Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 191 (2021) 17–24

Procedia Computer Science

www.elsevier.com/locate/procedia

The 18th International Conference on Mobile Systems and Pervasive Computing (MobiSPC)
August 9-12, 2021, Leuven, Belgium

# A new Kappa Architecture for IoT Data Management in Smart Farming

Jean Bertin Nkamla Penka[a,*], Saïd Mahmoudi[a], Olivier Debauche[a,b,c]

[a] University of Mons, Faculty of Engineering - ILIA / Infortech, Place du parc 20, Mons 7000, Belgium
[b] University of Liège - GxABT, Terra, Passage des déportés 2, Gembloux 5030, Belgium
[c] University of Liège - GxABT, BioDynE - DEAL, Passage des déportés 2, Gembloux 5030, Belgium

**Abstract**

Agriculture 4.0 is a domain of IoT in full growth which produces large amounts of data from machines, robots and sensors networks. This data must be processed very quickly, especially for the systems that need to make real-time decisions. The Kappa architecture provides a way to process Agriculture 4.0 data at high speed in the cloud, and thus meets processing requirements. This paper presents an optimized version of the Kappa architecture allowing fast and efficient data management in Agriculture. The goal of this optimized version of the classical Kappa architecture is to improve memory management and processing speed. the Kappa architecture parameters are fine tuned in order to process data from a concrete use cases. The results of this work have shown the impact of parameters tweaking on the speed of treatment. We have also proven that the combination of Apache Samza with Apache Druid offers the better performances.



*Keywords:* agriculture 4.0; IoT; Internet of Things; kappa architecture; smart farming; smart farming.

## 1. Introduction

Agriculture 4.0 is a natural evolution of precision agriculture which saw the integration of ICT in the field of agriculture. Nowadays with Agriculture 4.0, the interaction between Wireless Sensors and Actuators Network (WSAN), agricultural machinery, robots (Milking, UGVs), UAVs (drones) [7], geo-services and external sources of data and services allowing to propose new services for farmers reducing the time spent on technical interventions and improving their well-being. These new applications automate a series of tasks that farmers previously performed and enlightens them in their decision-making. Automation and rapid decision-making require the fastest possible processing of data. These treatment are critical when they impact the control of the environmental conditions in which biological

* Corresponding author. Tel.: +32 65 374 059. Fax: +32 71 140 041.
*E-mail address:* JeanBertin.NKAMLAPENKA@student.umons.ac.be

1877-0509 
10.1016/j.procs.2021.07.006

objects (animal or plant) evolve. For example, in agriculture without substratum (aeroponic system) a failure in water or nutrient supply system quickly causes dieback or even death of cultivated plants. A second example come from industrial henhouse, where the control of the ammonia level, the temperature or the $CO_2$ level are not finely controlled, this implies the appearance of diseases and / or an increase in mortality. A third example is the supply chain ensured between farms and agro-industries that must known the availability of biological material used in the composition of their products to plan their production. Lambda and Kappa architectures are conventionally used for processing IoT data in smart farming.

The Lambda architecture is composed of two processing steps. The first process data in real-time while the second is specifically dedicated to batch processing of data in deferred time or for large-scale treatments. The main drawback of the Lambda architecture is the need to maintain two separate processing branches which leads to an increase in costs. Indeed, this architecture is well adapted if different processing operations are carried out on the two branches, for example when the current data is processed in real time and the old data stored in files or databases are processed in batch processing. The Kappa architecture is well suited for the online processing of data flows produced by IoT devices, but can also process offline data in the form of micro batches [10]. In this architecture, one way of processing is to ensure the treatment for the real-time and batch processing of data. This approach is advantageous because it uses the same code to achieve the treatment in batch and real-time processing. This approach is implemented when an estimated value must quickly calculate in real-time and in a second time, a more precise value is calculated by batch processing and replace the value obtained in first approach, in a second time.

However, this generalist architecture is not specifically optimized for smart farming. Wingerath et al. mentioned that Kappa architecture is viable only with fine tuned data retention or data compression or if high power computing is available. Referring us to Wingerath et al. [15], who attempt to optimize the performance of the Kappa architecture at message queue level namely where the data is temporarily stored before its processing. They noticed that the way with which the message queue which stores temporary data before their processing is configured directly impacts the speed of data ingestion and processing. They have also analyzed the influence of the allocated memory and the offset commit period at the message queue level on the global speed of treatment.

In this paper, we propose a fine tuned Kappa architecture on the basis of a concrete use case in Precision Livestock of behaviors classification. We will study the impact of each parameters on the overall performance of the architecture. The rest of this paper is organized as follow: In section 2, we summarize the works related to Kappa architecture. In section 3, we present the proposed modified Kappa architecture. Afterwards, in section 4, we present our experiments applied to a concrete use case, then the results are presented and analyzed. Finally, we conclude the paper and draw the future research perspectives.

## 2. Related works

In the domain of Internet of Things in particular in Smart Farming, different architectures exist which allow to collect, process and store data. One of the major architecture used is the Kappa Architecture. In the following paragraph, we will summarize the principal contributions about Kappa Architecture and stream processing. Persico et al. [11] benchmarked Lambda and Kappa architecture with three different configurations (horizontal scalability with standard deployment and optimized deployment, and vertical scalability). Experimentation was achieved on the Yahoo Flickr Creative Commons 100 Millions (YFFC100M) divided in subsets Small (1M), Medium (10M), Large (60M) and Extra-Large (100M) of tuples. Results show that Lambda architecture perform better than Kappa architecture on all datasets, on horizontal and vertical scalability tests [11]. Bixio et al. presented an architecture based on proxy, adapter and data processing microservices to manage stream data from IoT at edge and cloud level and able to manage dynamically and relocate microservices. This proposed architecture extends the IoT platform Senseioty[1], use Java OSGi microservice framework to develop microservices, and Siddhi[2] and Apache Flink as stream processing engines [2]. Zschörnig et al. suggested a personal analytics IoT platform based on a Kappa architecture where Kafka is the log data storage, Kafka stream is used for stream processing deployed as microservices developed in Java, Druid is the database

[1] Senseioty: https://senseioty.com/
[2] Siddhi: https://siddhi.io/

for the serving layer, API Services are written in Python, Metabase allows the visualization of data, and a data lake allows long term storage of data [16]. Persson et al. proposed a Kappa architecture based on Serverless deployment for IoT to push computation to the very edge of the network. The framework used to design these architecture is the distributed IoT-framework Calvin [12]. Feick et al. presented the state-of-the-art real-time architectures Lambda and Kappa for data processing. After a short description of each architecture, they made an experimentation with both architectures with a case study based on the Twitter's streaming API as data source. Their conclusion shown that the choice between these architectures depends on the use case and the constraints defined by the application [8]. Sanla et al. presented a comparative performance between the Lambda and the Kappa architectures for real-time Big Data analytic. Experimentation has been done with data size 3 MB, 30 MB and 300 MB. The results shown that Lambda architecture outperforms Kappa architecture for around 9% of accuracy test but it takes approximately 2.2 times more than Kappa architecture. They concluded also that Lambda architecture uses more 10-20% of CPU usage and 0.5 GB of RAM usage more than Kappa architecture. Therefore, they recommended to use Lambda architecture when accuracy is needed for the business and Kappa architecture when it is not the case but quick results is required [14]. Roukh et al. developed WalleSmart, an architecture dedicated to Smart Farming and based on an adapted Lambda Architecture and a datalake. Indeed in contrary to classical Lambda, they implemented different code for batch and real-time analysis [13]. Fote et al. presented a review of Big Data storage and analysis tools applied to Smart Farming. They proposed a Smart Farming architecture defined by Data sources (sensors, IoT devices, robots, etc.), by process tools (MQTT, Kafka and Storm), by storage point (Cassandra, PostgreSQL) and by a view dashboard built on NodeJS and Python [9].

These related works focused on optimized deployment and scalability, on comparison between Kappa and Lambda architectures performances or on a different Kappa architecture implementation based on Serverless deployment. This paper will focus on the impact of the message processing queue optimization on the global architecture performances in terms of ingesting speed.

Moreover, the related works focused on processing components choice, but rarely on the message queue optimization and its impact on global performances of the architecture. This part of the architecture is generally under studied. That is why; we proposed to investigate it in this work.

## 3. The modified Kappa architecture

The analysis of the related works presented in the previous section shows that there is no works that clearly attempts to optimize the message processing queue located upstream of the data processing component.

On the basis of the literature review, we propose in this paper a new optimized data processing pipeline based on a classical Kappa architecture, and composed of four majors parts: a message queue (**1**) which stores temporary data before their treatment by the data processing software (**2**) which produces a result stored in database (**3**). While an orchestrator (**4**) ensures the coordination of the operation of the different software and monitors their operating status. In this proposition, the memory usage and processing speed were fine tuned by optimizing Kafka parameters used as data log storage. Different combinations of data processing software and database have also been tested in order to choose the better ones.

The Fig. 1 presents the conceptual organization of the proposed architecture composed of a message queue which allows to collect data coming from external services, databases, files (CSV, TSV, and so on), agriculture machinery, UAVs, UGVs or sensors [7]. Afterwards, the data is ingested by data processing software and finally, the result of the treatment is stored in a database where the data can be queried by applications. The duration of data storage depends on its nature and value. Indeed, some data immediately lose its value after consumption while others can retain a value over time [7].

In order to optimize the memory usage and the processing speed of our Kappa, we decided to focus on the message queue which is Apache Kafka in our case. According to the Kafka's documentation, we found that we can improve its performances and therefore the processing time of the Kappa architecture, by using the parameter **OffsetcommitPeriodMs**. The offset in Apache Kafka represents the number assigned to each message (see message queue in Fig. 1). The **OffsetcommitPeriodMs** is the delay in milliseconds (ms) before the update of the offset in Kafka as treated by the consumer of the message. The Apache Kafka's documentation point out that Kafka needs 8 bytes of the RAM per offset to store its messages. Therefore we have decided to analyze the impact of the RAM's memory allowed to Kafka

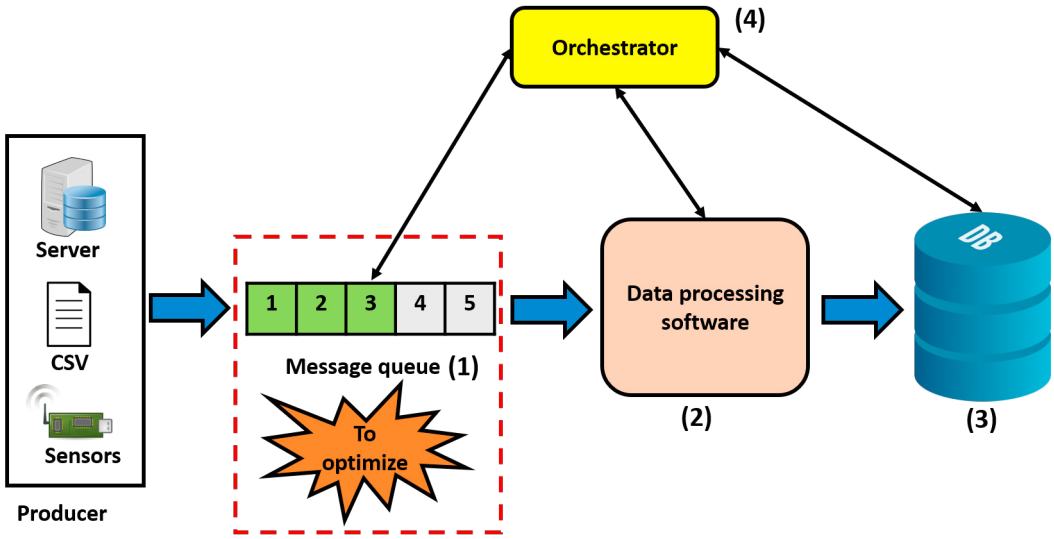



Fig. 1. General scheme of a Kappa architecture.

on the processing time of the messages. Considering the existing software that can be used to build the Kappa, we have compared the well known software for data processing and for data storage with a most recent software. They are described in the following section.

## 4. Experimentation

The use case implemented to achieve our experimentation is the behavior of farm animal's analysis in pasture based on IMU and GPS data [4, 5, 6, 3]. This analysis is particularly important to early detect signs of illness, distress and fertility periods. It allows also to detect especially bogged down animals, an escaped animals from its enclosure, lameness which is the source of economic losses and suffering for the animal, etc.

In the experimentation, the impact on processing time of the Kappa architecture was evaluated in function of Ram allocated and the offset commit period of the Kafka message queue.

The experimentation has been achieved on data produced by IMU 9-DOF and GPS posting logged by means of an iPhone 5s placed upside the neck of a cow. The iPhone 5s thanks to Data Sensor v1.26 allowing to collect 41 parameters at a rate up to 100Hz. Collected data are (1) Acceleration on x, y, z; (2) Euler angles (pitch, roll, yaw); (3) Attitude quaternion on x, y, z; (4) Rotation matrix (3x3); (5) Gravitational component of acceleration; (6) User component of acceleration; (7) Rotation rate; (8) Magnetic data; (9) Magnetic and true heading; (10) Latitude and longitude; (11) Altitude and accuracies; (12) Course; (13) Speed; (14) Sensor proximity. We have implemented a Decision Tree based algorithms described in [1] to classify cow feeding behaviors. Hardware configuration used for the implementation of the architecture was a High Performance VPS XXL Contabo with followings characteristics: 10 vCPU Cores, 60 GB RAM, 1.6 TB SSD, 1 Gbit/s Port [3]. Four different configurations of our architectures were benchmarked. All of theses combinations implement Kafka as log data storage. Kafka temporary stores data before their processing by Apache Storm or Apache Samza. Two kinds of databases were tested Apache HBase and Apache Druid in combinations with Storm and Samza to obtains 4 configurations (see Fig. 2).

Apache Kafka is an open-source distributed event streaming platform which is mainly used as temporary log storage. Apache Storm and Apache Samza are two open source distributed real-time computation systems. Apache Storm is also scalable, fault-tolerant which guarantees that the data will be processed in case of incident. An Apache Storm topology ingests streams of data and processes those streams in arbitrarily complex ways. Apache Samza is a scal-

[3] https://contabo.com/en/vps/

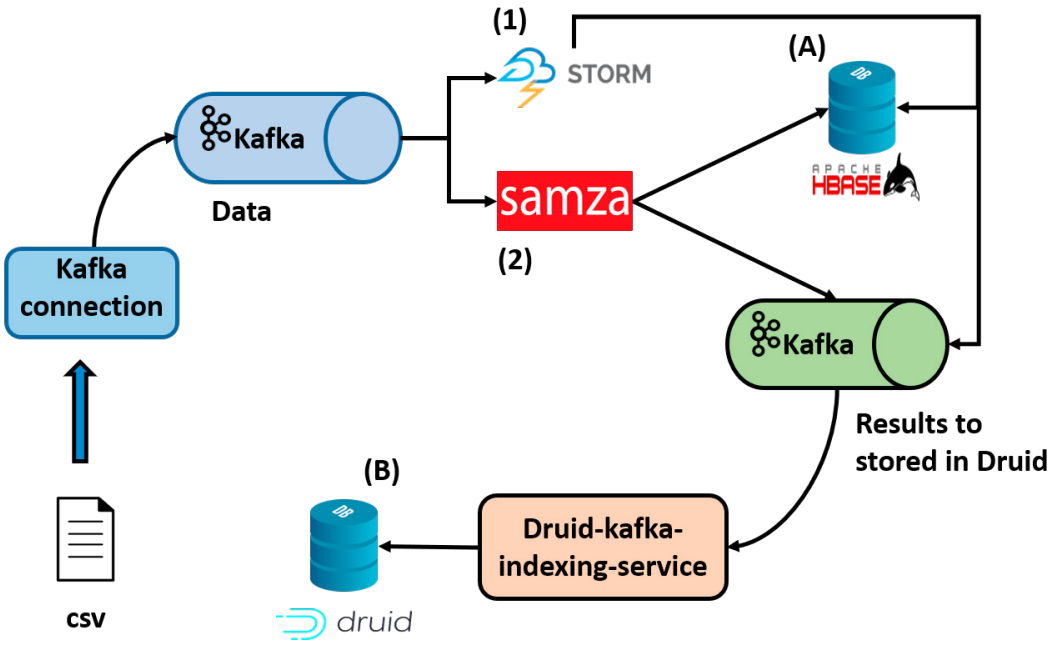



Fig. 2. Implementation of the Kappa architecture to be optimized.

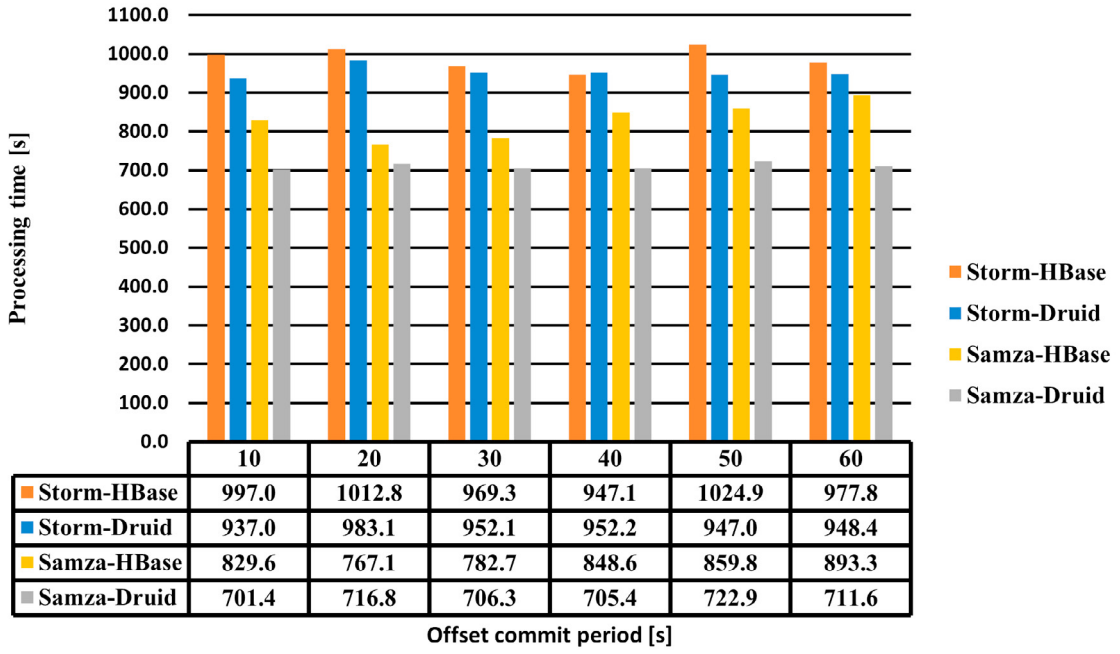
able data processing engine that allows to process and analyze data in real-time. Apache HBase is an open-source, distributed, versioned, non-relational database built to provide random, real-time read/write access to Big Data with large tables composed of billions of rows and millions of columns. Apache Druid is an open source distributed data store with high performance real-time analytics capabilities and combining concepts of data warehouses, time series databases, and search systems. We have chosen Apache Storm because it is a well known and used tool for data processing while Apache Samza is a promising stream processing software. We have also decided to work with NoSQL databases to be more flexible and adaptable to the evolution of data structure over the time. Tested configurations are: (*1A*) Apache Kafka - Apache Storm - HBase; (*1B*) Apache Kafka - Apache Storm - Apache Kafka - Druid; (*2A*) Apache Kafka - Apache Samza - Apache HBase; (*2B*) Apache Kafka - Apache Samza - Apache Kafka - Apache Druid. In configurations 1B and 2B results of stream processing is store in a new Kafka topic and then ingested by the service Druid-Kafka-Indexing service which ingests data directly in the Kafka. While in configuration 1A and 1B data is directly store in HBase.

## 5. Results

The results of the experimentation allows us to demonstrate the impact of the memory (ram) allocated and offset commit at Kafka level on global performances of Kappa architecture.

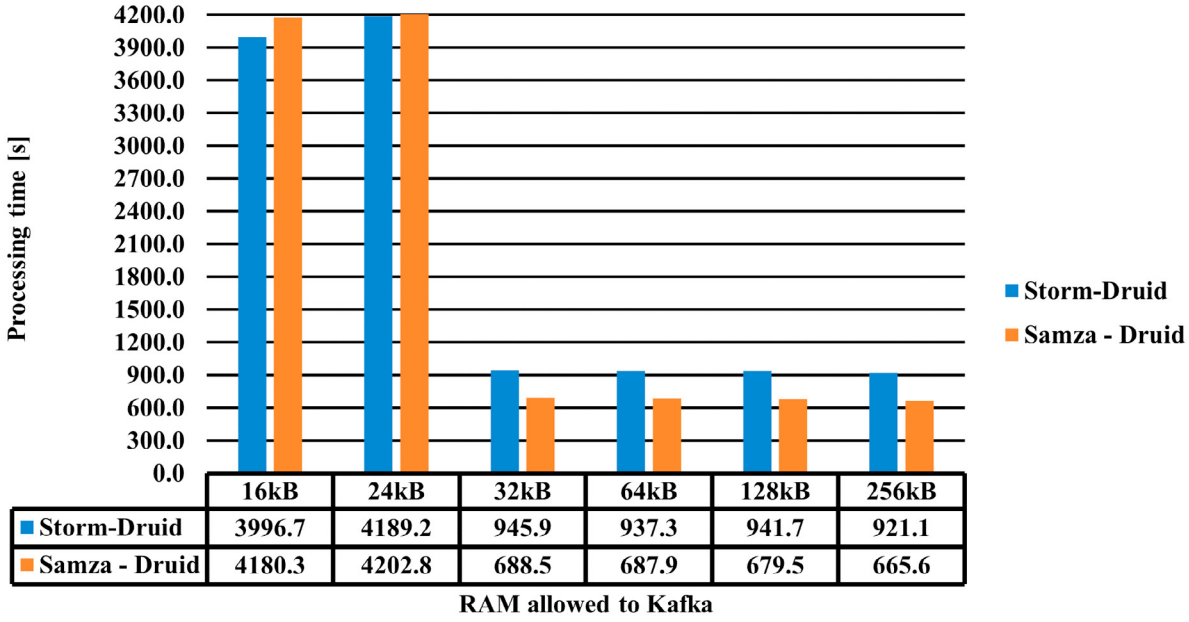
The Fig. 3 presents the results of the benchmark used. For each Kappa's configurations defined on Fig. 2, the processing time per Kafka's offset commit period has been measured. The processing time measured is the time elapsed between data reading from csv file, the treatment by Apache Storm or Apache Samza and the storage into one NoSQL database. The experiments were repeated ten times per offset commit period and the values are the mean values.

On Fig. 3, by comparing the performances, we can notice that the combinations with Apache Druid gave better processing time than the others with Apache HBase. The comparison between the data processing tools shows that Apache Samza gave better performance than Apache Storm.




| | 10 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|
| Storm-HBase | 997.0 | 1012.8 | 969.3 | 947.1 | 1024.9 | 977.8 |
| Storm-Druid | 937.0 | 983.1 | 952.1 | 952.2 | 947.0 | 948.4 |
| Samza-HBase | 829.6 | 767.1 | 782.7 | 848.6 | 859.8 | 893.3 |
| Samza-Druid | 701.4 | 716.8 | 706.3 | 705.4 | 722.9 | 711.6 |

Fig. 3. Impact of the offset commit period on the processing time.




| | 16kB | 24kB | 32kB | 64kB | 128kB | 256kB |
|---|---|---|---|---|---|---|
| Storm-Druid | 3996.7 | 4189.2 | 945.9 | 937.3 | 941.7 | 921.1 |
| Samza - Druid | 4180.3 | 4202.8 | 688.5 | 687.9 | 679.5 | 665.6 |

Fig. 4. Impact of the RAM allowed to Kafka on the processing time.

The combination Apache Samza - Apache Druid gave the best result of our benchmark. This result could be explained by the fact that Apache Druid is designed and optimized to ingest and read data. Apache Samza is also designed to optimize data treatment by its internal parallelism strategy.

The combination Apache Storm - Apache HBase presents less good result of our benchmark. Indeed, Apache Storm implements an internal mechanism to avoid congestion during the treatment of data. This mechanism could introduce a delay in the data treatment.

Another reason for this bad performance in comparison to the previous combination could be explained by the fact that Apache HBase is not optimized to quickly ingest data but to store a large amount of data.

The Fig. 4 presents the impact of the RAM allowed to the Kafka on the processing time. Regarding to our previous benchmark results, this choice was made to analyze the impact only for the combinations Apache Storm - Apache Druid and Apache Samza - Apache Druid. The experiments were repeated ten times per ram values and the represented values are the mean values.

On the Fig. 4, we can see that the processing time is only impacted with the RAM values lower than 32KB. This behavior could be explained by the fact that the memory allowed to Apache Kafka is a kind of buffer for messages processing. Therefore, when the memory allowed to Apache Kafka is insufficient to process incoming messages, Kafka stores the messages into his buffer with a delay greater than expected. The consequence is the increase of the data processing time. In this paper, the analysis of the impact on the processing time of the Kappa architecture in function of the offset commit period and the RAM allocated to the Kafka message queue was presented in the use case. At the end, we can conclude that the best optimized combination is the Apache Kafka as the message queue, Apache Samza as the data processing and Apache Druid as NoSQL database.

## 6. Conclusion and perspectives

Agriculture 4.0 is an IoT domain in full growth with needs to process large volumes of data in the shortest possible time. The performances of 4 combinations of software associating Kafka as message queue and mixing (Samza or Storm as processing software) with (HBase or Druid as database) to build the Kappa architecture have been evaluated. Afterwards, we have described the improvement of the Kappa architecture to optimize the speed of data ingestion and processing. This optimal Kappa architecture is implemented with Apache Samza which process data and Apache Druid to store them. Apache Kafka is used as a first message queue which play the role of temporary log storage before data ingesting by Samza. Then, data processing results are stored in a second message queue before its insertion in Druid database. This operation is achieved by the ”Kafka-indexing-service”, a Druid service that ingests data stored in Kafka message queue and insert it in Druid database. Moreover, in this paper, we have shown that this association of software outperforms classically associated Apache Storm with Apache HBase in Kappa Architecture. Afterwards, we have shown the impact on the overall performance of this optimal architecture of both RAM allocation and offset commit period at Kafka level.

In our future works, the architecture will be completed with a data lake to store raw data on long term and also develop an edge computing complement to process data at fog level in order to improve performances.

## Acknowledgements

This research is partially funded by Infortech and Numediart institutes. Authors would like thanks to Prof Jérôme Bindelle (ULiège - GxABT) which has provided us the dataset of diary cows used to achieve our experimentation.

## References

[1] Andriamandroso, A.L.H., Lebeau, F., Beckers, Y., Froidmont, E., Dufrasne, I., Heinesch, B., Dumortier, P., Blanchy, G., Blaise, Y., Bindelle, J., 2017. Development of an open-source algorithm based on inertial measurement units (imu) of a smartphone to detect cattle grass intake and ruminating behaviors. Computers and electronics in agriculture 139, 126–137. doi:10.1016/j.compag.2017.05.020.

[2] Bixio, L., Delzanno, G., Rebora, S., Rulli, M., 2020. A flexible iot stream processing architecture based on microservices. Information 11, 565. doi:10.3390/info11120565.

[3] Debauche, O., Mahmoudi, S., Andriamandroso, A.L.H., Manneback, P., Bindelle, J., Lebeau, F., 2017a. Web-based cattle behavior service for researchers based on the smartphone inertial central. Procedia Computer Science 110, 110–116.
[4] Debauche, O., Mahmoudi, S., Andriamandroso, A.L.H., Manneback, P., Bindelle, J., Lebeau, F., 2019. Cloud services integration for farm animals' behavior studies based on smartphones as activity sensors. Journal of Ambient Intelligence and Humanized Computing 10, 4651–4662.
[5] Debauche, O., Mahmoudi, S., Mahmoudi, S.A., Manneback, P., Bindelle, J., Lebeau, F., 2020. Edge computing for cattle behavior analysis, in: 2020 Second International Conference on Embedded & Distributed Systems (EDiS), IEEE. pp. 52–57.
[6] Debauche, O., Mahmoudi, S., Manneback, P., Tadrist, N., Bindelle, J., Lebeau, F., 2017b. Improvement of battery life of iphones inertial measurement unit by using edge computing application to cattle behavior .
[7] Debauche, O., Trani, J.P., Mahmoudi, S., Manneback, P., Bindelle, J., Mahmoudi, S., Lebeau, F., 2021. Data management and internet of things : A methodological review in smart farming. Internet of Things , 100378URL: https://www.sciencedirect.com/science/article/pii/S2542660521000226, doi:https://doi.org/10.1016/j.iot.2021.100378.
[8] Feick, M., Kleer, N., Kohn, M., 2018. Fundamentals of real-time data processing architectures lambda and kappa, in: Becker, M. (Ed.), SKILL 2018 - Studierendenkonferenz Informatik, Gesellschaft für Informatik e.V., Bonn. pp. 55–66.
[9] Fote, F.N., Mahmoudi, S., Roukh, A., Mahmoudi, S.A., 2020. Big data storage and analysis for smart farming, in: 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), IEEE. pp. 1–8. doi:10.1109/CloudTech49835.2020.9365869.
[10] Kreps, J., 2014. Questioning the lambda architecture. Online article, July 205.
[11] Persico, V., Pescapé, A., Picariello, A., Sperlí, G., 2018. Benchmarking big data architectures for social networks data processing using public cloud platforms. Future Generation Computer Systems 89, 98–109. doi:10.1016/j.future.2018.05.068.
[12] Persson, P., Angelsmark, O., 2017. Kappa: serverless iot deployment, in: Proceedings of the 2nd International Workshop on Serverless Computing, pp. 16–21. doi:10.1145/3154847.3154853.
[13] Roukh, A., Fote, F.N., Mahmoudi, S.A., Mahmoudi, S., 2020. Wallesmart: Cloud platform for smart farming, in: 32nd International Conference on Scientific and Statistical Database Management, pp. 1–4. doi:10.1145/3400903.3401690.
[14] Sanla, A., Numnonda, T., 2019. A comparative performance of real-time big data analytic architectures, in: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), IEEE. pp. 1–5. doi:10.1109/ICEIEC.2019.8784580.
[15] Wingerath, W., Gessert, F., Friedrich, S., Ritter, N., 2016. Real-time stream processing for big data. it-Information Technology 58, 186–194. doi:10.1515/itit-2016-0002.
[16] Zschörnig, T., Wehlitz, R., Franczyk, B., 2017. A personal analytics platform for the internet of things-implementing kappa architecture with microservice-based stream processing, in: International Conference on Enterprise Information Systems, SCITEPRESS. pp. 733–738. doi:10.5220/0006355407330738.