# An Agent-based Approach for Dynamic Big Data Processing in a Smart City Environment

Zakarya Elaggoune LIRE Laboratory Constantine 2 University 25000 Constantine, Algeria zakarya.elaggoune@univ-constantine2.dz Ramdane Maamri LIRE Laboratory Constantine 2 University 25000 Constantine, Algeria ramdane.maamri@univ-constantine2.dz

Imane Boussebough LIRE Laboratory Constantine 2 University 25000 Constantine, Algeria iboussebough@gmail.com

#### Abstract

The big data era brought us new processing and information management challenges to face. The existing tools managed to control the ongoing challenges, and the current architectures are close to meeting the needs of the users. But the volume rate at which new data is generated leads to new rising challenges. This is especially true in the context of smart cities, where gathering information in an energy-efficient manner to prolong the lifetime of Wireless Sensor Networks (WSNs); and adapting the analytical mechanism to support the speed at which new data is generated to deliver real-time results dynamically are the two key rising challenges. This article aims at exploring and describing how Multi-Agent Systems (MAS) can handle a large amount of data with a dynamic analytics capabilities and in an energy-efficient manner.

### **1** INTRODUCTION

The prospects for smart cities are very promising, and various smart device manufacturing groups, for example, IBM and Intel, are launching various initiatives to strengthen their focus in this sector. They recognized ten important areas that will play a key role in creating a smart city: smart lifestyle, smart security system, smart home, smart building, smart environment, smart government, smart grid, smart tourism, smart transportation and smart health [CDBN09]. Each component of smart cities is based on large-scale data analysis that show public safety, economic development, pollution, traffic conditions, and so on.

Smart cities are an imminent need, and are the true form of smart earth applied to custom areas to achieve intelligent and integrated city management. In smart cities, different sets of data are continually analyzed to present intelligent planning ideas, intelligent building models and intelligent management, where big data is treated as the fuel of any smart system [Coc14].

At the beginning of the Big Data era, three main challenges inherent to the characteristics of big data appeared (the "3V" initial Big Data):

Volume: data sets with enormous size and complexity (many features),

Velocity: fast generation of data arriving in continuous flows,

Variety: Different types of data come in different forms.

These challenges, also known as "data flood", have pushed storage systems and processing techniques to their limits at that time. After becoming familiar with the first three challenges, the new techniques began to perform well, but soon the flood of data overwhelmed these

Copyright  $\textcircled{}{}^{\odot}$  by the paper's authors. Copying permitted for private and academic purposes.

In: Proceedings of the 3rd Edition of the International Conference on Advanced Aspects of Software Engineering (ICAASE'18), Constantine, Algeria, 1,2-December-2018, published at http://ceur-ws.org





techniques. Indeed, as the volume of data increased and sources multiplied, raw data became increasingly poor and useful information became scarcer. Increasingly, the usefulness and reliability of data and their sources have been questioned. Hence the emergence of two new challenges taking the "3V" challenges of big data to "5V". [JGL<sup>+</sup>14] define the new 'Vs' as follows:

Value: the usefulness of the data or more precisely the amount of useful information among the flooded data,

Veracity: Reliability and confidence attributed to the data and its sources.

With the recent increase in the number of smart and portable devices and other measuring instruments in ambient applications and smart cities, we are just beginning to address every aspect of this new big data. In the smart cities context, we can extract two main rising challenges from this new big data:

Gathering data from WSN in an energy-efficient manner.A WSN consists of a large number of sensor nodes with limited batteries, which are randomly deployed over an area to collect data. The lifetime of the network decreases because of these limited batteries. Therefore, it is important to minimize the energy consumption of each node, which leads to the extension of the lifetime of the WSN. Since many of the detected data could be redundant or unimportant, collecting only relevant data could be a good technique for saving energy in sensor nodes and extending network lifetime.

Managing the dynamicity of the data in an adaptive way. One of the advantages of big data is the exploitation of the large volume of data in several purposes, like business strategies and healthcare. For efficient data exploitation, the data processing process stops and restarts each time new data arrives, and this to integrate the new data sensed into the processing cycle. Restarting the analytical process periodically consumes energy and time, therefore, processing data continuously without stopping in an adaptive way is a necessary task.

Thence, our goal is to propose a new approach for smart cities, that can gather relevant information in a smart manner and can adapt to changes that occur in the data without having to restart the entire process. Therefore, we use a multi-agent approach to design a two-tiers system: the first tier for data gathering and preprocessing (a smart wireless sensor network); and the second one is a real-time multi-agent system for dynamic big data analytics.

The rest of this article is organized as follows. In Sect. 2, we describe the two-tiers multi-agent-approach. In Sect. 3 the smart WSN is presented, describing in detail the different steps of relevant data extraction. Then we discuss the dynamic big data mechanism In Sect. 4. Lastly, we conclude our study in Sect. 5.

# 2 An Overview of the System

In this system we propose the use of fuzzy agents for the data relevance estimation. To communicate the data between sensor nodes with low energy consumption, we use the technique of clustering, where in each Cluster-Head(CH) an instance of a fuzzy agent is embedded. After gathering the data, each CH sends the extracted relevant data to the sink node, this last one dispatch the relevant data to the second-tier (processing agent) for real-time analysis.

Concerning the second-tier, which is the big data processing, we use a multi-agent system to build a threelayer big data processing system: a real-time processing layer; an adaptive batch processing layer; and a service layer that combines the results of the two previous layers.

The aimed system is composed of the following set of components (see Figure 1):

- First-tier- a smart wireless sensor network: sensor node; fuzzy agent; cluster-head; sink node.
- Second-tier- a dynamic big data processing: data node; processing agent; knowledge; service agent.

# 3 First-Tier: Smart Wireless Sensor Network

the basic role of sensor nodes is to collect information from the environment and send them to the base station in order to perform calculations. This collection must respect the battery life of each node to maintain the lifetime of the network.

The traditional model of data collecting is the Client/Server (C/S) approach. In the C/S approach; when the sensors capture the data, they send it directly to the base station as unprocessed raw data. in addition, to send data to the base station, the communication goes through a multi-hop communication. This multi-hop communication causes additional power consumption, because intermediate nodes relay information on more distant nodes. Several studies have been done to optimize the architecture of this model, some works are listed below:

- Incremental data fusion of a maximum number of sensors [PDN04]: when a node sends its data to the sink, the intermediate nodes merge their data with others coming from the first node. Therefore, this data is fused into a single message. this solution is not scalable, and it is suitable only for networks which does not contain a large number of nodes. Furthermore, the intermediate nodes do not have always relevant information to send and they do not filter out redundant and irrelevant information.
- Data aggregation for clustered WSN [CMM08]: the authors propose a clustering algorithm in which sensors choose themselves as cluster heads with a certain probability and disseminate their decisions. their work focuses on incorporating adaptive behavior into protocols in such a dynamic network. Once the data from each node is received, the cluster head transmits it directly to the sink. This solution based in the cluster heading paradigm which consumes a large amount of energy. Furthermore, the authors did not address the problem of complexity and neglected the importance of scalability of such kind of networks.
- The ant agent [LKF08]: the authors present a data aggregation based on ant colonies for wireless sen-

sor networks. they try to tackle the problem of building an aggregation tree for a group of source nodes in the WSN to send sensory data to the base station. However, the construction of this tree largely depends on the deployment of the nodes, which is generally random, and consumes a large amount of energy. Since the communication range of a node is limited, the nodes can only communicate with their one hop neighbors, so the euclidean distance between the source node and the receiving node is unreliable

- Mobile agent based directed diffusion (MADD) [CKY<sup>+</sup>06]: The authors considered mobile agents (MA) in multi-hop environments and adopted direct broadcast to dispatch the MA. In directed broadcasting, a detection task is broadcast through the sensor network as requests of interest for named data, i.e. the interests of the users are diffused through the sensor network. The sink node floods a request to the interest sensors and the intermediate nodes set gradients to send data around the routes to the sink node[IGE<sup>+</sup>03]. however, the current MADD framework is only suitable when the data is retrieved directly from the network whenever there are request from the users. some enhancement for the framework is needed to retrieve requests only from the active area.
- There are several works that have proposed a structured strategy like multicast tree[AKUMK09, UG07]. However, because of excessive communication costs and centralized management of the sensor network structure, structured approaches are not good for dynamic scenarios.

After having analyzed the solutions presented above, we can deduce that there is still a lot of work in terms of energy efficiency in the wireless sensor networks field, and since preprocessing data and eliminating irrelevant information contributes to lower energy consumption, our goal is to propose a wireless sensor network based on the relevance of data. We use the agent technique for intelligent and adaptive management.

For more efficiency, we have proposed the use of the clustering technique to send data easily to the Sink and for batter organization. We can use the algorithm Low Energy Adaptive Clustering Hierarchy (LEACH) or any other efficient algorithm to decompose the network into clusters, each with a Cluster-Head (CH). To achieve our objective, we propose to integrate, into each CH, a fuzzy agent to process data, eliminate non-useful data, and reduce redundancy. Each CH in the network is seen as an autonomous fuzzy agent with its own attitudes and characteristics towards the different events they receive.

#### 3.1 Fuzzy Agent Role Behaviors



Figure 2: Degree of relevance of data

The aim of the WSN is to collect the maximum data and eliminate the irrelevant or redundant ones.

Each Cluster-Head in the network is associated with a fuzzy agent (FA), the principal role of the FA is to use fuzzy logic to estimate the relevance of the data and to eliminate the unimportant data. Hence, we have defined two main points for fuzzy agent to extract the relevant information, which means to reduce the power of each node and to extend the life of the WSN:

- 1. *Degree of relevance of data:* the degree of relevance of the data strongly depends on the desired application. This parameter is calculated locally in the sensor node. The fuzzy agent can estimate the degree of relevance of the data collected. This information is taken into account if it's the primary information containing the required information. for example, for air pollution monitoring, the node records the latest collected data to compare with the new ones collected. The fuzzy agent considers data as relevant if the difference between the two values is greater than a predetermined threshold. However, if the difference increases, the fuzzy agent consider that these data have a higher priority, so the degree of relevance increases.
- 2. *Inter-sensor-nodes redundancy elimination:* typically, the sensor nodes are randomly deployed. so, many sensor nodes will cover the same geographical points, which means that they will give the same information (redundancy). In this case, the fuzzy agent will compare the values collected by each sensor node with its neighbors for eliminating the intersensor-nodes redundancy.

Figure 2 illustrates the fuzzy logic used by the agent to estimate the relevance of the data.

# 4 Second-Tier: Dynamic Big Data Processing

#### 4.1 Big Data Architectures

The most used process for big data analysis is the distributed pipeline (Figure 3-a). this model has been proposed to circumvent the rigidity problem by reducing the processing time by means of parallelism. This pipeline is based on the MapReduce pattern and its famous Hadoop framework.

However, applying this model does not solve the problem of data dynamicity, moreover, this model relies on batch processing and does not really focus on realtime processing, which leaves always a portion of nonprocessed data (Figure 3-b).

Other architectures have extended this model, trying to support the real-time processing, in the following paragraphs we will discuss the two most used architectures: Lambda Architecture(LA) and Kappa Architecture(KA).

• lambda Architecture (LA): "The LA aims to satisfy the needs for a robust system that is fault-tolerant, both against hardware failures and human mistakes, being able to serve a wide range of workloads and use cases, and in which low-latency reads and updates are required. The resulting system should be linearly scalable, and it should scale out rather than up." [HB]

This is what it looks like, from a high level point of view [HB]:

- All streamed data is sent to both the batch layer and the speed layer,
- The Batch layer pre-calculate the batch views,
- The serving layer indexes the batch views so that they can be queried in low-latency way,
- The speed layer indemnify the high latency of updates to the serving layer and process only recent data,
- Any incoming query can be resolved by merging results from real-time views and batch views.

The idea behind these layers was that the speed layer will be providing real-time results into serving layer, and if any data is missed while stream processing or any data errors, then batch job will compensate that and updates the serving layer, so providing accurate results. But it is very hard to build the pipeline and maintain analysis logic in both batch and speed layer.





Figure 3: (a) distributed Big Data Analytics pipeline [BGG16];(b) Big data processing-Batch

# Batch Layer Batch Engine Streaming Layer Real-time Engine Data storage Raw data

Figure 4: Lambda Architecture [SV16]

• Kappa Architecture (KA): "Kappa Architecture is a simplification of Lambda Architecture. A Kappa Architecture system is like a Lambda Architecture system with the batch processing system removed. To replace batch processing, data is simply fed through the streaming system quickly." [Ues]

One of the disadvantage of the lambda architecture, as detailed above, is to have to keep coding and executing the same logic twice, and this is avoided in the kappa architecture. However, the kappa architecture should only be considered an alternative to the lambda architecture in applications that do not require unbounded retention.



Figure 5: Kappa Architecture [SV16]

# 4.2 The need of new dynamic approaches

After having analyzed the solutions presented above, we <sup>Queries</sup> can deduce that available big data architectures do not really adapt to the dynamism of data. Furthermore, they must restarting periodically to take into account the realtime data streamed and does not integrate the new data in adaptive way.

The MAS technology, with the cooperative interaction process of its autonomous agents, gives us the means to break the rigidity problem in the other big data architectures, and can offer an adaptive management of big data streaming without the need to restarting the process periodically.

When an agent receives new data, it starts processing data directly to deliver real-time results. And after this agent consumes all the data stored in his node, he creates a link with the last agent in the batch-layer to contribute to the batch processing (distributed data mining), and another agent with an empty data node takes his place for real-time data processing. This translates into data analysis tasks in interaction, mainly through communication, then each task can help and work with other tasks for the sake of continuous real-time adaptation of the analytic process to changes in data.

The cooperation between the agents is described in the following steps (Figure 6) :

- 1. Each node in the system is associated with a processing agent. The node that receives the captured data from the WSN is responsible for rel-time processing and returns real-time views as a results, the other nodes in the system work on the batch processing and return the batch views.
- 2. Agents in the batch-layer are partitioned into neighborhood groups. The neighborhood is defined by



Figure 6: Multi-Agent Cooperation

time, from which two neighboring agents represent two successive periods. Each group represents a full batch period, from where agents of the same group apply distributed data mining and display batchviews.

3. Whenever the data stored in the real-time node is processed, the real-time agent updates the real-time views and creates a link with the last agent in the batch-layer to contribute to the batch processing. Another agent with an empty data node takes his place for real-time data processing.

Another way to achieve this goal, is to use the property of System-of-Systems (SoS) by combining one or several MASs for each step of Big Data analytics and represent them with an agent in one super MAS (see figure 7). this property is used to widen the batch period.



Figure 7: MAS of MAS based Big Data Analytics

# 4.3 Service Agent

The service agent is responsible for serving the views computed by the real-time and batch layers. This process can be facilitated by additional indexing of data to speed up readings [TR14]. The real-time Views and the batch Views are created for a specific use case. This use case problem is resolved in the serving layer (Figure 8). Querying from the users is managed by a dedicated service agent. For each new query the service agent is created.

To prepare the response and solve the given problem, service agent is collects the needed data. Fresh online data are provided by the real-time views. A similar processing is done to collect historical data (batch-views). Both views are combined together to display the whole picture of the data.

After combining all required data from the real-time and batch views, the response is presented. In this point the life-cycle of service agent ends.



Figure 8: Second-tier: dynamic big data processing

# 5 Conclusion

This two-tiers approach allow building the smart city as an agent community that can work in distributed and complex systems. The first-tier describes the construction and effective used of fuzzy agents in the wireless sensor network, with the consideration of the relevance of collected data, which can help enormously in the prolongation of the lifetime of the network by decreasing the energy consumption of each sensor node. In the processing layer, we described and discussed how multi agent system can be applied to process big data dynamically without the need to restarting the process periodically. As systems architecture and agent behaviors were designed, in our future research, we move into the implementation and validation phases.

# References

- [AKUMK09] Jamal N. Al-Karaki, Raza Ul-Mustafa, and Ahmed E. Kamal. Data aggregation and routing in wireless sensor networks: Optimal and heuristic algorithms. *Comput. Netw.*, 53(7):945–960, May 2009.
- [BGG16] E. Belghache, J. P. Georgé, and M. P. Gleizes. Towards an adaptive multiagent system for dynamic big data analytics. In 2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress [SV16] (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pages 753–758, July 2016.
- [CDBN09] A. Caragliu, C. Del Bo, and P. Nijkamp. Smart cities in europe. Serie Research Memoranda 0048, VU University Amsterdam, Faculty of Economics, Business Administration and Econometrics, 2009.
- [CKY<sup>+</sup>06] Min Chen, Taekyoung Kwon, Yong Yuan, Yanghee Choi, and Victor C.M. Leung. Mobile agent-based directed diffusion in wireless sensor networks. EURASIP Journal on Advances in Signal Processing, 2007(1):036871, Oct 2006.
- [CMM08] Huifang Chen, Hiroshi Mineno, and Tadanori Mizuno. Adaptive data aggregation scheme in clustered wireless sensor networks. *Comput. Commun.*, 31(15):3579–3585, September 2008.
- [Coc14] Annalisa Cocchia. Smart and Digital City: A Systematic Literature Review, pages 13–43. Springer International Publishing, Cham, 2014.
- [HB] M. Hausenblas and N. Bijnens. Lambda architecture.
- [IGE<sup>+</sup>03] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva. Directed diffusion for wireless sensor networking. *IEEE/ACM Transactions on Networking*, 11(1):2–16, Feb 2003.

- [JGL<sup>+</sup>14] H. V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Commun. ACM*, 57(7):86– 94, July 2014.
- [LKF08] Wen-Hwa Liao, Yucheng Kao, and Chien-Ming Fan. Data aggregation in wireless sensor networks using ant colony algorithm. *Journal of Network and Computer Applications*, 31(4):387–401, 2008.
- [PDN04] S. Patil, S. R. Das, and A. Nasipuri. Serial data fusion using space-filling curves in wireless sensor networks. In 2004 First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004., pages 182–190, Oct 2004.
  - N. SeyvetIgnacio and M. Viela. Applying the kappa architecture in the telco industry, 2016.
- [TR14] B. Twardowski and D. Ryzko. Multi-agent architecture for real-time big data processing. In 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), volume 3, pages 333–337, Aug 2014.
- [Ues] Shu Uesugi. Kappa architecture.
- [UG07] S. Upadhyayula and S. K. S. Gupta. Spanning tree based algorithms for low latency and energy efficient data aggregation enhanced convergecast (dac) in wireless sensor networks. *Ad Hoc Netw.*, 5(5):626–648, July 2007.