A cloud-based foundational infrastructure for water management ecosystem

Rayane El Sibai, Jacques Bou Abdo, Chady Abou Jaoude, Jacques Demerjian, Yousra Chabchoub, Raja Chiky

A cloud-based foundational infrastructure for water management ecosystem

Rayane El Sibai
Faculty of Engineering
Al Maaref University
Beirut, Lebanon
rayane.elsibai@mu.edu.lb

Jacques Bou Abdo
Faculty of Natural and Applied Sciences
Notre Dame University
Deir El Kamar, Lebanon
jbouabdo@ndu.edu.lb

Chady Abou Jaoude
Ticket Lab., Faculty of Engineering
Antonine University
Baabda, Lebanon
chady.aboujaoude@ua.edu.lb

Jacques Demerjian
LaRRIS, Faculty of Sciences
Lebanese University
Fanar, Lebanon
jacques.demerjian@ul.edu.lb

Yousra Chabchoub
LISITE
ISEP
Issy-les-Moulineaux, 92130, France
yousra.chabchoub@isep.fr

Raja Chiky
LISITE
ISEP
Issy-les-Moulineaux, 92130, France
raja.chiky@isep.fr

***Abstract*—Water monitoring is one of the critical battles of sustainability for a better future of humanity. 44 countries are considered at high risk of the water crisis, 28 of which are developing countries and have limited capacity to deploy a national scale solution. As a response to the United Nation's sustainability goals and initiatives, this paper proposes an intelligent water monitoring service which acts as a foundational infrastructure for all future water management systems. It also provides municipalities, Non-Governmental Organization and other private initiatives with the tools needed to establish local water monitoring in the scale of villages or rural areas with a very small initial investment.**

***Index Terms*—Cloud computing, data streams, data quality, sampling techniques, Software as a Service (SaaS).**

I. Introduction

In 2015, the United Nations General Assembly approved the following: "Transforming our world: the 2030 Agenda for Sustainable Development" [1]. In order to guide the actions of the international community towards global sustainability, several objectives related to resource management and its consequence were defined. The water-related goals were the following: (1) Eradicate hunger, ensure food security and nutrition, and promote sustainable agriculture (2) Ensure the availability and sustainable management of water and sanitation for everyone, and finally, (3) Protect, restore and improve the sustainable use of terrestrial ecosystems, manage forests sustainably, combat desertification, etc.

ICT solutions for water management have become extremely important as a global initiative. Intelligent water management systems [2], dynamic water pricing systems [3], peer-to-peer water trading systems [4], and e-agriculture [5] are at the core of UN's sustainability goals and are becoming very hot research topics. The development of efficient ICT solutions for water management relies on continuous updates of accurate and timely information, but the literature is short on water monitoring systems specialized in providing such information. It is inefficient and impractical to require every ICT solution to collect and process massive streams of data as a prerequisite to performing its task. The overhead resulting from adding a water monitoring component to every water management system is simply too expensive.

In this work, we are interested in designing an intelligent water monitoring service that helps in complementing the water management ecosystem and facilitates the development of ICT solutions for water management. The designed service collects, processes and delivers accurate, meaningful, timely, dynamic, and cost-efficient data, through standardized interfaces, to the systems needing it, as shown in Figure 1.

To do so, we propose a new cloud service that offloads the complexity of the central monitoring system with all its overheads to the cloud. These overheads include huge initial investment which could prevent NGOs and private initiatives from deploying a functional monitoring system, scalability issues, and maintenance issues. This service supports new business models and an ecosystem that can be created around the generated data while providing water network operators with a plethora of tools to make their decisions smarter, faster, or automatic. Artificial Intelligence can be designed to utilize standardized interfaces. The standardized interface is fundamental for the cloud computing ecosystem. It provides end-users with the capability to access cloud services. Analytics tools, big data applications, security management, and monitoring tools are among the commonly provided services. In the rest of this paper, we refer to Intelligent water management systems, dynamic water pricing systems, peer-to-peer water trading systems, and e-agriculture by water management systems.

The rest of the paper is organized as follows. Section II

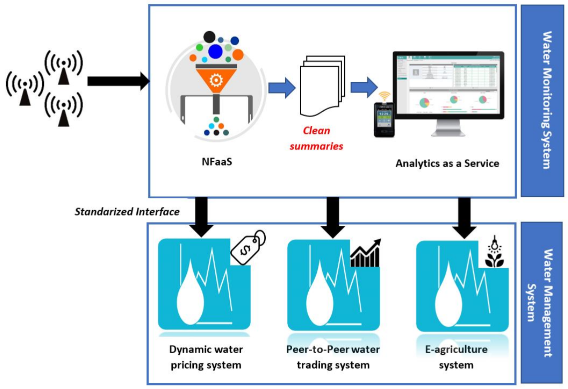

Fig. 1: Water management ecosystem.

presents the related works. Section III discusses the requirements for a functional central monitoring system. Section IV introduces our contribution, the "Native Filtering as a Service (NFaaS)". The paper ends with a conclusion.

II. Related Works

Cloud computing for environmental monitoring consists of using cloud infrastructure to control the data streams generated by Wireless Sensor Networks (WSNs). The adoption of cloud technology allows users and businesses to save time and costs since the cost of storage and the efforts of the infrastructure installation, configuration, and maintenance drops dramatically and become negligible. In addition, the estimation and planning of the required resources, and the use of excessive storage and computation capacities are no longer required as the resources can be flexibly adjusted as needed.

In [6], a cloud-based system has been proposed to monitor the changes that occurred in the environment. The system generates an alert and notifies the user when an anomaly is detected. The main benefit of this system is to reduce the amount of time needed to retrieve and analyze the data when compared to the traditional data analysis systems. [7] presented a cloud-based system for weather monitoring. It is composed of two layers: an interface layer to allow the interaction between the users and the cloud through a web server, and a database layer to allow storing and fetching the data coming from the WSNs base stations. [8] implemented a cloud system to store, process, and monitor the data related to the air quality and detect specific compounds and contaminants. Several parameters related to the air are observed and sent to the cloud by the sensors. [9] designed a semantic Extract-Transform-Load (ETL) framework on the cloud platform for monitoring and predicting air quality. At first, different data analysis techniques are applied to clean the data and remove duplicated values. Secondly, a semantic model is built so that to have a significant relationship between the data. Finally, several data mining algorithms are used to analyze and predict air quality.

A framework based on blockchain technology for dynamic water pricing has been proposed by [3], in which the water price is a function of the energy prices, which drops as the price of the energy required to pump the water drops. The deployment of water management systems in African countries requires several specifications such as continuous monitoring, remote control of water distribution, etc. The impact of using blockchain on intelligent water management systems in these areas was investigated by [2]. In [10], the authors concluded that it is difficult to deploy an Integrated Water Resources Management (IWRM) in real life. The required monitoring, management, and maintenance operations, on a national scale, required are very expensive.

III. Central monitoring system requirements

In this section, we show the infrastructure requirements for a functional central monitoring system as calculated in our previous work [11]. As will be shown, these requirements can be an obstacle facing the deployment of a water monitoring system by states, municipalities, NGOs, or private initiatives in developing nations, who are in desperate need of such a system.

The calculated infrastructure includes the computational resources required to run the central monitoring system which is responsible for data cleaning and summarization, and the number of servers needed to ensure high availability of the service. The used dataset is issued from the water sensors deployed in the Paris region. The recorded data are structured data streams that have both spatial and temporal characteristics. Each record observation is composed of two fields: the recording date of the measure, and the value of the measure (water consumption volume). These data are regularly generated by the sensors with a frequency of one observation every 15 minutes. The specifications of our server on which we run the central monitoring system are the following: RAM: 4 GB, System Disk: 120 GB, and Processor: 2.7 GHz Intel Core $i5-8305G$.

1) Response time: The execution time of the sampling process includes the time of reading, sampling, and writing the data in the summary, and is dependent on the following factors: (1) Number of received observations for each stream, which depends on the stream rate or frequency of the sensor, (2) Number of streams received simultaneously, (3) Sampling rate, (4) Window size, and (5) Sampling technique.

We evaluate in Figure 2 the response time taken by the central monitoring system to construct a summary for each received data stream, using the sampling algorithms previously discussed. One can notice that the response time increases linearly with the increase of the number of streams to be processed. The following equation stands for the response time of the sampling filter as a function of the number of streams to be processed:

$$Response\ time = a \times Number\ of\ streams + b \quad (1)$$

where b is a fixed amount of time needed to initialize the sampling filter, and $a \times Number\ of\ streams$ is an incremental component proportional to the number of received data streams.

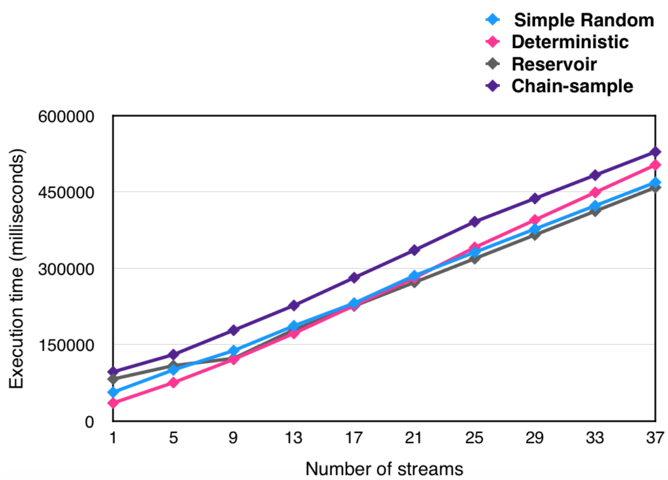

Fig. 2: Execution time of the sampling process according to the number of streams, locally.

Figure 2 also shows that the Chain-sample algorithm has the highest execution time compared to other sampling algorithms. This is due to the collision problem that it suffers, and the number of data to be written in the summary, which is higher compared to that of other sampling algorithms [11]. Notice that the size of the window has a high impact on the execution time taken by Chain-sample. This time increases as the window size increases, as demonstrated in [12]. The figure also shows that the Deterministic sampling algorithm has the smallest execution time, as it is the simplest sampling method.

Figure 3 represents the cleaning process's response time. One can conclude that the cleaning time needed is not linearly dependent on the number of received streams. In fact, for a single data stream, this cleaning time depends on the following parameters: the number of received observations, the amount and the distribution of missing data and outliers. As these parameters vary widely from one stream to another, the cleaning time cannot be predicted in advance.

2) Service availability: Increasing the availability of a system involves maximizing the percentage of time during which the system is operational. To protect the system against unexpected overloads and to avoid the system balancing can be used to ensure high availability. It consists of using a set of servers in which the incoming streams are evenly distributed to help reduce the load on a single server. We show in Figure 4 the calculated number of servers needed to run the central monitoring system while ensuring high availability for the stream rates discussed above.

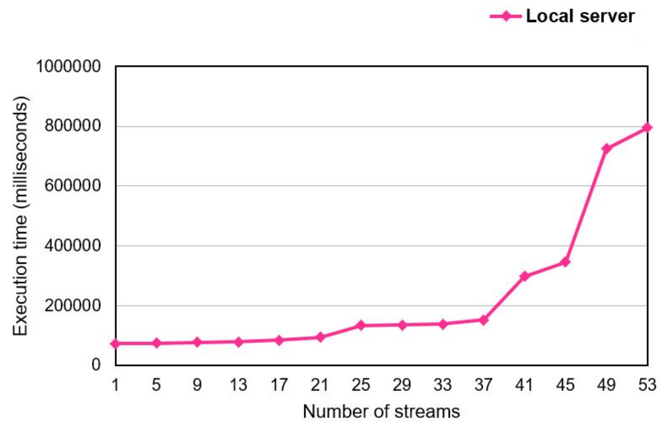

Fig. 3: Response time of the cleaning process according to the number of streams.

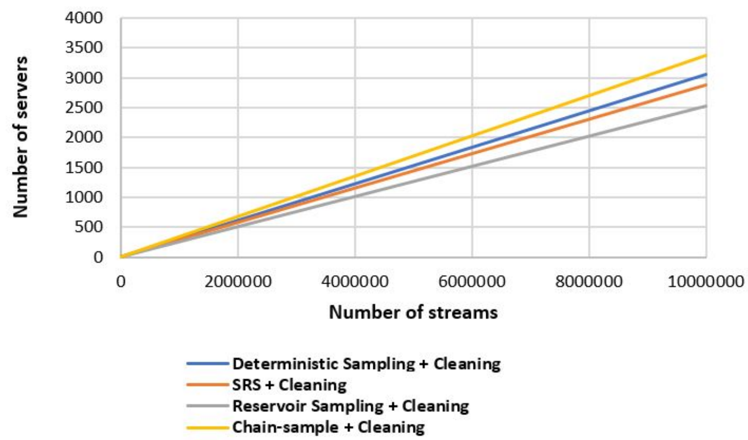

Fig. 4: Required number of servers to ensure high availability.

As shown in Figure 4, a local small-scale monitoring system of 1 million sensors requires a data center of 253 to 337 servers for stream cleaning and processing. This data center is too expensive for private initiatives, NGOs, or municipalities to install, operate, and maintain thus jeopardize their willingness to deploy such a critical system. Consequently, water management systems relying on the data generated from the monitoring system will be jeopardized as well.

IV. Native Filtering as a Service (NFaaS)

To provide states, municipalities, NGOs, or private initiatives the possibility to deploy a water management system without the burdens of water monitoring systems, we propose a new cloud service called Native Filtering as a Service, which is presented in this section. Sensor data streams are arriving at a high-speed. They produce a huge mass of data, impossible to store entirely in due time. On their side, erroneous, inaccurate, or inconsistent data cause much damage as for the supervision and detection of abnormal phenomena in the monitored network. It is, therefore, necessary to filter these data on the fly and to store only those that are relevant by producing summaries. Native filtering consists of filtering qualitatively and quantitatively, in real-time, the received sensor data to overcome the poor data quality and huge data volume problems.

The native filtering process consists of two sub-processes: the cleaning filter, also called a qualitative filter, and the sampling filter, also called a quantitative filter. When the qualitative filter receives the data from the sensors, it proceeds to check and to improve their quality. Thereafter, the data are summarized by the quantitative filter. Native Filtering as a Service (NFaaS) consists of providing the cleaning and sampling filters as services via cloud computing. Therefore, clients can access and benefit from the cleaning and sampling applications over the Internet, on-demand, with a payment based on their use. Once the data are clean and summarized, it is essential to store them, so that they are available for future analysis either by their owners or by other users interested in these data. Eventually, the visualization of the obtained data analysis results is essential to help the network explorer to understand the monitored environment. That is where the Big Data as a Service (BDaaS) comes in.

BDaaS provides the users with a remote cloud storage server to store and edit their summaries, as well as an analytics platform for the analysis of these summaries and the visualization of the results. The cloud-based sensors data streams processing flow is shown in Figure 5. Every instance of NFaaS has its own geographical coverage, specialized analytics, and scope. Water management systems may be interested in retrieving data from multiple NFaaS instances or consortiums. Multiple NFaaS instances can store their output on a consortium blockchain, as shown in Figure 5, and this provides endless options for an ecosystem of brokers, regulators, investors, and innovators.

A. Overview of native filtering services

The native filter takes as input several data streams, issued from the sensors in their native format, and gives in output new data streams in the same format. It consists of two sub-filters: the qualitative filter (Cleaning as a Service) to clean the data by detecting and deleting errors and duplicated data and predicting missing and erroneous data, and the quantitative filter (Sampling as a Service) which is responsible for reducing the large volume of data by using several sampling algorithms. Our approach for managing the data stream quality is to first evaluate the data quality according to several dimensions, and then, improve it to obtain reliable and effective data analysis results. Since data streams are volatile, once expired, they are no longer available for analysis. Therefore, all the needed queries have to be defined before the arrival of data streams.

However, new requirements may appear after the arrival of the stream. In this case, the data stream management system cannot answer new queries. One of the solutions to palliate this problem is to store an extract of the stream in a compact structure, called summary. The cleaning and sampling services will be discussed in Section IV-B.

In figure 6, we show the use case diagram for NFaaS cloud service. The class diagram of the native filtering services is presented in Figure 7.

B. Services explanation

Data analysis results depend on data quality. Sensors data are often of bad quality. Since the conclusions and decisions that the network explorer will take are based on the data, this can lead to erroneous results and faulty decisions. One solution to deal with this problem is to use sensors with high precision and to deploy redundant sensors to cover the breakdown of a given sensor. Nevertheless, this approach is very expensive as it requires very high costs for the sensors. That is why we opted for a software-based solution where the data quality is evaluated and improved using several complementary methods. Our solution is based on the Total Data Quality Management (TDQM) methodology [13]. At first, data quality dimensions are defined and measured. Then, a set of quantitative metrics are produced. Finally, several actions are taken to enhance data quality.

Data quality metrics we are dealing with are the accuracy, completeness, reliability, and confidence. These dimensions are chosen because of their important impact on the data analysis results. The accuracy dimension represents the point at which the sensor readings are close to the real observations. In sensor networks, sensor readings may deviate from the real observations due to several factors such as the physical sensor failure, sensor malfunction, etc. Data accuracy can be improved by detecting and removing these erroneous data.

Errors are of three types: outliers, spikes, and stuck-at. Stuck-at errors can be detected using the CUSUM algorithm that we proposed previously [14]. Outliers and spikes errors are detected using a set of static rules. In the context of the water distribution network, any reading value that exceeds the maximum theoretical flow value (called $debit_Max$) that a sensor can emit will be considered as an outlier. The $debitMax$ value is calculated using the Hazen-Williams formula. This latter determines the maximum transportable flow rate through a given pipe according to the physical properties of the pipe and the pressure drop caused by the friction of the used material. The $debitMax$ value is calculated according to the Hazen-Williams formula, where DH is the altitude difference between the two extremities of the pipe, in $meters$, L is the length of the pipe, in $meters$, C is the friction coefficient of the used material, and D is the diameter of the pipe, in $meters$. The accuracy degree for each sensor reading takes the following values: 0 if the value is erroneous, otherwise, 1. The $debitMax$ value is calculated as follows:

$$debitMax = 0.28 \times ((DH/L)^{0.54}) \times (C \times D)^{2.63} \quad m^3/sec \tag{2}$$

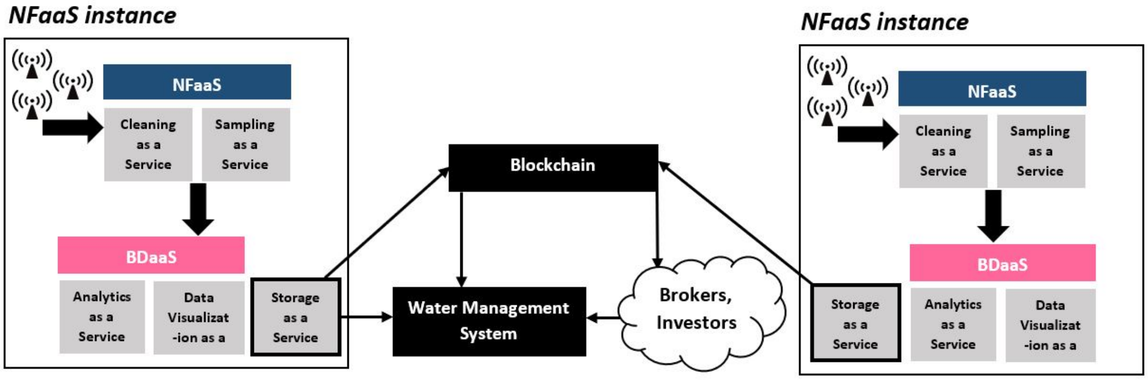

Fig. 5: Cloud-based sensors data streams processing.

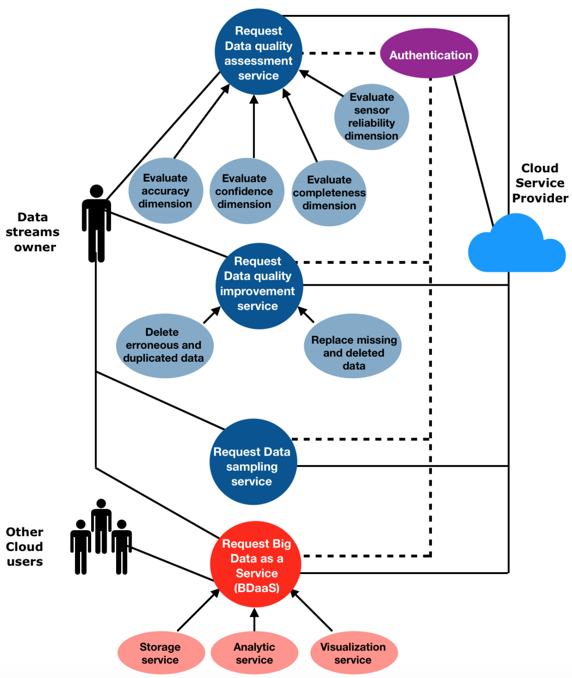

Fig. 6: Use case diagram of the Native filtering as a Service (NFaaS).

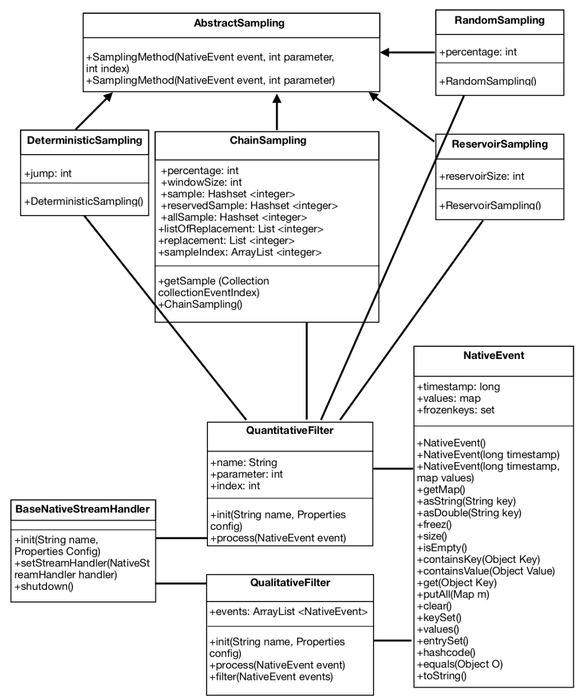

Fig. 7: Class diagram of the Native filtering services.

The completeness dimension depends on the frequency for recording the data in the environment and the frequency for communicating the recorded observations to the data center or the server. Based on the known stream arrival rate, the amount of missing data at each instant t (in *hour*) can be computed as follows:

$$cumulativeMissing_t = [\frac{\sum_{t_0}^{t} Number\ of\ missing\ data}{(t - t_0) \times streamRate}] \tag{3}$$

The sensor reliability dimension is based on the accuracy dimension. It is measured by the $cumulativeError$ degree which depicts the percentage of cumulated erroneous values.At each instant t, the $cumulativeError$ degree is calculated as follows:

$$cumulativeError_t = [\frac{\sum_{t_0}^{t} Number\ of\ erroneous\ data}{\sum_{t_0}^{t} Number\ of\ received\ data}] \tag{4}$$

The confidence dimension depicts the degree of trustiness of the sensor reading. This degree depends on the originality of the data. When a value is missing or erroneous, it has to be regenerated to enhance its quality. In this case, its $confidence$ degree is the proportion of consecutive missing or erroneous values around it. x is calculated relative to a given time period as follows, where m is the number of consecutive missing values, IAD is the Interpolation Allowed Duration (in $hours$) representing the maximum missing data period allowed to replace the data.

$$confidence = [1 - \frac{m}{IAD \times streamRate}] \quad (5)$$

Once the quality dimensions are evaluated, the cleaning filter proceeds to enhance data quality by deleting erroneous and duplicated data and replacing erroneous and missing data. The native filtering solution uses sampling algorithms to summarize the data and construct a summary. The singularity of this structure lies in its ability to perform various data analysis tasks on the stream and to answer in an approximate way to any request, whatever the period of time investigated. Different sampling algorithms can be applied by the native filtering solution. The choice of the sampling technique depends on the required precision regarding the data analysis results, and the constraints related to the processing time needed by the algorithm. Four sampling methods were implemented in the sampling filter: Simple Random Sampling (SRS), Deterministic sampling, Reservoir sampling, and Chain-sample. More details about these algorithms can be found in [15].

C. Big Data as a Service (BDaaS) in standardized interfaces

The built summaries from the NFaaS are sources of BDaaS. They can be published on the cloud allowing their owners to get a return on investment. Data is a valuable asset for every organization. The intelligent use of data can help make decisions based on real facts, and thus, improve the business processes by minimizing the risks and increasing the business in general. The concept of DaaS is that the data, regardless of its source and type, can be cleaned, enriched, integrated into a centralized infrastructure, and made available to different users and applications on the cloud. In this way, DaaS allows the users to obtain the appropriate data and to use it directly to meet their needs, and the data owners to get a return on investment.

Data can be used and analyzed to improve the performance of systems and decision support applications, as well as for risk assessment. While the benefits of Big Data are significant, many challenges remain. They are related to the high volume of the data, their velocity, confidentiality, and variety. Big Data analysis involves several distinct tasks such as data acquisition, data analysis, and data modeling. The use of the cloud environment for processing Big Data is an ideal solution, both scalable and adapted to huge data volume. In our monitoring system, data analytic tools can be used in detecting abnormal phenomena such as micro-variations of certain parameters that have an impact in terms of risk, variation, or non-nominal frequency, etc, as soon as possible, thus, allowing to save considerable amounts of potable water. Finally, data visualization is essential for interpreting the data. Cloud computing provides users and businesses with visualization tools. It allows them to spend their time in visualizing their data without having to worry about the infrastructure.

V. Conclusion

Water monitoring is a critical challenge for humankind since we share the same globe, breath the same air, drink the same water, and share the same DNA. We are not short on solutions, but we are in desperate need of practical ones that are implementable in developing countries and rural areas. Successful solutions are not necessarily the most technically advanced, but the ones with successful business models. In this paper, we proposed a water monitoring cloud service that facilitates the development of critically needed water management systems. The strength of this proposal is multi-folded. First, it doesn't rely on the government's maintenance which can be very inefficient in developing countries. The second strength is that it creates an ecosystem around the generated data which can help develop the areas using these services

Acknowledgment

This work was funded by the FUI 17 project WAVES[1]. The project aims to design and develop a monitoring platform for the supervision of water distribution networks. The goal is to develop a relevant solution for the realization of a decision support system for the network operators and explorers.

References

[1] UN Desa et al. Transforming our world: The 2030 agenda for sustainable development. 2016.

[2] Eustace M Dogo, Abdulazeez Femi Salami, Nnamdi I Nwulu, and Clinton O Aigbavboa. Blockchain and internet of things-based technologies for intelligent water management system. In *Artificial Intelligence in IoT*, pages 129–150. Springer, 2019.

[3] Gheorghe Grigoras, Nicu Bizon, Florentina Magda Enescu, José Manuel López Guede, Guadalupe Flores Salado, Raymond Brennan, Connie O'Driscoll, Megersa O Dinka, and Mohamed Gar Alalm. Ict based smart management solution to realize water and energy savings through energy efficiency measures in water distribution systems. In *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–4. IEEE, 2018.

[4] Seung Jae Pee, Jong Ho Nans, and Ju Wook Jans. A simple blockchain-based peer-to-peer water trading system leveraging smart contracts. In *Proceedings on the International Conference on Internet Computing (ICOMP)*, pages 63–68. The Steering Committee of The World Congress in Computer Science, Computer …, 2018.

[5] Yu-Pin Lin, Joy R Petway, Johnathen Anthony, Hussnain Mukhtar, Shih-Wei Liao, Cheng-Fu Chou, and Yi-Fong Ho. Blockchain: The evolutionary next step for ict e-agriculture. *Environments*, 4(3):50, 2017.

[6] K Prathibha and Pawan Hegde. A real-time system for environmental study based on cloud computing. In *9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2018.

[1] https://www.waves-rsp.org

[7] E Kanagaraj, LM Kamarudin, A Zakaria, R Gunasagaran, and AYM Shakaff. Cloud-based remote environmental monitoring system with distributed wsn weather stations. In *Sensors*, pages 1–4. IEEE, 2015.

[8] Patricia Arroyo, José Luis Herrero, José Ignacio Suárez, and Jesús Lozano. Wireless sensor network combined with cloud computing for air quality monitoring. *Sensors*, 19(3):691, 2019.

[9] Yue Shan Chang, Kuan-Ming Lin, Yi-Ting Tsai, Yu-Ren Zeng, and Cheng-Xiang Hung. Big data platform for air quality analysis and prediction. In *27th Wireless and Optical Communication Conference (WOCC)*, pages 1–3. IEEE, 2018.

[10] Asit K Biswas. Integrated water resources management: a reassessment: a water forum contribution. *Water international*, 29(2):248–256, 2004.

[11] Rayane El Sibai, Jacques Bou Abdo, Jacques Demerjian, Yousra Chabchoub, Raja Chiky, and Kablan Barbar. Information technology infrastructure for data streams native filtering. In *Middle East and North Africa Communications Conference (MENACOMM)*, pages 1–6, 2018.

[12] Rayane El Sibai, Yousra Chabchoub, Jacques Demerjian, Zakia Kazi-Aoul, and Kabalan Barbar. A performance study of the chain sampling algorithm. In *Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 487–494. IEEE, 2015.

[13] Richard Y Wang. A product perspective on total data quality management. *Communications of the ACM*, 41(2):58–65, 1998.

[14] Rayane El Sibai, Yousra Chabchoub, Raja Chiky, Jacques Demerjian, and Kablan Barbar. An in-depth analysis of cusum algorithm for the detection of mean and variability deviation in time series. In *International Symposium on Web and Wireless Geographical Information Systems*, pages 25–40. Springer, 2018.

[15] Rayane El Sibai, Yousra Chabchoub, Jacques Demerjian, Raja Chiky, and Kablan Barbar. A performance evaluation of data streams sampling algorithms over a sliding window. In *Middle East and North Africa Communications Conference (MENACOMM)*, pages 1–6, April 2018.