Tools	Performance Metrics	Datasets	Advantages	Disadvantages
Duan et al. (2023)[91]	Survey & literature review	Accuracy, latency, power consumption, communication cost, memory footprint, privacy	N/A	Comprehensive coverage of DAIoT applications	Focused on Distributed Artificial Intelligence
Walia et al. (2024)[92]	Survey & literature review, Theoretical analysis	Processing time, offloading latency, power consumption	N/A	Highlights resource management challenges	Limited Discussion on Energy Efficiency, Complexity in Practical Implementation
Wu et al. (2020)[93]	Survey & literature review, Cloud-edge orchestration framework	Accuracy, communication cost, power consumption, latency, resource utilization	N/A	Insight into research challenges and open issues	lack of cost-effectiveness analysis
Hossain et al. (2023)[94]	Experimental evaluation	Latency, accuracy, power consumption, network bandwidth	Multimodal AI optimization techniques	Detailed exploration of integrating multimodal AI with edge computing	Focused on multimodal AI with edge computing
Kumar et al. (2024)[95]	Experimental evaluation	Computation cost, energy consumption, processing time	Synthetic	Cost-effective, energy-efficient and scalable offloading strategies	Lack of analysis on security
Gill et al. (2025)[96]	Survey & literature review	Latency, energy consumption, accuracy, scalabiliy, resource utilization	N/A	Extensive literature review	Edge only approach, lack of scalability analysis
Sathupadi et al. (2024)[97]	Cloud-edge predictive maintenance framework, experimental evaluation	Latency, power consumption, network bandwidth, accuracy	Sensor network data	Hybrid edge-cloud architecture for AI models for managing sensor networks in industrial settings	The workload distribution algorithm is constrained by the limited edge resources
Campolo et al. (2021)[98]	Proof of concept, experimental evaluation	Latency, accuracy, data transferred	N/A	Scalable AI distribution using a virtualized deep edge	Requires robust network infrastructure and its complexity of implementation
Iftikhar et al. (2023)[99]	Survey & literature review	Resource utilization, throughtput, latency, power consumption	N/A	Comprehensive review of fog and edge platforms managed by AI	Focused on fog and edge platforms managed by AI
Gu et al. (2023)[100]	Survey & literature review, theoretical analysis	Latency, power consumption, cache hit rate, computation cost	N/A	comprehensive review on collaborative computing architecture for smart grids	Complex decision-making, edge computing limitations and device heterogeneity
Jazayeri et al. (2021)[101]	HMAO (Hidden Markov Model-based Auto-scaling Offloading), experimental evaluation, iFogSim toolkit	Execution time, power consumption	Synthetic	Optimized computation offloading for time and energy efficiency	High model complexity, autoscaling overhead
Ji et al. (2020)[102]	Experimental evaluation	Accuracy, execution time, memory usage, power consumption	Deepfashion2 dataset	Comprehensive experimental evaluation on a real-world application	Limited scope

Table 2. Evaluation of AI-based Edge Cloud Computing Studies.

Serverless model for efficient AI knowledge processing in the smart city

This section presents a serverless Edge-Cloud model for KM processes in a smart city context. While edge infrastructure is currently limited in cities, major urban areas would greatly benefit from such infrastructure to unify and enhance various smart services. The model leverages low-power edge servers distributed across the city to handle data from connected IoT devices and stakeholders. Through wireless connectivity, including Wi-Fi, 5G, and other radio methods, edge devices form a network capable of local data processing. This distributed setup minimizes latency for critical applications, with computation tasks dynamically offloaded to cloud servers during peak demand, supported by Content Delivery Network (CDN) edge services to optimize response times. The infrastructure enables seamless knowledge capture, processing, and distribution, creating an adaptive framework for smart city applications. The experiments evaluate the performance characteristics of the proposed model in the context of typical AI function applications.

Figure 2 shows the Edge-Cloud model for KM processes in the smart city. Low power edge servers are distributed around the city, with wireless network connection to these servers. This connection can be achieved via traditional Wi-Fi network, 5G, or other radio transmission method forming a Wireless Sensor Network, where the sensors and actuators can be connected directly, and their data managed directly by the application running on the edge infrastructure provisioned by the provider. The edge nodes also service stakeholders in the knowledge processes. However, the edge environment is resource-constrained. Offloading computations to the cloud is an option to use in the case of a spike in demand. Also, CDN edge services can also be used in this case to have better latency than cloud servers. The Smart City applications that use the infrastructure generate knowledge that is stored, processed and distributed. In that sense, knowledge travels up into the architecture, where it is stored in cloud data centres. There, it is processed and sent down the architecture to be distributed to the stakeholders. A major benefit of it being a serverless architecture is that processing can be seamlessly transferred up and down the architecture.

The ecosystem is composed of a set of edge platforms E, distributed across the city.

$$E = \{e_1, e_2, ..., e_m\} \tag{1}$$

Let D be all the IoT and stakeholder devices distributed across the city.

$$D = \{d_1, d_2, ..., d_n\} \tag{2}$$

Let F be all the serverless AI functions in the platform.

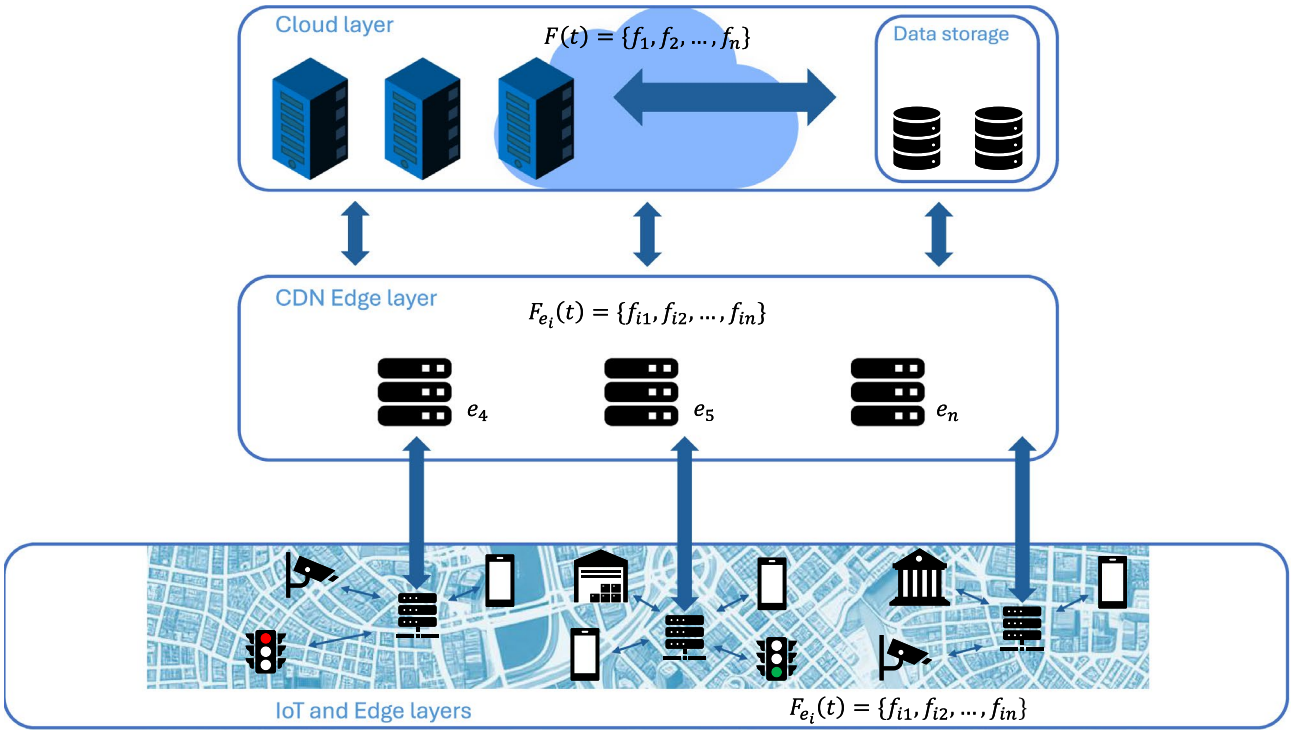

Fig. 2. Proposed service provider infrastructure for knowledge processing. Created with Powerpoint version 2502 (https://www.microsoft.com/en-us/microsoft-365/powerpoint). The city map was generated with Copilot AI version 18.2503.1061.0 (https://copilot.microsoft.com/).

$$F = \{f_1, f_2, ..., f_n\} \tag{3}$$

$$\text{Where } F_{e_i}(t) \subseteq F \tag{4}$$

Therefore, $||F|| = n$ is the total amount of AI functions that exist in the platform to provide service to all the applications. These are distributed between the different edge platforms around the city to provide the functionality where it is needed. As a consequence, the functions deployed at any time (t) change. The most optimal approach is the execution of these functions in the edge. However, in cases of peak demand, the execution of the function f_i may be allocated in a cloud server. Lastly, the cloud stores all AI functions, and distributes them to edge servers where needed.

The infrastructure is distributed in four layers.

- **IoT and stakeholder devices:** This layer corresponds to the sensoring devices and actuators. These are the devices that interact with the users and gather data to be processed. In a smart city environment, they are expected to be dispersed through the city in multiple smart applications that stakeholders interact with.
- **City edge servers:** As data is gathered by the sensoring devices, it needs to be offloaded to be processed. Here, local computing edge servers spread through the city offer the necessary computing resources to run the applications with local network connectivity latency. These computations harness useful knowledge from the data gathered by the IoT devices. Where necessary, computations and information can be sent to the next layers.
- **CDN edge servers:** Where computations are offloaded to the cloud, there is still the opportunity to use the existing CDN edge infrastructure. It presents itself as an intermediate computing layer between the local city edge servers and the remote cloud servers, located geographically closer than the cloud.
- **Cloud layer:** Traditional cloud infrastructure where the systems and applications default if a more nearby option is not available. Here, computing resources are plentiful. This layer is also used for data and knowledge storage. In this layer, knowledge is processed and distributed back to the stakeholders.

In this context, process time (PT) is defined as the total delay since the action is triggered until the action is completed. For example, a user would make a request for traffic status information. In this case, the delay perceived by the user is the time since it opens the service until the information is displayed. Therefore, PT can be expressed as a sum of different delays, as shown by Eq. 5.

$$PT = Comm.(data) + Setup(f_i) + Comp.(f_i(data)) + Comm.(f_i(data)) \tag{5}$$

Where $comp.(f_i(data))$ is the time it takes the function to compute with the associated data. $Setup(f_i)$ is the time it takes for the edge or cloud environment to be ready to execute the function. The first time a service is invoked, or after a long time, the corresponding function might not be present in the Edge environment. In that case, the function has to be retrieved from the cloud server, and the execution environment has to be created. This is a costly operation, but it is only performed the first time. Further invocations of the same function would find the execution environment ready. Lastly, there is the communication time. It is expressed as *Comm.*(*data*) and $Comm.(f_i(data))$. The first one is the time it takes for the data that wants to be processed to be transmitted to the edge or cloud environment. The second term is the time it takes for the result or response to reach its destination.

Figure 3 shows a sequence diagram of the interaction between the different components of the model. When a processing request is made on an IoT device, it first reaches the local edge layer. If processing cannot be completed there, it is offloaded to the CDN edge layer. If it is still necessary, further processing can be done in the cloud. This multilayered structure makes the architecture scalable. Techniques such as auto-scaling, load balancing, resource pooling, edge-cloud orchestration or predictive scaling can be employed to do the offloading decission from the lower layers to the upper layers[62]. After the request is completed, further processing is performed in the cloud to infer knowledge from the gathered data. Then, this knowledge is stored in knowledge databases, and IoT devices can make requests to retrieve it. Although a multi-layered architecture is able to allocate and distribute the execution of the task, it also introduces an overhead for the scheduling decision. In this case, each layer makes the decision whether to execute the request locally, or to offload the task to the upper layer. Therefore, in the worst case scenario, the time complexity is defined by Eq. 6, where the overhead time (O) of the scheduling decision in all the layers is aggregated. The overhead time in each layer is composed of the initial scheduling decision time (S), and the final response time (R) where the response from the upper layer is aggregated and sent to a lower layer of the architecture.

$$O = O_{Edge} + O_{CDN} + O_{Cloud} = (S_{Edge} + R_{Edge}) + (S_{CDN} + R_{CDN}) + (S_{Cloud} + R_{Cloud}) \quad (6)$$

The proposed framework integrates in every step of KM processes:

- **Knowledge creation:** IoT devices capture relevant data that is processed at edge nodes. The AI functions aggregate and analyse data in real time to produce actionable insights. On the other hand, cloud services are used to train and improve ML models that contribute to knowledge discovery, while also serve as support to process data in events of peak demand.
- **Knowledge storage:** The cloud offers many services related to data storage. Services such as Amazon S3 or Azure Blob Storage store information such as documents, logs, and other raw data that can be stored. Cloud databases store structured knowledge for efficient access.
- **Knowledge sharing:** Cloud APIs expose data through RESTful or GraphQL interfaces. Users receive real-time updates via WebSocket services.
- **Knowledge application:** The AI functions deliver personalized recommendations or insights based on user queries. Also, contextual data is used to refine outputs dynamically.
- **Knowledge feedback:** Users can rate or comment on knowledge outputs. That information is sent back to logging and analytics applications that capture it for further refinement.

Security is a major concern in all steps of knowledge management. Therefore, the proposed model must implement measures to secure the data at every step. One of the main concerns is to secure the communications. For that, data encryption techniques such as end-to-end encryption algorithms (e.g., TLS) can be employed. Authentication of users and services is also crucial. Role-Based Access Control (RBAC) and Attribute-Based

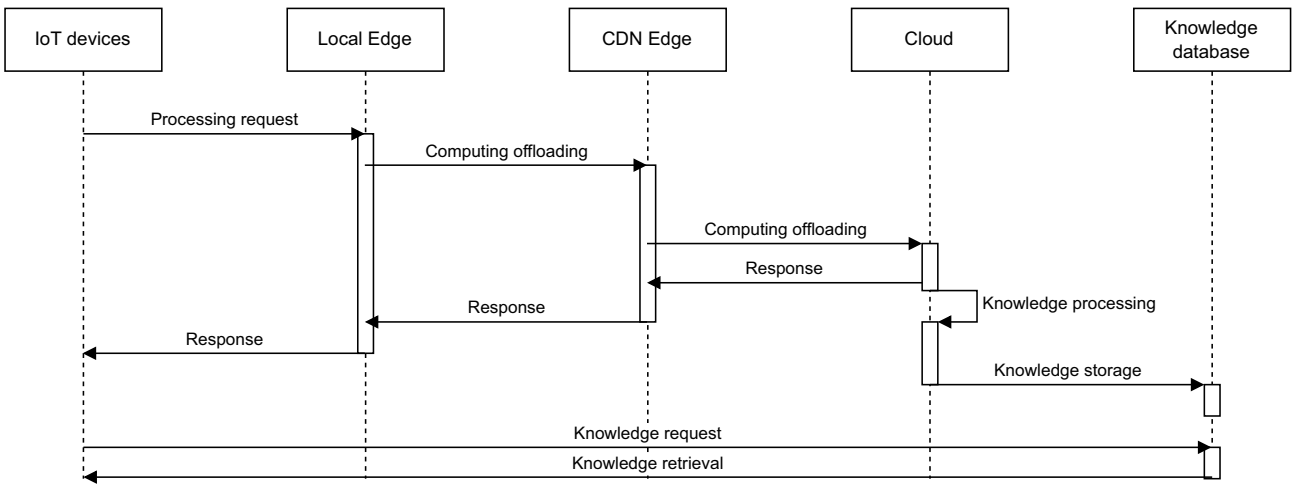

Fig. 3. Sequence diagram showing the interaction between the different components of the model.

Access Control (ABAC) should be used to authenticate APIs and databases. Additionally, decentralized identity mechanisms can be integrated to enhance user privacy and control over personal data. This approach leverages blockchain technology to create self-sovereign identities, reducing the risk of identity theft and unauthorized access[103].

To ensure data integrity and confidentiality, homomorphic encryption can be employed. This advanced encryption technique allows computations to be performed on encrypted data without decrypting it, thus maintaining data privacy throughout the processing lifecycle. Homomorphic encryption is particularly useful in scenarios where sensitive data needs to be processed in untrusted environments, such as edge devices or third-party cloud services[104].

Lastly, there is the need to comply with privacy regulations such as GDPR or CCPA, to which special attention should be given to protect sensitive data with secure and regulation-compliant storage. Implementing these security measures will provide comprehensive protection of sensitive data across the architecture layers, ensuring that the system is resilient against various security threats.

Another crucial aspect of the proposed model is the concern for interoperability, and avoiding the vendor lock-in problem. Ensuring seamless integration is essential for the efficiency and scalability of smart applications. It is also essential to share resources from a pool of cloud-service providers in a seamless fashion. To this end, multi-cloud deployments or federated cloud models could be used. Multi-cloud environments allow organizations to distribute workloads across multiple cloud providers, reducing dependency on any single vendor. This approach enhances flexibility, resilience, and cost optimization[105]. At the same time, Federated cloud models promote interoperability between different cloud providers through standardized APIs and governance frameworks. This enables seamless workload migration and data sharing while maintaining autonomy[106].

However, components from different cloud providers lack standardization, making it difficult to manage resources from different providers and legacy applications. To mitigate this problem, there are several strategies that may be used, such as unified resource management, security integration, transparent data governance policies, and cost optimization[107].

Finally, there are the environmental concerns of the architecture. Performing computations in such scale makes use of significant power resources and its corresponding carbon footprint. The proposed architecture improves over an all-cloud scenario. Traditional cloud computing infrastructures rely heavily on centralized datacenters, which consume significant amounts of electricity and often depend on non-renewable energy sources, contributing to substantial carbon emissions. The use of edge computing limits the demand for massive data centers by handling computations locally in energy-efficient devices[108,109].

Several experiments have been conducted to test characteristics of the architecture. The aim of the experiments is to demonstrate the benefits of the edge layer for smart applications. The experiments aim to model the *PT* of the cloud and edge platforms, measuring data transfer time and computation time. The experimental environment has been selected to be representative of the cloud and edge environments, with a cloud server and an edge device . The edge server is composed of a Raspberry Pi 4, with a Cortex-A72 64-bit 1.5GHz processor (Quad core), with 4GB of RAM. The Raspberry Pi is a low-cost, single-board computer widely used in applications such as robotics, IoT, scientific data acquisition. It is also commonly used as an edge device[23,110]. The cloud server is hosted on Amazon Web Services (AWS), one of the more popular cloud providers. The server model is m5.large, and it has 2vCPUs with 8GB of RAM. The experiment simulates a smart city application that captures images from a street camera, and processes them using an AI, deep learning algorithm to infer knowledge. As an example of a smart city application, car crash detection is performed over the images, using a Convolutional Neural Network (CNN) Deep Learning (DL) AI model. The model was developed by Escontrela et al., and it is available on a GitHub repository [https://github.com/Giffy/AI_CarCrashDetector]. The decision was made to use this model, as it is a model made for car crash detection, a typical smart city application. Different data sizes were used, using common image resolution sizes used to store pictures, while also covering a wide range of data sizes, from 99KB to 41.2MB. These are 99KB (224x224), 2MB (1920x1080), 6.6MB (3840x2160), 16.7MB (7680x4320) and 41.2MB (15360x8640) . Base latency measures the communication time when the smallest amount of data possible is sent. The network environment of the edge platform is composed of a household Wi-Fi communication network. In the cloud environment, standard internet network options were used. The experiments were performed in an "ideal" scenario for the edge and cloud servers, where the workload is low, and no other applications are being executed at the same time. For all the measurements taken, the experiment was repeated 5 times, and the times are averaged. Table 3 shows the 95% confidence interval ranges for upper and lower bound of the experiments. These values are within acceptable ranges for the following experiments to show meaningful results.

The first thing to be discussed is going to be *Comm.*(*data*). Figure 4 shows the transmission times obtained in the experiments. They are shown in blue for the cloud server, and in red for the edge server. As expected, the transmission time of the cloud server is higher than the transmission time of the edge server, as it takes less time to transmit data over a local Wi-Fi network than over the internet. Moreover, the variability of the transmission time is slightly higher for the cloud server, and it presents more outlier values. Factors that contribute to this are bigger geographic distances and internet bandwidth limitations.

Figure 5 includes all the values that contribute to the *PT*. Total computation times are shown with circle marks in the blue for the cloud server, and in red for the edge server. The results show the difference in computing capabilities for both platforms. The cloud server has a significant advantage over the edge server as it has more computing resources.

The last thing to be discussed is the total *PT* for both cloud and edge servers. Here, $Setup(f_i)$ is assumed to be 0, as the execution environment is presumed to have been set up in advance. Additionally, $Comm.(f_i(data))$ is 0, as this application only stores the knowledge gained from processing the image and does not send a response back to the camera. The diamond marked line shows the total *PT* for the cloud server, while the triangles shows

Experiment	Total Time (Range)	Processing Time (Range)	Comm. Time (Range)
Base latency 0B CLOUD	0.108–0.182s	0–0	0.108–0.183s
Base latency 0B EDGE	0.037–0.38s	0–0	0.037–0.113s
224x224 99.8KB CLOUD	0.69–1.086s	0.373–0.38s	0.31–0.706s
224x224 99.8KB EDGE	2.667–2.935s	2.521–2.539s	0.146–0.396s
1920x1080 (HD) 2.1MB CLOUD	1.199–1.812s	0.438–0.451s	0.76–1.374s
1920x1080 (HD) 2.1MB EDGE	3.224–3.37s	2.692–2.729s	0.525–0.641s
3840x2160 (4K) 6.6MB CLOUD	2.467–2.83s	0.62–0.871s	1.847–1.959s
3840x2160 (4K) 6.6MB EDGE	4.579–4.909s	3.219–3.248s	1.360–1.662s
7680x4320 (8K) 16.7MB CLOUD	5.077–5.523s	1.306–1.55s	3.772–3.972s
7680x4320 (8K) 16.7MB EDGE	8.502–8.871s	5.045–5.126s	3.387–3.789s
15360x8640 (16K) 41.2MB CLOUD	12.163–13.025s	3.772–3.803s	8.389–9.253s
15360x8640 (16K) 41.2MB EDGE	20.079–20.487s	12.042–12.093s	8.008–8.394s

Table 3. 95% confidence interval ranges for the experiments.

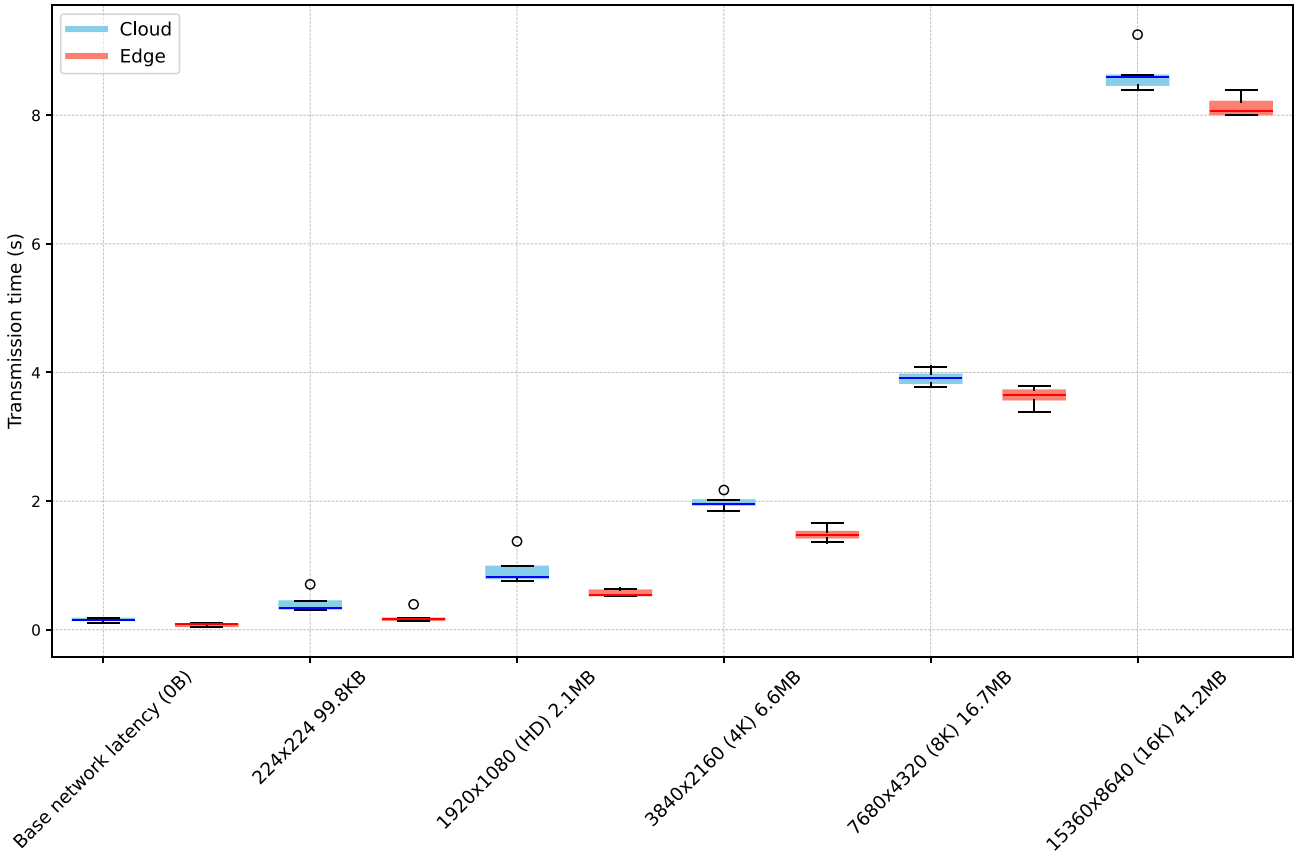

Fig. 4. Comparison of transmission times for different data sizes across Cloud and Edge platforms.

the total *PT* for the edge server. In a scenario with low workloads on both the edge and cloud servers, the cloud server exhibits lower *PT* than the edge server. Deep learning AI functions are tasks that need a significant amount of resources. The result is that the computations take long enough to make the transmission time advantage irrelevant in the edge server. However, this is an ideal scenario. In real environments, the workloads for the edge and cloud platforms change with time. The computation time is susceptible to change with the amount of workload of the platform. To quantify various workload scenarios, the experiments also display the critical computation times. The circle-marked dotted blue line shows the minimum critical computation time for the cloud. It is the minimum amount of computation time that would make the *PT* of the cloud bigger than the edge. On the other hand, the square-marked dotted red line shows the maximum non-critical computation time for the edge. It is the maximum amount of computation time possible that does not make the edge server have a bigger *PT* than the cloud server. For this application, the computation time in the edge in the ideal scenario is greater than this threshold. Other lighter applications may have an ideal computation time lower than the

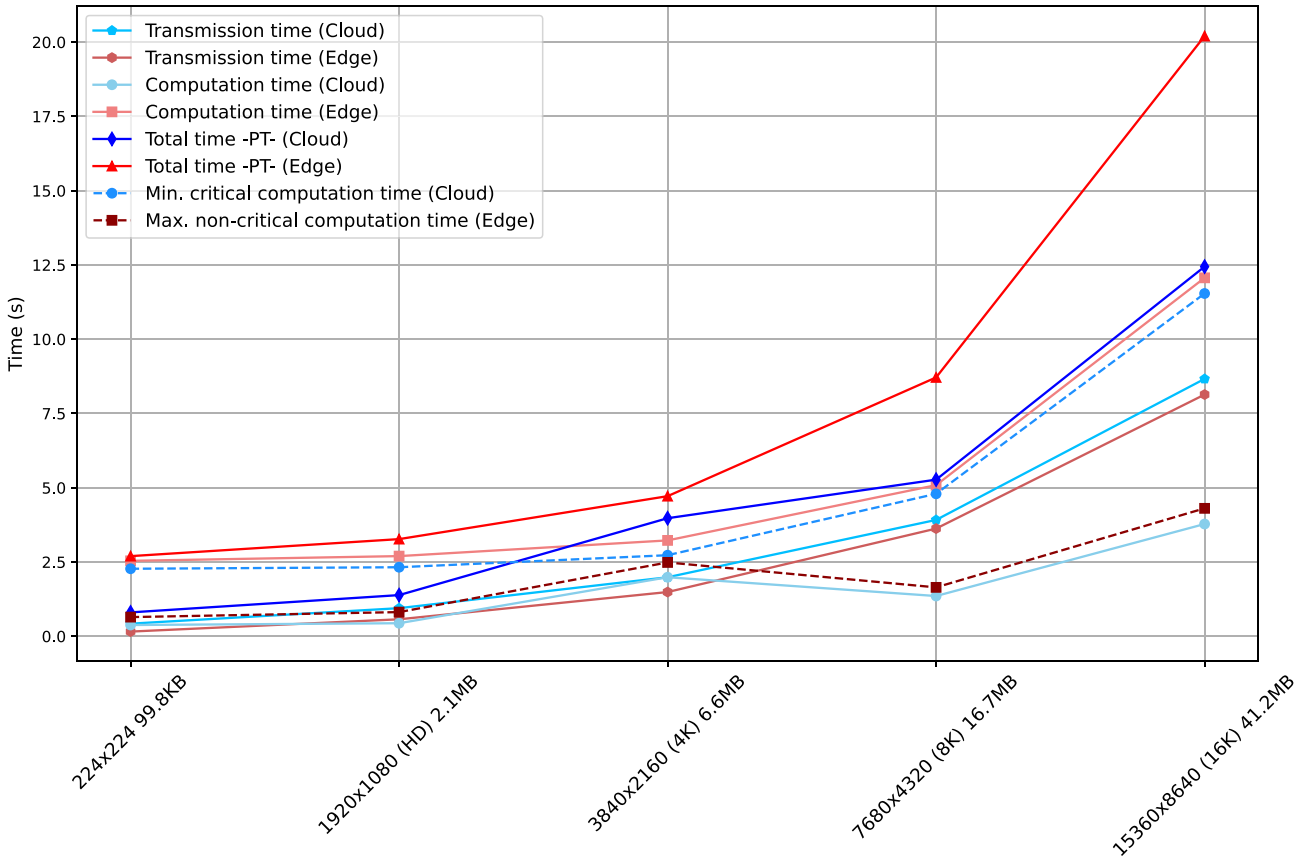

Fig. 5. Comparison of transmission time, computation and total processing times for different data sizes across Cloud and Edge platforms.

threshold. The thresholds can be used as a way to balance work to the cloud when the workload increases in the edge.

Although the cloud platform exhibited lower *PT* than the edge in the ideal scenario, the proposed Edge-Cloud architecture offers several other key advantages of edge computing. Price of computations in the edge depends on the power consumption, and the fixed price of the device. The base price of a Raspberry Pi 4 starts at $35. We got that value from the official website. The power consumption of the Raspberry depends on the CPU load and peripheral device usage. Experiments from Ali Süzen et al.[111] show that the average power consumption of a Raspberry Pi 4 during CNN work is 3.9W. Using the average electricity price of $0.31/kWh in the European Union in 2024, the price per hour of computation in the Raspberry edge server is of $0.001198. The m5.large server average price is of $0.1162 per hour. Using that data, Fig. 6 shows the price to compute the crash detection service with different image sizes for the cloud and edge servers. This experiment shows where the edge really shines. The results show that the edge server has a cost of computation up to 59.72 times less than the cloud server. By performing computation tasks to edge servers, the architecture reduces network congestion and optimizes resource allocation, contributing to overall cost savings. This shows the cost-effectiveness of the architecture.

On top of that, edge servers play a crucial role in reducing network congestion and providing more scalable and efficient data processing closer to the source. In this architecture, by offloading computations to local edge nodes, the system can handle real-time applications more effectively, even during peak demand periods. Furthermore, the distributed nature of edge servers allows for better resilience and fault tolerance, ensuring that critical services remain operational despite potential cloud service disruptions or network latency spikes. Therefore, while the cloud may offer advantages in total process time under specific conditions, the role of the edge server in optimizing resource allocation and ensuring continuous operation of smart city applications makes it an essential component of the proposed model. Furthermore, the integration of the cloud and edge layers improves the long term scalability of the applications.

Discussion on limitations and future work

There are some aspects of this work that could be improved in future research. First,

While this article provides a comprehensive analysis of the serverless edge computing model for KM processes in the smart city, several limitations must be considered:

First, the experiments were made on a limited set of devices, a single raspberry pi and cloud server, representing only two scenarios, the edge and the cloud. It may leave out other configurations used in real-world

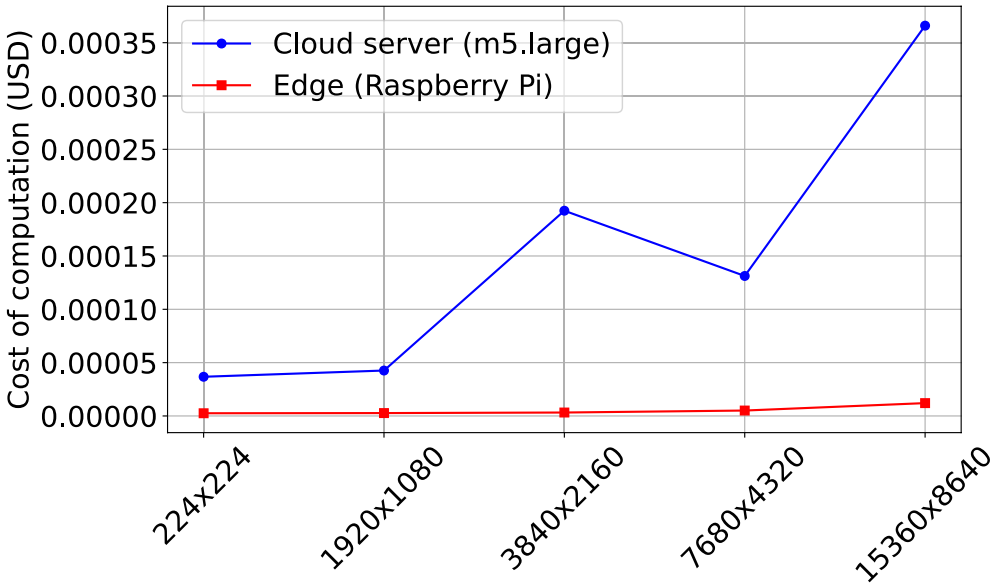

Fig. 6. Price of service computation in the edge and cloud servers.

scenarios. Second, the proposed architecture is a general model for KM AI applications, but does not account for the specifics of different applications. Some applications may require specialized computing devices, such as GPUs, which may require adaptations to the architecture. Also, the performance analysis conducted in the experiments is based on ideal conditions, where edge and cloud servers operate under minimal workloads. In real-world scenarios, dynamic workloads, network conditions, and environmental factors could lead to significantly different outcomes. Although the study addresses this by calculating critical computation time thresholds, it does not account for real-time workload analysis or adaptive measures to dynamically balance tasks. Furthermore, although this architecture is designed to be general, future iterations of the study could explore how the architecture can be tailored to meet the specialized requirements of varied applications, providing guidelines for implementing such adaptations in practice.

The long-term scalability and cost-effectiveness of the proposed architecture remain unexplored, which are critical factors for large-scale deployments. Similarly, regional differences in server locations, which could influence network latency and service quality, are not addressed, potentially impacting implementation in geographically diverse areas.

The paper also identifies security and privacy challenges in managing and storing sensitive data across multiple layers of the architecture. Each layer (edge, CDN, and cloud) has distinct security requirements, including encryption, authentication, and regulatory compliance (e.g., GDPR or CCPA). However, further discussion and a more focused approach of these security challenges and the implementation of measure would be beneficial.

The dependence of the architecture on skilled human resources for deployment and management introduces workforce challenges, as cities may lack personnel with expertise in serverless computing, edge infrastructure, and AI technologies. Furthermore, the distributed nature of edge nodes increases complexity in maintenance and reliability, requiring effective strategies to address hardware failures, software updates, and security patches.

In less-dense or underdeveloped regions, the economic and technical feasibility of expanding edge infrastructure is a concern. While urban areas might readily adopt such systems, sparsely populated areas could face significant barriers due to limited connectivity and sparse edge deployments. Even in well-connected areas, shared resource contention on edge networks during peak demand could cause latency variations, potentially impacting real-time responsiveness for critical applications.

Legal and regulatory hurdles, particularly those involving cross-border data storage and processing, are also overlooked. Multi-jurisdictional compliance beyond GDPR and CCPA could affect implementation and operational flexibility. Additionally, the reliance of the architecture on proprietary serverless frameworks and edge solutions raises the risk of vendor lock-in, making future transitions between providers costly and complicated.

Lastly, while edge nodes reduce latency, their limited processing power constrains their scalability for complex or resource-intensive AI applications. As applications grow in sophistication, the architecture may require significant upgrades to meet demands, further increasing operational costs. The dependency on stable network connectivity also poses challenges, particularly in environments with unreliable networks.

Future research should address these limitations by exploring dynamic workload balancing strategies, adaptive security protocols for distributed systems, and strategies for maintaining interoperability and scalability. The security aspect of the architecture could be expanded with advanced methods such as decentralized identity and authentication, homomorphic encryption, or zero trust architecture. Then this work only studies the use of the cloud layer as a single entity. However, there are multiple cloud providers, each with its own cloud

ecosystem. Further work could study how to avoid vendor lock-in using alternative architectures based on more heterogeneous cloud models such as the federated cloud. In addition, this architecture could be integrated with emerging technologies such as 5G or blockchain. These advancements will be essential for fully realizing the potential of serverless edge-cloud architectures in smart cities.

Conclusions

Society 5.0 involves interconnection networks to share knowledge between different stakeholders. In the smart city, it has kickstarted collaborative platforms and applications where information is shared between citizens, government, and organizations and companies. Smart applications gather data, process it, and exploit it and share it through data interconnection networks. Knowledge management plays a crucial role in enhancing public services, citizen engagement, and infrastructure governance. Technologies such as AI, big data and IoT form the backbone of knowledge management systems, enabling organizations to harness, store, and share information effectively.

However, smart applications require infrastructure to work. To this aim, there are several service provisioning models. Cloud computing serves as a foundational technology for KM in smart cities by providing cost savings, scalability, and flexibility, especially beneficial for SMEs by reducing the need for significant IT investments. Serverless computing builds on cloud computing by simplifying application development, allowing developers to focus on functionality rather than infrastructure. Edge computing, with servers located closer to data sources, offers faster and more cost-effective processing, despite higher maintenance costs. Together, serverless and edge computing enhance KM in smart cities by reducing latency, improving data processing, and lowering entry barriers for organizations. Various service providers are emerging to offer serverless edge computing solutions, leveraging existing CDN infrastructures and providing frameworks for easy integration with client infrastructure.

This study presents a comprehensive model of a serverless edge-cloud architecture to optimize KM processes in smart cities, with the aim of enhancing urban management and delivering intelligent, data-driven applications. The technologies that compose the architecture contribute in all steps of knowledge processes. The proposed model, which leverages the scalability and processing power of cloud computing alongside the low-latency and cost-efficient benefits of edge servers, provides a flexible framework to support a wide range of applications, from traffic management to public safety monitoring. This hybrid approach is especially well-suited to the latency-sensitive nature of many smart city applications, ensuring that critical data processing tasks can be handled locally whenever possible, reducing transmission time, and preserving bandwidth for other essential services.

The experimental results demonstrate that edge computing, while more resource-limited compared to centralized cloud servers, significantly reduces data transmission time by processing information closer to the data source. This approach reduces network congestion and improves the responsiveness of smart city services. However, it is notable that, for more computationally intense AI functions, cloud servers are better suited, where processing power outweighs network latency concerns. These findings suggest that optimal performance in smart city applications may require dynamic load balancing strategies that can shift workloads between edge and cloud resources based on real-time demand and resource availability. This flexibility underscores the need for a nuanced approach to resource allocation in urban data architectures, balancing computational efficiency, energy consumption, and speed.

Ultimately, this study underscores the significant potential of an edge-cloud architecture for supporting advanced KM processes in smart cities, offering a flexible, scalable and cost-effective framework capable of addressing the unique demands of urban data environments. By integrating AI, edge computing, and serverless infrastructure, this approach not only enhances data processing and sharing capabilities but also aligns with the goals of Society 5.0 in fostering efficient, inclusive, and responsive urban ecosystems. This work has the potential to impact several fields in the smart city environment, such as smart city management, environmental monitoring, industrial and manufacturing applications, retail and customer experience, education and research, and financial and business services. As smart cities continue to evolve, this model could serve as a foundational architecture for deploying adaptive, citizen-centered applications that can sustainably manage and leverage the vast data resources of the future.

Data availability

The datasets generated and/or analysed during the current study are available in the Github repository, https://github.com/cloudlab-aia/Edge-Cloud_KM.

Received: 23 December 2024; Accepted: 31 July 2025
Published online: 01 September 2025

References

1. Girard, J. & Girard, J. Defining knowledge management: Toward an applied compendium. *Online J. Appl. Knowl. Manag.* **3**, 1–20 (2015).
2. Benjamins, R., Fensel, D. & Gómez-Pérez, A. Knowledge Management through Ontologies. In *Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management (PAKM-98), International Conference on Practical Aspects of Knowledge Management (PAKM-98), 29–30 October 1998*, vol. 13, 12 (Facultad de Informática (UPM), 1998).
3. Ramírez-Gordillo, T., Mora, H., Maciá-Lillo, A., Amador, S. & Gil, D. Human-Centric Solutions and AI in the Smart City Context: The Industry 5.0 Perspective. In *Research and Innovation Forum 2023, Springer Proceedings in Complexity* (eds Visvizi, A. et al.) 193–203 (Springer, Cham, 2024). https://doi.org/10.1007/978-3-031-44721-1_16.
4. Harahap, N. J., Limbong, C. H. & Simanjorang, E. F. S. The education in era society 5.0. *J. Edusci. (JES)* **10**, 237–250. https://doi.org/10.36987/jes.v10i1.3959 (2023).
5. Smuts, H. & Van der Merwe, A. Knowledge management in society 5.0: a sustainability perspective. *Sustainability* **14**, 6878. https://doi.org/10.3390/su14116878 (2022).

6. Roblek, V. & Meško, M. Smart city knowledge management: Holistic review and the analysis of the urban knowledge management. In *The 21st Annual International Conference on Digital Government Research*, 52–60 (ACM, Seoul Republic of Korea, 2020). https://doi.org/10.1145/3396956.3398263.
7. Ramírez-Gordillo, T., Mora, H., Pujol-Lopez, F. A., Jimeno-Morenilla, A. & Maciá-Lillo, A. Industry 5.0: Towards Human Centered Design in Human Machine Interaction. In Visvizi, A., Troisi, O. & Corvello, V. (eds.) *Research and Innovation Forum 2023*, Springer Proceedings in Complexity, 661–672 (Springer International Publishing, Cham, 2024). https://doi.org/10.1007/978-3-031-44721-1_50.
8. Vishnu, S. A world with Cloud Computing. *Int. Sci. J. Eng. Manag.* **02**, 1–10 (2023).
9. Ghorbian, M. & Ghobaei-Arani, M. A survey on the cold start latency approaches in serverless computing: an optimization-based perspective. *Computing* **106**, 3755–3809. https://doi.org/10.1007/s00607-024-01335-5 (2024).
10. Simelane, B. & Smuts, H. Selecting a knowledge management methodology in Society 5.0. *EPiC Ser. Comput.* **93**, 164–173. https://doi.org/10.29007/ljbz (2023).
11. Elouali, A., Mora, H. & Gimeno, F. J. M. Data Transmission Reduction Model for cloud-based IoT Systems. In *2021 IEEE International Conference on Smart Internet of Things (SmartIoT)*, 252–256. https://doi.org/10.1109/SmartIoT52359.2021.00046 (2021).
12. Choi, B. & Lee, H. Knowledge management strategy and its link to knowledge creation process. *Expert Syst. Appl.* **23**, 173–187. https://doi.org/10.1016/S0957-4174(02)00038-6 (2002).
13. Imran, M. K., Fatima, T., Sarwar, A. & Amin, S. Knowledge management capabilities and organizational outcomes: contemporary literature and future directions. *Kybernetes* **51**, 2814–2832. https://doi.org/10.1108/K-12-2020-0840 (2021).
14. Lovrenčić, S. The Role of Knowledge Management in Transition to Industry 5.0. In *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, 1076–1082 (IEEE, Opatija, Croatia, 2023). https://doi.org/10.23919/MIPRO57284.2023.10159790.
15. Beniiche, A., Rostami, S. & Maier, M. Society 50: Internet as if people mattered. *IEEE Wirel. Commun.* **29**, 160–168. https://doi.org/10.1109/MWC.009.2100570 (2022).
16. Kashef, M., Visvizi, A. & Troisi, O. Smart city as a smart service system: Human-computer interaction and smart city surveillance systems. *Comput. Hum. Behav.* **124**, 106923. https://doi.org/10.1016/j.chb.2021.106923 (2021).
17. Mora, H., Pujol, F. A., Ramírez, T., Jimeno-Morenilla, A. & Szymanski, J. Network-assisted processing of advanced IoT applications: challenges and proof-of-concept application. *Clust. Comput.* **27**, 1849–1865. https://doi.org/10.1007/s10586-023-04050-6 (2024).
18. Tambuskar, D. P., Jain, P. & Narwane, V. S. An exploration into the factors influencing the implementation of big data analytics in sustainable supply chain management. *Kybernetes* **53**, 1710–1739. https://doi.org/10.1108/K-07-2022-1057 (2023).
19. de Bem Machado, A., Secinaro, S., Calandra, D. & Lanzalonga, F. Knowledge management and digital transformation for Industry 4.0: a structured literature review. *Knowl. Manag. Res. Pract.* **20**, 320–338. https://doi.org/10.1080/14778238.2021.2015261 (2022).
20. An, X., Deng, H., Chao, L. & Bai, W. Knowledge management in supporting collaborative innovation community capacity building. *J. Knowl. Manag.* **18**, 574–590. https://doi.org/10.1108/JKM-10-2013-0413 (2014).
21. Jagtar Singh, J. S. Sense-making: Information literacy for lifelong learning and knowledge management. *DESIDOC J. Libr. Inf. Technol.* **28**, 13–17. https://doi.org/10.14429/djlit.28.2.161 (2008).
22. Rinkus, S., Johnson-Throop, K. A. & Zhang, J. Designing a Knowledge Management System for Distributed Activities: A Human Centered Approach. In *AMIA Annual Symposium Proceedings* **2003**, 559–563 (2003).
23. Sahu, B., Tiwari, A., Raheja, J. L. & Kumar, S. Development of Machine Learning & Edge IoT Based Non-destructive Food Quality Monitoring System using Raspberry Pi. In *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*, 449–455 (2020).
24. Pal, S. Artificial Intelligence-Based IoT-Edge Environment for Industry 5.0. In Pal, S., Savaglio, C., Minerva, R. & Delicato, F. C. (eds.) *IoT Edge Intelligence*, 111–148 (Springer Nature Switzerland, Cham, 2024). https://doi.org/10.1007/978-3-031-58388-9_4.
25. Skobelev, P. O. et al. Development of a knowledge base in the "Smart Farming" system for agricultural enterprise management. *Procedia Comput. Sci.* **150**, 154–161. https://doi.org/10.1016/j.procs.2019.02.029 (2019).
26. Putra, A. S. & Warnars, H. L. H. S. Intelligent Traffic Monitoring System (ITMS) for Smart City Based on IoT Monitoring. In *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*, 161–165. https://doi.org/10.1109/INAPR.2018.8626855 (2018).
27. Chuang, S.-M., Chen, C.-S. & Wu, E. H. Implementation of Non-intrusive Intelligent Sensor System and 5G Edge Computing Gateway for Smart Factory. In *2022 IEEE 4th Eurasia Conference on IOT, Communication and Engineering (ECICE)*, 64–69, https://doi.org/10.1109/ECICE55674.2022.10042853 (2022).
28. Pham, Q.-V. et al. A survey of multi-access edge computing in 5G and beyond: fundamentals, technology integration, and state-of-the-art. *IEEE Access* **8**, 116974–117017. https://doi.org/10.1109/ACCESS.2020.3001277 (2020).
29. Sajid, A., Abbas, H. & Saleem, K. Cloud-assisted IoT-based SCADA systems security: a review of the state of the art and future challenges. *IEEE Access* **4**, 1375–1384. https://doi.org/10.1109/ACCESS.2016.2549047 (2016).
30. Stone, M., Knapper, J., Evans, G. & Aravopoulou, E. Information management in the smart city. *Bottom Line* **31**, 234–249. https://doi.org/10.1108/BL-07-2018-0033 (2018).
31. Kolding, M. et al. Information management: a skills gap?. *Bottom Line* **31**, 170–190. https://doi.org/10.1108/BL-09-2018-0037 (2018).
32. Mora, H., Ramírez, T., Pujol, F. A. & Jimeno-Morenilla, A. Mobile cloud computing paradigm: A survey of operational concerns, challenges and open issues. *Trans. Emerg. Telecommun. Technol.* **35**(12), e70020 (2024).
33. Telang, T. Introduction to Cloud Computing. In Telang, T. (ed.) *Beginning Cloud Native Development with MicroProfile, Jakarta EE, and Kubernetes: Java DevOps for Building and Deploying Microservices-based Applications*, 1–27, (Apress, Berkeley, CA, 2023). https://doi.org/10.1007/978-1-4842-8832-0_1.
34. Mora Mora, H., Gil, D., Colom López, J. F. & Signes Pont, M. T. Flexible framework for real-time embedded systems based on mobile cloud computing paradigm. *Mob. Inf. Syst.* **2015**, 652462. https://doi.org/10.1155/2015/652462 (2015).
35. Wang, J., Zhao, L., Liu, J. & Kato, N. Smart resource allocation for mobile edge computing: a deep reinforcement learning approach. *IEEE Trans. Emerg. Top. Comput.* **9**, 1529–1541. https://doi.org/10.1109/TETC.2019.2902661 (2021).
36. Kumar, M., Walia, G. K., Shingare, H., Singh, S. & Gill, S. S. AI-based sustainable and intelligent offloading framework for IIoT in collaborative cloud-fog environments. *IEEE Trans. Consum. Electron.* **70**, 1414–1422. https://doi.org/10.1109/TCE.2023.3320673 (2024).
37. Kumar, M., Walia, G. K., Shingare, H., Singh, S. & Gill, S. S. AI-based sustainable and intelligent offloading framework for IIoT in collaborative cloud-fog environments. *IEEE Trans. Consum. Electron.* **70**, 1414–1422. https://doi.org/10.1109/TCE.2023.3320673 (2024).
38. Hussain, W., Sohaib, O., Naderpour, M. & Gao, H. Cloud marginal resource allocation: a decision support model. *Mob. Netw. Appl.* **25**, 1418–1433. https://doi.org/10.1007/s11036-019-01457-7 (2020).
39. Qureshi, M. S. et al. Time and cost efficient cloud resource allocation for real-time data-intensive smart systems. *Energies* **13**, 5706. https://doi.org/10.3390/en13215706 (2020).
40. Carlucci, D., Renna, P., Materi, S. & Schiuma, G. Intelligent decision-making model based on minority game for resource allocation in cloud manufacturing. *Manag. Decis.* **58**, 2305–2325. https://doi.org/10.1108/MD-09-2019-1303 (2020).
41. Zhang, Y., Yao, J. & Guan, H. Intelligent cloud resource management with deep reinforcement learning. *IEEE Cloud Comput.* **4**, 60–69. https://doi.org/10.1109/MCC.2018.1081063 (2017).

42. Mostefai, M. A., Annane, A., Kissoum, L. & Ahmed-Nacer, M. Implementing knowledge management systems in cloud-based environments: A case study in a computer science high school. In *2015 International Conference on Cloud Technologies and Applications (CloudTech)*, 1–6, https://doi.org/10.1109/CloudTech.2015.7337008 (2015).
43. Khayer, A., Jahan, N., Hossain, M. N. & Hossain, M. Y. The adoption of cloud computing in small and medium enterprises: a developing country perspective. *VINE J. Inf. Knowl. Manag. Syst.* **51**, 64–91. https://doi.org/10.1108/VJIKMS-05-2019-0064 (2020).
44. Almehrzi, M. Cloud Computing Based in Knowledge Management in Higher Education Institutions: Benefit and Risks. In Arai, K. (ed.) *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 3*, Lecture Notes in Networks and Systems, 636–650 (Springer International Publishing, Cham, 2022). https://doi.org/10.1007/978-3-030-89912-7_49.
45. Saratchandra, M. & Shrestha, A. The role of cloud computing in knowledge management for small and medium enterprises: a systematic literature review. *J. Knowl. Manag.* **26**, 2668–2698. https://doi.org/10.1108/JKM-06-2021-0421 (2022).
46. Khan, W. Z., Ahmed, E., Hakak, S., Yaqoob, I. & Ahmed, A. Edge computing: A survey. *Future Gener. Comput. Syst.* **97**, 219–235. https://doi.org/10.1016/j.future.2019.02.050 (2019).
47. Yu, W. et al. A survey on the edge computing for the internet of things. *IEEE Access* **6**, 6900–6919. https://doi.org/10.1109/ACCESS.2017.2778504 (2018).
48. Escamilla-Ambrosio, P. J., Rodríguez-Mota, A., Aguirre-Anaya, E., Acosta-Bermejo, R. & Salinas-Rosales, M. Distributing Computing in the Internet of Things: Cloud, Fog and Edge Computing Overview. In Maldonado, Y., Trujillo, L., Schütze, O., Riccardi, A. & Vasile, M. (eds.) *NEO 2016: Results of the Numerical and Evolutionary Optimization Workshop NEO 2016 and the NEO Cities 2016 Workshop held on September 20-24, 2016 in Tlalnepantla, Mexico*, 87–115 (Springer International Publishing, Cham, 2018). https://doi.org/10.1007/978-3-319-64063-1_4.
49. El-Sayed, H. et al. Edge of things: the big picture on the integration of edge, IoT and the cloud in a distributed computing environment. *IEEE Access* **6**, 1706–1717. https://doi.org/10.1109/ACCESS.2017.2780087 (2018).
50. Li, C., Xue, Y., Wang, J., Zhang, W. & Li, T. Edge-oriented computing paradigms: a survey on architecture design and system management. *ACM Comput. Surv.* **51**, 1–34. https://doi.org/10.1145/3154815 (2018).
51. Zhang, C. et al. A multi-access edge computing enabled framework for the construction of a knowledge-sharing intelligent machine tool swarm in Industry 40. *J. Manuf. Syst.* **66**, 56–70. https://doi.org/10.1016/j.jmsy.2022.11.015 (2023).
52. Zhang, H., Li, S., Yan, W., Jiang, Z. & Wei, W. A Knowledge Sharing Framework for Green Supply Chain Management Based on Blockchain and Edge Computing. In Ball, P., Huaccho Huatuco, L., Howlett, R. J. & Setchi, R. (eds.) *Sustainable Design and Manufacturing 2019*, 413–420 (Springer, Singapore, 2019). https://doi.org/10.1007/978-981-13-9271-9_34.
53. Li, Z. et al. Toward open manufacturing: A cross-enterprises knowledge and services exchange framework based on blockchain and edge computing. *Ind. Manag. Data Syst.* **118**, 303–320. https://doi.org/10.1108/IMDS-04-2017-0142 (2018).
54. Tian, Z. & Wang, X. Construction of enterprise innovation performance model using knowledge base and edge computing. *J. Supercomput.* **78**, 9570–9594. https://doi.org/10.1007/s11227-021-04211-7 (2022).
55. Pansara, R. R. Edge computing in master data management: enhancing data processing at the source. *Int. Trans. Artif. Intell.* **6**, 1–11 (2022).
56. Wang, R., Yan, J., Wu, D., Wang, H. & Yang, Q. Knowledge-centric edge computing based on virtualized D2D communication systems. *IEEE Commun. Mag.* **56**, 32–38. https://doi.org/10.1109/MCOM.2018.1700876 (2018).
57. Coppino, E. *Unlocking the potential of Industry 4.0 in Italian SMEs: A Knowledge Manageme perspective*. Thesis, Politecnico di Torino (2024).
58. Stadnicka, D. et al. Industrial needs in the fields of artificial intelligence, internet of things and edge computing. *Sensors* **22**, 4501 (2022).
59. Li, Z. et al. The serverless computing survey: a technical primer for design architecture. *ACM Comput. Surv.* **54**, 1–34. https://doi.org/10.1145/3508360 (2022).
60. Ghorbian, M., Ghobaei-Arani, M. & Esmaeili, L. A survey on the scheduling mechanisms in serverless computing: a taxonomy, challenges, and trends. *Clust. Comput.* **27**, 5571–5610. https://doi.org/10.1007/s10586-023-04264-8 (2024).
61. Ebrahimi, A., Ghobaei-Arani, M. & Saboohi, H. Cold start latency mitigation mechanisms in serverless computing: Taxonomy, review, and future directions. *J. Syst. Archit.* **151**, 103115. https://doi.org/10.1016/j.sysarc.2024.103115 (2024).
62. Tari, M., Ghobaei-Arani, M., Pouramini, J. & Ghorbian, M. Auto-scaling mechanisms in serverless computing: A comprehensive review. *Comput. Sci. Rev.* **53**, 100650. https://doi.org/10.1016/j.cosrev.2024.100650 (2024).
63. Ghorbian, M. & Ghobaei-Arani, M. Function offloading approaches in serverless computing: A Survey. *Comput. Electr. Eng.* **120**, 109832. https://doi.org/10.1016/j.compeleceng.2024.109832 (2024).
64. Ghorbian, M., Ghobaei-Arani, M. & Asadolahpour-Karimi, R. Function placement approaches in serverless computing: a survey. *J. Syst. Archit.* **157**, 103291. https://doi.org/10.1016/j.sysarc.2024.103291 (2024).
65. Benedict, S. Serverless blockchain-enabled architecture for IoT societal applications. *IEEE Trans. Comput. Soc. Syst.* **7**, 1146–1158. https://doi.org/10.1109/TCSS.2020.3008995 (2020).
66. Kumari, A., Behera, R. K., Sahoo, B. & Misra, S. Role of Serverless Computing in Healthcare Systems: Case Studies. In Gervasi, O., Murgante, B., Misra, S., Rocha, A. M. A. C. & Garau, C. (eds.) *Computational Science and Its Applications - ICCSA 2022 Workshops*, 123–134 (Springer International Publishing, Cham, 2022) https://doi.org/10.1007/978-3-031-10542-5_9.
67. Aslanpour, M. S. *et al.* Serverless Edge Computing: Vision and Challenges. In *Proceedings of the 2021 Australasian Computer Science Week Multiconference*, ACSW '21, 1–10 (Association for Computing Machinery, New York, NY, USA, 2021). https://doi.org/10.1145/3437378.3444367.
68. Xie, R. et al. When serverless computing meets edge computing: architecture, challenges, and open issues. *IEEE Wirel. Commun.* **28**, 126–133. https://doi.org/10.1109/MWC.001.2000466 (2021).
69. Nastic, S. et al. A serverless real-time data analytics platform for edge computing. *IEEE Internet Comput.* **21**, 64–71. https://doi.org/10.1109/MIC.2017.2911430 (2017).
70. Perez, A. *et al.* Accelerating and Scaling Data Products with Serverless. In Krishnamurthi, R., Kumar, A., Gill, S. S. & Buyya, R. (eds.) *Serverless Computing: Principles and Paradigms*, Lecture Notes on Data Engineering and Communications Technologies, 149–173 (Springer International Publishing, Cham, 2023). https://doi.org/10.1007/978-3-031-26633-1_6.
71. El Ioini, N., Hästbacka, D., Pahl, C. & Taibi, D. Platforms for Serverless at the Edge: A Review. In Zirpins, C. *et al.* (eds.) *Advances in Service-Oriented and Cloud Computing*, Communications in Computer and Information Science, 29–40 (Springer International Publishing, Cham, 2021). https://doi.org/10.1007/978-3-030-71906-7_3.
72. Arnarsson, I. O., Frost, O., Gustavsson, E., Jirstrand, M. & Malmqvist, J. Natural language processing methods for knowledge management-Applying document clustering for fast search and grouping of engineering documents. *Concurr. Eng.* **29**, 142–152. https://doi.org/10.1177/1063293X20982973 (2021).
73. Shu, X. & Ye, Y. Knowledge discovery: Methods from data mining and machine learning. *Soc. Sci. Res.* **110**, 102817. https://doi.org/10.1016/j.ssresearch.2022.102817 (2023).
74. Zdravković, M., Panetto, H. & Weichhart, G. AI-enabled enterprise information systems for manufacturing. *Enterprise Inf. Syst.* **16**, 668–720. https://doi.org/10.1080/17517575.2021.1941275 (2022).
75. Gabrani, G., Sabharwal, S. & Singh, V. K. Artificial Intelligence Based Recommender Systems: A Survey. In *Advances in Computing and Data Sciences* (eds Singh, M. et al.) 50–59 (Springer, Singapore, 2017). https://doi.org/10.1007/978-981-10-5427-3_6.
76. Siderska, J. Robotic process automation: a driver of digital transformation?. *Eng. Manag. Prod. Serv.* **12**, 21–31. https://doi.org/10.2478/emj-2020-0009 (2020).

77. Tarmizi, W. A. M. B. A., Rashid, A. N. Z., Sapri, N. A. A. M. & Yangkatisal, M. Natural language processing (NLP) application for classifying and managing tacit knowledge in revolutionizing AI-driven library. *Inf. Manag. Bus. Rev.* **16**, 1103–1119. https://doi.org/10.22610/imbr.v16i3(I)S.3949 (2024).
78. Esteva, A. et al. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *Npj Digit. Med.* **4**, 1–9. https://doi.org/10.1038/s41746-021-00437-0 (2021).
79. Bakhshizadeh, M. Supporting Knowledge Workers through Personal Information Assistance with Context-aware Recommender Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, 1296–1301 (Association for Computing Machinery, New York, NY, USA, 2024). https://doi.org/10.1145/3640457.3688010.
80. Trunk, A., Birkel, H. & Hartmann, E. On the current state of combining human and artificial intelligence for strategic organizational decision making. *Bus. Res.* **13**, 875–919. https://doi.org/10.1007/s40685-020-00133-x (2020).
81. Marikyan, D., Papagiannidis, S., Rana, O. F., Ranjan, R. & Morgan, G. "Alexa, let's talk about my productivity": The impact of digital assistants on work productivity. *J. Bus. Res.* **142**, 572–584. https://doi.org/10.1016/j.jbusres.2022.01.015 (2022).
82. Waseel, A. H., Zhang, J., Shehzad, M. U., Sarki, I. H. & Kamran, M. W. Navigating the innovation frontier: ambidextrous strategies, knowledge creation, and organizational agility in the pursuit of competitive excellence. *Bus. Process Manag. J.* **30**, 2127–2160. https://doi.org/10.1108/BPMJ-02-2024-0081 (2024).
83. Prasetyo, Y. A. *et al.* Implementation of Service Platform For Smart City As A Service. In *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*, 416–422. https://doi.org/10.1109/ICITSI50517.2020.9264955 (2020).
84. Yoon, W. *et al.* HERMES: GS1-based Smart City Service Intercommunity. In *2018 IEEE International Smart Cities Conference (ISC2)*, 1–8. https://doi.org/10.1109/ISC2.2018.8656925 (2018).
85. Kim, H. Y., Jo, S. S. & Lee, S. H. A comparative analysis on the smart city service regarding urban types. *J. Korea Acad.-Ind. Coop. Soc.* **23**, 107–117. https://doi.org/10.5762/KAIS.2022.23.10.107 (2022).
86. Weber, M. & Podnar Žarko, I. A Regulatory View on Smart City Services. *Sensors* **19**, 415. https://doi.org/10.3390/s19020415 (2019).
87. Sadhukhan, P. An IoT based Framework for Smart City Services. In *2018 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, 376–379. https://doi.org/10.1109/IC3IoT.2018.8668103 (2018).
88. Storck, C. R. & Duarte-Figueiredo, F. A 5G New Smart City Services Facilitator Model. In *2019 IEEE Latin-American Conference on Communications (LATINCOM)*, 1–6. https://doi.org/10.1109/LATINCOM48065.2019.8937947 (2019).
89. Caputo, F., Walletzký, L. & Ge, M. Modelling the Service Value Chain for Smart City. In *DIMT-2017 Digitalization in Management, Society and Economy* (Gerhard Chroust, 2017).
90. Herbaut, N., Negru, D., Chen, Y., Frangoudis, P. A. & Ksentini, A. Content Delivery Networks as a Virtual Network Function: A Win-Win ISP-CDN Collaboration. In *2016 IEEE Global Communications Conference (GLOBECOM)*, 1–6. https://doi.org/10.1109/GLOCOM.2016.7841689 (2016).
91. Duan, S. et al. Distributed artificial intelligence empowered by end-edge-cloud computing: a survey. *IEEE Commun. Surv. Tutor.* **25**, 591–624. https://doi.org/10.1109/COMST.2022.3218527 (2023).
92. Walia, G. K., Kumar, M. & Gill, S. S. AI-empowered fog/edge resource management for IoT applications: a comprehensive review, research challenges, and future perspectives. *IEEE Commun. Surv. Tutor.* **26**, 619–669. https://doi.org/10.1109/COMST.2023.3338015 (2024).
93. Wu, Y. Cloud-edge orchestration for the internet of things: Architecture and ai-powered data processing. *IEEE Trans. Cloud Comput.* https://doi.org/10.1109/JIOT.2020.3014845 (2020).
94. Hossain, M. E. et al. Integrating AI with edge computing and cloud services for real-time data processing and decision making. *Int. J. Multidiscip. Sci. Arts* **2**, 252–261. https://doi.org/10.47709/ijmdsa.v2i1.2559 (2023).
95. Kumar, M., Kishor, A., Singh, P. K. & Dubey, K. Deadline-aware cost and energy efficient offloading in mobile edge computing. *IEEE Trans. Sustain. Comput.* **9**, 778–789. https://doi.org/10.1109/TSUSC.2024.3381841 (2024).
96. Gill, S. et al. Edge AI: A taxonomy, systematic review and future directions. *Clust. Comput.* https://doi.org/10.1007/s10586-024-04686-y (2025).
97. Sathupadi, K., Achar, S., Bhaskaran, S. & Faruqui, N. Edge-cloud synergy for ai-enhanced sensor network data: A real-time predictive maintenance framework. *Sensors* https://doi.org/10.3390/s24247918 (2024).
98. Campolo, C., Genovese, G. & Iera, A. Virtualizing ai at the distributed edge towards intelligent iot applications. *J. Sens. Actuator Netw.* https://doi.org/10.3390/jsan10010013 (2021).
99. Iftikhar, S., Gill, S., Song, C., Xu, M. & Aslanpour, M. Ai-based fog and edge computing: A systematic review, taxonomy and future directions. *Future Gener. Comput. Syst.* https://doi.org/10.1016/j.iot.2022.100674 (2023).
100. Gu, H., Zhao, L., Han, Z. & Zheng, G. Ai-enhanced cloud-edge-terminal collaborative network: Survey, applications, and future directions. *IEEE Access* https://doi.org/10.1109/COMST.2023.3338153 (2023).
101. Jazayeri, F., Shahidinejad, A. & Ghobaei-Arani, M. A latency-aware and energy-efficient computation offloading in mobile fog computing: a hidden Markov model-based approach. *J. Supercomput.* **77**, 4887–4916. https://doi.org/10.1007/s11227-020-03476-8 (2021).
102. Ji, H., Alfarraj, O. & Tolba, A. Artificial intelligence-empowered edge of vehicles: architecture, enabling technologies, and applications. *IEEE Trans. Intell. Veh.* https://doi.org/10.1109/ACCESS.2020.2983609 (2020).
103. Tadjik, H., Geng, J., Jaatun, M. G. & Rong, C. Blockchain Empowered and Self-sovereign Access Control System. In *2022 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, 74–82. https://doi.org/10.1109/CloudCom55334.2022.00021 (2022).
104. Kumari, K. A., Sharma, A., Chakraborty, C. & Ananyaa, M. Preserving health care data security and privacy using Carmichael's theorem-based homomorphic encryption and modified enhanced homomorphic encryption schemes in edge computing systems. *Big Data* **10**, 1–17. https://doi.org/10.1089/big.2021.0012 (2022).
105. Bhatnagar, S. et al. Understanding multi-cloud: maximizing efficiency in user requirement analysis and package distribution. *Int. J. Syst. Assur. Eng. Manag.* https://doi.org/10.1007/s13198-025-02729-0 (2025).
106. Zheng, C., Wang, L., Xu, Z. & Li, H. Optimizing Privacy in Federated Learning with MPC and Differential Privacy. In *Proceedings of the 2024 3rd Asia Conference on Algorithms, Computing and Machine Learning*, CACML '24, 165–169 (Association for Computing Machinery, New York, NY, USA, 2024). https://doi.org/10.1145/3654823.3654854.
107. Vadisetty, R. Multi Layered Cloud Technologies to achieve Interoperability in AI. In *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC)*, 1–5. https://doi.org/10.1109/ICEC59683.2024.10837471 (2024).
108. Xu, M. & Buyya, R. Managing renewable energy and carbon footprint in multi-cloud computing environments. *J. Parallel Distrib. Comput.* **135**, 191–202. https://doi.org/10.1016/j.jpdc.2019.09.015 (2020).
109. Alharbi, H. A. & Aldossary, M. Energy-efficient edge-fog-cloud architecture for IoT-based smart agriculture environment. *IEEE Access* **9**, 110480–110492. https://doi.org/10.1109/ACCESS.2021.3101397 (2021).
110. Kaur, A. & Jasuja, A. Health monitoring based on IoT using Raspberry PI. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 1335–1340. https://doi.org/10.1109/CCAA.2017.8230004 (2017).
111. Süzen, A. A., Duman, B., Şen, B. Benchmark. & Analysis of Jetson TX2, Jetson Nano and Raspberry PI using Deep-CNN. In,. International Congress on Human-Computer Interaction. *Optimization and Robotic Applications (HORA)* 1–5, 2020. https://doi.org/10.1109/HORA49412.2020.9152915 (2020).

Acknowledgements
This work was supported by the Spanish Research Agency (AEI) (DOI: 10.13039/501100011033) under project Serverless4HPC PID2023-152804OB-I00.

Author contributions
A.M-L has performed the review of cloud technologies and contributed to the development of the architecture and experiments. H.M has contributed to the writing of the manuscript, and the development of the Knowledge Management and Society 5.0 sections. He also coordinated the team, and supervised the experiments. A.J-M has contributed to the writing of the manuscript, and supervision of the development of the architecture and the experiments. N.G-D has contributed to the development of the architecture and experiments. J.A-L has contributed to the writing and supervision of the final version of the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to A.M.-L.