Intelligent Monitoring of Water Consumption in Urban Environments using Computing and FOG-SWM

Renan S. Okada[1], Wesley dos Reis Bezerra[1], Carlos Becker Westphall[2]

[1]Instituto Federal Catarinense (IFC)

Rio do Sul – SC – Brazil

[2]Universidade Federal de Santa Catarina (UFSC)

Florianópolis – SC – Brazil

{renanokada2000,wesleybez,carlosbwestphal}@gmail.com

Abstract. *The increasing urbanization has brought significant challenges to water resource management, requiring advanced technological solutions for monitoring and controlling water consumption. In this work, we propose a system based on the FOG-SWM architecture, which combines fog computing and real-time data analysis to optimize water distribution and anomaly detection. We employ data processing and filtering techniques using Python, MapReduce in Pig Latin, and pattern recognition with the Isolation Forest algorithm. The results demonstrate the model's effectiveness in identifying unusual consumption variations and assisting strategic decision-making for urban water sustainability.*

1. Introduction

Excessive water consumption can lead to problems such as waste of natural resources and increased water treatment and distribution costs. According to Marçal, Brito and Diniz (2023), its causes include leaks in pipes, poorly regulated taps and faulty toilet flushes, among others. Unfortunately, this is a common problem in Brazil, where around 40% of tap water is lost during pumping in the public system. Despite being a vital resource for life, many people have not yet adopted sustainable habits regarding water use. According to data from the National Sanitation Information System (2019), water consumption in Brazil per person is 154 liters per day, while the UN recommendation is 110 liters per day. The growing demand for water and the scarcity of water resources in many regions of Brazil are serious problems that need to be addressed. According to Braga (2023), the water sector is currently facing significant challenges, including the need to improve the efficiency of water use in different sectors, both in urban and rural environments. In some cities, losses in water supply systems can reach more than 50%.

Computing is crucial, offering tools and methods to transform water resource management. Water consumption monitoring systems, Big Data, and machine learning can identify consumption patterns and detect anomalies automatically and accurately. Such technologies enable data collection and processing, while data analysis platforms allow the visualization and interpretation of this data to make informed decisions about water distribution and use.

In this context, the question is: "How can we detect irregularities in the water consumption flow?" This work aims to determine and focus on solving this problem, as the lack of detection leads to excessive use and waste of fresh water.

In this sense, an organized algorithm for detecting intelligent water anomalies can be a viable solution for water management. According to Vasconcelos (2023), a system that enables the detection of anomalies in water consumption has benefits such as reduced consumption, water waste, and cost savings.

Therefore, this work proposes to develop a system that can process historical water consumption data with the Hadoop tool for detecting water consumption anomalies. We can list other contributions, such as the following items:

- Collect and preprocess water consumption data for use in the Hadoop tool with Map-reduce;
- Use the Isolation Forest algorithm to flag and expose anomalous points in the water consumption flow.
- Perform a critical analysis of the results obtained and discuss the implications of using the chosen model in the computational, technological, and water resource management environments;
- Propose recommendations to improve the proposed approach and indicate possible directions for future work in the area.

The remainder of the work is organized as follows: Section 2 presents important concepts to help readers better understand the work carried out; Section 3 presents the development of the project; and finally, Section 4 presents the conclusions and proposed future work.

2. Background

Da Paz, Esquerre, and Sartori (2018) highlight that anomalies, also known as outliers, refer to inconsistent examples in a data set. Thus, considering the problem presented and the research question defined, the proposal is to develop a solution capable of detecting anomalies in water consumption efficiently and accurately to support more effective decision-making regarding this natural resource.

The solution seeks to contribute to the field of studies of a distributed system with cloud storage and decision-making in the computing area and, consequently, promote environmental sustainability. In this sense, it is a viable alternative to using technologies suitable for the scope of this work to solve the problem, presenting the following:

1. FOG-SWM architecture;
2. Integration with Map-reduce in the Pig Latin language;
3. Data processing and filtering with Python;
4. Matplotlib: Analysis and Plotting of consumption data;
5. Implementation of Isolation Forest to detect anomalies;
6. Visualization of Results.

The illustration of the FOG-SWM (1) (Smart Water Management) architecture together with the Map-Reduce processing model (2), part of the Hadoop ecosystem, in the Pig Latin language (2), becomes an interesting combination. Silva (2020) highlights that from its origin in IoT (Internet of Things) end devices to the cloud, they are used for data processing, storage, and control distribution.

With the Python language (3), data cleaning, filtering, and preprocessing were possible in the dataset-processing environment with its specific and influential libraries. In addition to enabling data visualization, libraries such as Matplotlib (4) result in graphs that facilitate data interpretation.

The Isolation Forest algorithm was used to detect anomalies (5), which was used to identify water consumption patterns that deviate significantly from normal behavior, allowing for the isolation and highlighting of anomalous observations. This favored identifying possible problems, such as leaks or fraud in the water supply system. Finally, the results (6) of the analyses were presented in a clear and accessible manner through intuitive graphs.

It should be noted that this document does not detail all of these steps due to the limitation in the number of pages.

2.1. Methodology

This work followed a methodological approach that involved bibliographic research, data collection, data preprocessing, Map-Reduce implementation, results analysis, and critical discussion.

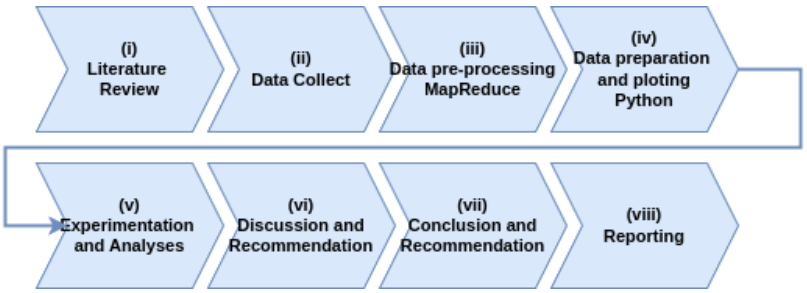

Figure 1. Project Development Phases.

As seen from the workflow above, Figure 1 presents the workflow developed in this work.

Literature Research and Review (**item i**) involved a literature review to understand the existing techniques for detecting anomalies in water consumption and the main approaches used in the literature. This review also served as a basis for justifying the choice of the FOG-SWM architecture and Hadoop tool to solve the proposed problem. In addition to providing a solid theoretical basis, it allowed the understanding of relevant concepts, techniques, and approaches in the area.

Continuing with Data Collection and Preprocessing (**items ii** and **iii**), water consumption data was collected to implement Map-Reduce on the Hadoop platform. After that, queries were made in academic databases such as IEEE Xplore, Google Scholar, and Scopus, as well as books, scientific articles, dissertations, and theses that address the topic in question. Finally, the data was preprocessed using Map-Reduce, shaping the data to the desired model.

The use of the Python programming language (**item iv**) and machine learning and distributed computing libraries was crucial for cleaning, filtering, and visualizing the data in graphs after item iii, making it possible to highlight each location's water consumption situation. This resulted in the creation of components for data collection, storage, large-scale processing, and analysis of water consumption data.

After completing the entire data preparation process and allowing the model to reach the state closest to ideal for identifying patterns and trends, the Isolation Forest algorithm (item v) was applied to detect anomalies in water consumption, highlighting points that deviated from the established normality.

Following this, experiments were carried out (**item v**) to evaluate the developed system's performance and effectiveness. Actual water consumption data were collected, and appropriate metrics were used to verify the system's ability to detect anomalies accurately and efficiently.

The work's conclusions (items vi and vii) were drawn up as evidence, highlighting the main findings, contributions, and limitations. Recommendations for future research and improvements to the system were presented, considering technical, methodological, or application aspects.

Finally, all stages of the work were documented (**item viii**), including a detailed description of the methodology, the results of the experiments, the analyses performed, and the conclusions obtained. The final text of the work was written, following the norms and guidelines of the educational institution, ensuring the quality and clarity of the presentation.

2.2. The Data Set

In the search for efficiency in detecting anomalies in distributed systems, the quality and integrity of the dataset used are of utmost importance. With this in mind, this project is based on the distributed processing of an extensive and detailed dataset, which reflects the efficiency in collecting and processing water consumption from one or more distribution networks, homes, buildings, industries, and businesses located in locations in New York City, belonging to the New York City Housing Authority (NYCHA).

The New York City Housing Authority (NYCHA) is recognized as the most significant public housing entity in North America. Its history dates back to 1935. According to the New York government (2023), it aims to provide adequate and affordable housing for low—and moderate-income New Yorkers.

Representing a significant portion of the city's population, NYCHA houses approximately 1 in 17 New Yorkers, translating to more than half a million individuals

residing in public housing units and participating in programs like PACT (Permanent Commitment to Affordability Together).

Figure 2. Leadership and Departments - NYCHA

As illustrated in Figure 2, NYCHA has over 177,000 apartments spread across more than 2,400 buildings within 335 housing projects. It is not only a housing provider but also a hub for essential services focused on economic development, support for youth and seniors, and a variety of social services. The breadth of NYCHA's services reflects its strategic importance in the urban fabric, functioning almost as a microcosm within the New York metropolis (NYC GOV, 2023).

The dataset used in this project, titled "New York Housing & Development Water Consumption And Cost," includes water consumption and cost data from 2013 to 2020 collected and made available by NYCHA. It is a comprehensive and detailed compilation of information allowing in-depth analyses of water consumption patterns. This dataset was obtained from Kaggle.com, a widely recognized online platform hub for data scientists and Machine Learning (ML) enthusiasts worldwide.

Borough	BROOKLYN
Development Name	FENIMORE-LEFFERTS
Account Name	FENIMORE-LEFFERTS
Location	BLD 15
Meter Scope	BLD 15
TDS # (Identifier for NCYHA Housing Complexes)	205
Vendor Name	NEW YORK CITY WATER BOARD
Revenue Month	2019-09
Service Start Date	08/22/2019
Service End Date	09/22/2019
Number Days	31
Meter Number	E20527605
Current Charges	144.68
Rate Class	Basic Water and Sewer
Consumption HCF	14
Water&Sewer Charges	144.68
Other Charges	0

Table 1. Columns present in a row of the set New York Housing & Development Water Consumption And Cost (2013 - 2020)

3. Development

This section presents the relevant plotting results after data filtering and application of the Isolation Forest method to identify anomalies in water consumption patterns in some studied locations in New York. This machine learning method isolated observations that deviate from the normal consumption pattern, providing valuable insights for understanding and managing water resources.

3.1 Data Compilation

Data compilation was a crucial step in the process, in a distributed manner, analyzing large data sets and involving several tools and techniques. An example of four rows extracted from the total set to be compiled can be seen in Table 1:

Name	Date	Consumption_HCF	Consumption_KiloLiters
WILLIAMSBURG	2017-10-01	45.0	127.44
WILLIAMSBURG	2018-11-01	44.0	124.608
WILLIAMSBURG	2019-12-01	6.0	16.992
WILLIAMSBURG	2020-03-01	6.0	16.992

Table 2 - Sample of data extracted from the original file of NYCHA residences - Columns Name, Date, Consumption in Hcf, and Consumption in Kilo Liters.

The process was divided into a few main steps, each using specific tools to perform different data processing and analysis tasks:

1. Use of the Apache Pig tool;
    a. MapReduce operations with Pig Latin;
2. Refinement of the resulting data with Python;
3. Anomaly Detection with Isolation Forest.

As a first step, the Apache Pig tool, an integral part of the Hadoop ecosystem, was used to perform MapReduce operations (**1**) within the Hortonworks Data Platform (HDP), a well-known system for distributed processes, in Virtual Box. The Apache Pig tool (1) allowed the writing of scripts for batch data processing, which are converted into MapReduce tasks, taking advantage of the efficiency and scalability of Hadoop. Together with the sandbox tool, the Hadoop file system was used to store the data in a distributed manner, ensuring high availability and fault tolerance. In order to distribute the system efficiently within Hadoop, the Pig tool proved beneficial since it transforms large volumes of data and performs join and sort operations. Pig Latin (1.a), the language used in Pig, is a data procedural language that simplifies technical aspects of MapReduce.

In **step 2**, the data was refined with Python. After the initial processing with Pig, the data was refined, cleaned, sorted, and suppressed null or zero fields using the Python programming language, which is widely recognized for its flexibility and libraries for data analysis.

Pandas were used to manipulate and prepare the data, such as removing rows where the "Account Name" had consumption in HCF equal to zero and selecting accounts with consumption records for at least 84 months (7 years).

Afterwards, a new column called 'Consumption_KiloLiters' was implemented. Using the Pandas library, water consumption values in HCF (Hundred Cubic Feet) were converted to kiloliters (KL), using the formula HCF is 100 FEET = 2.831685 cubic meters or 2831.685 liters. This allowed a more universal understanding of water volumes since the liter is a common unit of measurement for water consumption.

Matplotlib.pyplot made it possible to visualize the data and better understand consumption trends before and after data processing. NumPy was essential for scientific

computing, supporting multidimensional arrays and matrices and providing a vast collection of mathematical functions to operate on these arrays.

In **step 3**, there is Anomaly detection with Isolation Forest to identify potential anomalies in water consumption patterns. It isolated specific points by randomly selecting a feature and then selecting a division value between the maximum and minimum values of the selected feature. Anomalies are those instances that have short paths in the isolation trees.

With the Sklearn.ensemble module, Isolation Forest, was used to build a model that efficiently identifies anomalies in water consumption. This is a module of the Scikit-learn library that provides machine learning algorithms for ensembling and bagging.

Each of these tools played a vital role in the processing and analysis of the data. Early development using Pig and Hadoop provided the ability to process and filter large volumes of data in a distributed and efficient manner. With its robust libraries, Python allowed for data refinement to ensure that only the most relevant information was used. Finally, Isolation Forest helped to isolate and identify anomalies, providing valuable insights that could be visualized and interpreted through the generated graphs. This integrated workflow was critical to the success of the water consumption anomaly detection project.

3.2. Partial Results

After applying the filtering process, distinct water consumption patterns by inhabited locations emerged. Visualizing these patterns is essential to interpret trends and identify atypical behaviors.

A series of graphs illustrates water consumption in the 20 selected locations. Each graph provides valuable insights into the consumption profile and serves as a tool to guide potential water management decisions and sustainability policies.

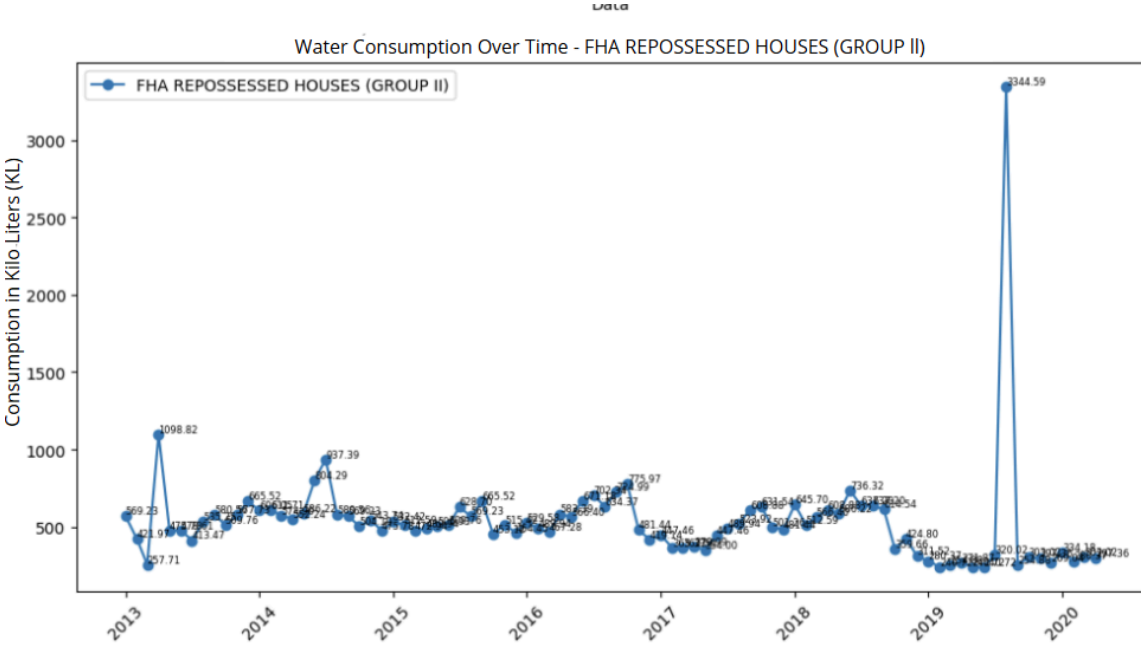

Figure 3. Consumption information processed for one of the locations entitled FHA REPOSSESSED HOUSES (GROUP II), New York - Consumption in KiloLiters on the Y axis and Date on the X axis.

Figure 3 visually represents water consumption for Group II of FHA Repossessed Houses over the years, highlighting the general trend and occasional usage spikes. Stability with moderate variations is noted, except for a significant spike at the end of the observed period, mid-2019, which requires special attention to determine the cause of such an increase in consumption.

Table 3. Results of Consumer Statistics processed for the locality of FHA REPOSSESSED HOUSES (GROUP II), New York

count	88
mean	534,73 KL
std (Standard Deviation)	341,68 KL
min (Minimum Value)	240.72 KL
25% (First Quartile)	402,85 KL
50% (Median)	509,76 KL
75% (Third Quartile)	591,18 KL
max (Highest Value)	3344.60 KL

Table 3 complements the visual analysis with descriptive statistics, providing a quantitative view of the 88 water consumption records. The average consumption is 534.73 KL, with a standard deviation of 341.68 KL, indicating high variability.

Consumption ranged from a minimum of 240.72 KL to a surprising maximum of 3344.60 KL, highlighting the need to investigate extremes to ensure efficiency and identify possible leaks or inefficiencies in the system.

3.3. Resulting Graph With Application Of Isolation Forest

This section presents the relevant results of applying the Isolation Forest method to identify anomalies in water consumption patterns in some locations studied in New York. This machine learning method isolated observations that deviated from the normal consumption pattern, providing valuable insights for understanding and managing water resources.

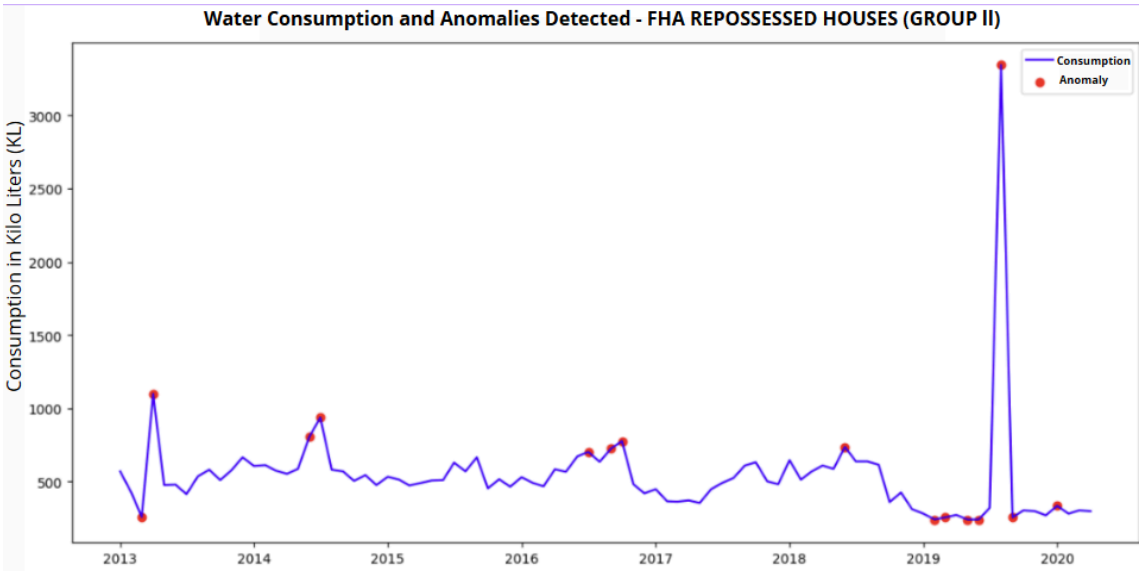

Figure 4. Consumption Information with the application of Isolation Forest for the location of FHA REPOSSESSED HOUSES (GROUP II), New York - Consumption in KiloLiters on the Y axis and Date on the X axis.

In Figure 4, the extraordinary consumption that deviates from the ordinary is noticeable. There are minor anomalies throughout the records until 2019. However, a surprising record is made, exceeding 3 million liters of consumption, while the average would be 534,730 liters, increasing by 6 times the standard value for the location of FHA REPOSSESSED HOUSES (GROUP II).

4. Conclusion and Future Works

Based on the results obtained and the analyses performed throughout this study, it is possible to conclude that the work achieved its objective of developing a robust system for detecting anomalies in water consumption. Theories and articles that adopted the idea of the FOG-SWM architecture and the implementation of the Hadoop and

MapReduce tools were reviewed. These proved effective in identifying atypical water consumption patterns, offering a new approach to monitoring and managing water resources.

With a preview of the functionality in mind, it was possible to delve deeper into the tools for collecting and detecting anomalies in water consumption. The FOG-Computing architecture was used as a fictitious model in data collection, while Map-Reduce was employed in preprocessing, mapping, and reducing the collected information to the desired model.

Executing the algorithm within the Hadoop environment ensured that large volumes of data were stored and processed efficiently, a crucial aspect for large-scale water resource management entities. The collection, storage, processing, and analysis of data were interdependent elements that allowed for more efficient development of such data.

After collecting and applying Map-Reduce, the filtering and cleaning of the data, carried out by algorithms developed in the Python language, were possible thanks to its vast libraries. Then, applying the Isolation Forest algorithm within this framework proved to be an assertive choice, demonstrating its ability to effectively and accurately discern between what would be considered regular consumption and anomalous consumption.

Thus, satisfaction with the results is significant, especially when considering the potential impact of early detection of irregularities on water conservation and cost savings. Specific challenges were encountered during the development period, such as the need for a reliable repository with valid data, which meant that the data quality was generous, and the detection of anomalies was fundamentally based on faithful and consistent data.

This system contributes to the academic field as a case study in the application of emerging technologies and presents valuable practical implications for society. In the technological context, this work contributes to the evolution of distributed data processing methodologies, showing the importance of integrating different tools and techniques to achieve accurate and reliable results.

In addition, there is room for continuous optimization of the project. Future research could further explore incorporating more advanced machine learning and artificial intelligence techniques to improve detection accuracy and speed. Adapting the system to be more interactive and user-friendly could also be a focus, allowing water resource managers to not only receive anomalous consumption alerts but also gain insights on how to optimize water use.

Looking to the future, we see a vast field of possibilities for the expansion and application of this work:

• External integrations;

• More robust AI applications;

• Graphical interface;

Integrating the system (1) developed in entities responsible for managing water consumption in homes, businesses, and industries could improve water use efficiency and help prevent significant losses due to leaks and other forms of waste.

With the continuous evolution of technologies, the application of more robust A**rtificial Intelligence (2)** trained to deal with large amounts of data focused on detecting anomalies in water consumption would be a viable improvement to the current project since there is a clear model, concept, and proposal for future projects.

Finally, future versions of the system can consider developing a **graphical user interface (3)** and implementing a real-time feedback system. Such improvements would facilitate the system's adoption by a wider range of users, making water resource management more accessible and effective.

Acknowledgment

The authors sincerely thank the Federal University of Santa Catarina (UFSC). This study was partially funded by the Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC), Edital 20/2024.

References

Braga, B. Desafios Futuros da Gestão da Água no Brasil. 2023. Disponível em: <https://aguasdobrasil.org/wp-content/uploads/2023/04/Revista-Aguas-do-Brasil-29-web2.pdf> Acesso em: 04 Maio. 2023

Brandão, Igor Antônio de Paiva et al. Dispositivo IoT de micromedição integrado com cloud computing para monitoramento do consumo de água. 2022.

Da Paz, Lívia Menezes; Esquerre, Karla; Sartori, Isabel. DETECÇÃO DE ANOMALIAS ATRAVÉS DO MONITORAMENTO DO CONSUMO DE ÁGUA. Revista do Seminário Internacional de Estatística com R, v. 3, n. 2, 2018.

Diaconita, Vlad; Bologa, Ana-Ramona; Bologa, Razvan. Hadoop oriented smart cities architecture. Sensors, v. 18, n. 4, p. 1181, 2018.

Diagnóstico dos Serviços de Água e Esgoto. Secretaria Nacional de Saneamento, 2019 Disponível em: <http://antigo.snis.gov.br/downloads/diagnosticos/ae/2019/Diagnostico-SNIS-AE-2019-Capitulo-16.pdf>. Acesso em 13 mar. 2023

Jach, Tomasz; Magiera, Ewa; Froelich, Wojciech. Application of HADOOP to store and process big data gathered from an urban water distribution system. Procedia Engineering, v. 119, p. 1375-1380, 2015.

Leguizamón Rojas, Gloria Ayde et al. Análisis de datos y modelos de aprendizaje para monitorizar el consumo de agua en redes de abastecimiento usando tecnologías Big Data. 2018.

Marçal, Pedro Henrique Silva; Brito, Vinicius Gabriel Pereira; Diniz, Débora Pelicano. HIDRÔMETRO INTELIGENTE: auxiliar na visualização do consumo e desperdício de água e seu impacto financeiro. Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica, v. 13, n. 1, 2023.

Niches, David et al. Proposta de estrutura de monitoramento de vulnerabilidades para sistemas de abastecimento de água. 2023. nov. 2023.

NYCHA. NYCHA Partners with HPD, Restored Homes HDFC, and Neighborhood Housing Services of Queens to Provide Homeownership Opportunities to Public Housing Residents through the Small Homes Rehab-NYCHA Program, Cluster III. O que é e sobre o FHA. Disponível em:<https://www.nyc.gov/site/nycha/about/press/pr-2022/pr-20221116.page#> Acesso em: 10 jan. 2024

Pinto, Renata Costa. Real-time business intelligence para um uso mais eficiente da água em ambientes urbanos. 2020. Tese de Doutorado.

Silva, Thiago Pereira et al. Plataformas de Fog Computing: da Teoria à Prática. Sociedade Brasileira de Computação, 2020.

Vasconcelos, Bruno Kevin. Smart campus: um estudo exploratório na UFPB—Campus V. Trabalho de conclusão de curso, 2023