IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 27, No. 6, December, 2022, pp. 949-956
Journal homepage: http://iieta.org/journals/isi

Big Data Enabling Fish Farming Data-Driven Strategy

Mohamed El Mehdi El Aissi*, Sarah Benjelloun, Younes Lakhrissi, Safae El Haj Ben Ali

Faculty of Science and Technology, SIGER Laboratory, Sidi Mohamed Ben Abdellah University, Fes 36000, Morocco

Corresponding Author Email: mohamedelmehdi.elaissi@usmba.ac.ma

https://doi.org/10.18280/isi.270611

Received: 29 August 2022
Accepted: 8 November 2022

Keywords:
big data, data lake, data driven strategy, digital transformation, fish farming, data value chain

ABSTRACT

The last two decades have witnessed an exponential data generation in tremendous amounts. The digital transformation in many domains leads to massive amounts of heterogeneous data. In order to benefit from this generation, value is extracted through data processing. Meanwhile, the fish farming functioning activity generates data with such volume, speed, heterogeneous sources and structures but it is not fully exploited. The traditional data warehouse solutions are not able to manage complex data with these characteristics. Thus, the concept of the data lake has emerged for more flexible and powerful data exploitation. Indeed, big data technologies and techniques are used to extract, process and analyze data. Since big data technologies have proved their benefits in other domains, it is irrefutable that using them in the fish farming domain will help it to reach its full potential. For this purpose, we propose in this paper a dedicated data lake architecture for handling fish farming data to initiate the adoption of a data driven strategy.

1. INTRODUCTION

Many countries started aquaculture production to keep up with the rapid world population growth. Indeed, food demand is increasing at a very fast pace [1]. Fish is preferred because it is a source of protein leaner and lower in calories and also cheaper. Fish farming is not only helping secure food safety as a basic need but is also offering several employment opportunities and is contributing on an economic scale. Asian countries hold the leading role in terms of fish production. China holds the first place followed by other Asian countries like Indonesia, India, Vietnam, the Philippines, Bangladesh, South Korea, Thailand, and Japan [2]. Unfortunately, Morocco is still in its embryonic phase in this domain despite the existence of large perspectives [3]. In order to accelerate the development of aquaculture, it is a necessity to adopt new technologies of digitalization.

The digital transformation changes the way an organization operates [4]. It has a profound effect not only on individual industries, but also on the perception of value as a whole. The transformation affects every level of the organization to bring together across areas and work effectively [5]. Systems, processes, workflows and culture are all part of the transformation.

In fact, data driven techniques have proved their efficiency in improving nearly all industries. According to the study [6], data technologies and business revenue worldwide reached 189.1 billion us dollars and is forecasted to reach 274.3 billion us dollars in 2022.

Big data analysis has been used in various industries such as banking, insurance, medicine, industry and marketing [7]. Despite all the success that big data has achieved in the mentioned fields, it started being applied to agriculture only recently [8], but not in the fish farming domain [9]. Therefore, using big data related technologies became an inevitability to tackle the challenges of productivity, environmental impact, food security and sustainability [10].

The scope of this article is to design a data management system capable of handling massive data generated by fish farming systems. The reason behind this architecture proposal is to initiate a data-driven strategy for the fish farming domain. Our inspiration comes from the fact that data driven strategies have become a key element for field development such as banking, insurance and healthcare. Furthermore, big data techniques are not widely treated yet in fish farming despite their benefits as they proved a significant potential in other domains [11]. It has to be noted that our proposed architecture is not restricted to managing data generated by only one farm, but also data gathered from multiple farms. This strategy has the objective of building a knowledge repository for fish farmers on one hand and researchers on the other hand. To clarify, the proposed architecture will allow performing advanced analytics on massive and rich data repositories through a dedicated platform, which is not possible with traditional data management systems. Therefore, extracting valuable information and benefiting from predictions and revealed hidden patterns.

This article is organized into seven major sections. After the introduction, the second section contains the overview of fish farming and its main challenges. The third section presents the data driven strategy as well as the existing data handling systems. The fourth section describes the different data sources in the fish farming system and the characteristics of big data. In the fifth section, we present the value chain of the data-driven fish farming strategy. In the sixth section, we propose a dedicated data lake end-to-end functional architecture. The seventh section contains a conclusion and future work.

2. OVERVIEW OF FISH FARMING DOMAIN AND ITS MAIN CHALLENGES

2.1 Fish farming overview

The global appetite for fish and fishery products shows no signs of slowing down. The fisheries and aquaculture have an important and growing role in food, nutrition and employment. According to the Food and Agriculture Organization of the United Nations (FAO), the global fishing and aquaculture industry has grown significantly over the past decades and total production, trade and consumption reached an all-time high in 2018 of 179 million ton with 58.8 million tons in China only. Aquaculture, inland waters and marine waters, have reached 115 million tons. In 2018, inland aquaculture produced 51.3 million tons of seafood, accounting for 62.5% of the total global production of farmed fish for human consumption. These are aquaculture produced from natural inland water sources, such as rivers, lakes, and fish farms [12].

The world production of aquaculture fish is mainly in Asia, with an 89% market share over the past two decades. Among the major producing countries, China, India, Indonesia, Vietnam, Bangladesh, Egypt, Norway, and Chile have consolidated their share in regional or global production to varying degrees in the past few years over the past two decades [12].

Since Morocco is the country of residency and provides practicalities of contacts, it is where our research is based. Moroccan continental aquaculture production levels are not very well recorded, but according to The World Bank [13], the Moroccan production level has increased from 1.403 tons in 2001 to around 2.250 tons in 2005. Then, the production level dropped to 579 tons in 2008 to then increase to 1.267 tons in 2018.

With the population growth, the demand for fishery products has significantly increased. It is the case for Morocco and according to [14] fish consumption per capita has reached 20 kilograms. This means a total of approximatively 693 million kilograms. As a result, in 2021, Moroccan fish imports' value reached 230 621 million American dollars, according to the International Trade Center [15].

2.2 Challenges

Many challenges are still faced in the fish farming domain. According to the FAO, it is necessary to transform and adapt agricultural systems to achieve food safety while facing the threat of climate change [16]. In Morocco, the nature of the climate obliges to take precautions in terms of water and land use, especially in semi-arid and arid areas. This leads to shed light on the use of these resources in the sectors that use them most. Moreover, according to [17], it is primordial to always keep a good water quality. The different parameters - Pressure, Temperature, PH, Dissolved Oxygen, Alkalinity, etc.- need to be monitored in order to keep mortality rate very low. Also, these parameters can be modified to impact the fish reproduction by simulating an adequate environment for reproduction. Furthermore, feed and feeding strategies are very important for a high feed conversion ratio (FCR) [18], slight changes can have impacts, so it is profitable to have this as automated actions, tightly controlled. Another challenge is that fish grow at a different rate. Once this happens, big fish feed on smaller ones. This results in a big loss of profits since fish feed is expensive. Knowing this, it is important to sort fish according to their sizes frequently. On top of that, fish farming should be based on predictive actions and not only corrective actions. That is why the use of Artificial Intelligence (AI) is primordial and it can be used for species classification, behavioral analysis, feeding decisions, size or biomass estimation, water quality prediction and disease identification [19]. These challenges can be resumed in the following:

- Control consumption and use of water
- Control water quality parameters at all times
- Automate and manage fish feeding
- Sort fish according to their size automatically
- Avoid manual actions to reduce variability and/or mistakes
- Predict and act accordingly to avoid loss

The need to adopt new technologies in fish production is more and more manifest [20]. Indeed, new technologies and especially data related technologies have earned one's spurs in other domains. Industry 4.0 relies heavily on the internet of things, artificial intelligence, cloud infrastructure and big data analytics, forming the big four technologies [21]. The aim is to overcome the challenges discussed above by:

- Improve accuracy and repeatability in fish farming operations
- Reduce resource wastage for sustainability
- Facilitate more autonomous and continuous fish monitoring
- Increase feed conversion ratio
- Reduce mortality and sickness rate
- Reduce dependencies on manual labor
- Provide more reliable decision support

Through these means, fish health and welfare will improve while increasing productivity and environmental sustainability [22].

Since the fish farming industry generates data continuously, big data technologies will have a huge impact on the way and result of production. A data driven strategy is making data in the center of interest and getting most out of it [23]. Indeed, data driven fish farming systems will not only overcome existing challenges, but will also help identify hidden patterns and correlations, develop actionable insights and predict future needs and trends.

3. DATA DRIVEN STRATEGY FOR FISH FARMING DOMAIN

3.1 Big data in data driven strategies

The concept of data driven strategies has been used in many domains such as marketing, banking, insurance, etc. and it offered an interesting increase of profitability and risks anticipation [24].

The global revenue from the big data market in 2020 is 56 billion U.S. dollars and is forecasted to reach 93 billion U.S dollars in 2027 [6]. In a survey conducted by Statista on over a thousand companies worldwide, 38% respondents adopt data-driven decision-making in 2018. In 2020, the respondents surveyed demonstrated a 12 percent increase in implementing data-driven decision making within their global organizations when compared to 2018 as shown in Figure 1 [25]. This increase will continue as data becomes the center of decision-making worldwide.

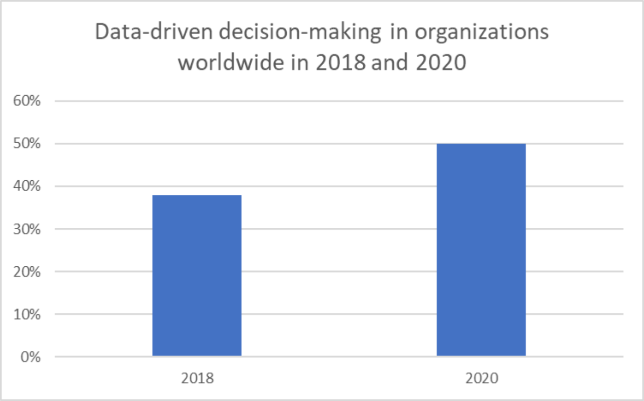

Figure 1. Data-driven decision-making in organizations worldwide in 2018 and 2020 [25]

Furthermore, the survey conducted by Statista in 2020 presented in Figure 2, shows the distribution of organizations implementing data-driven decision making. The banking sector led in terms of data-driven decision making within organizations, with 65 percent of respondents indicating as such. Other noteworthy sectors for data-driven decision making within organizations are insurance (55%) and telecom (54%). It has to be noted that the agricultural domain is not present in this ranking, thus it needs to invest more in data-driven decision-making.

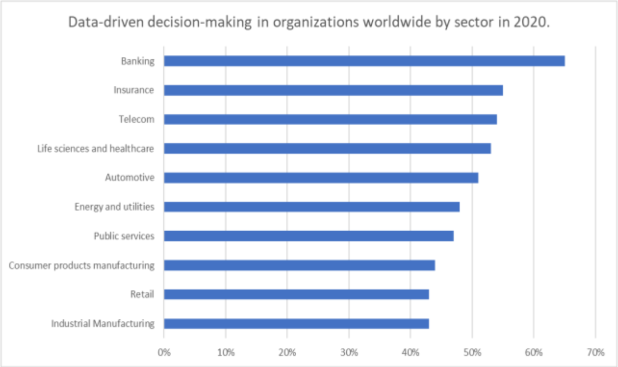

Figure 2. Data-driven decision-making in organizations worldwide by sector in 2020 [26]

Big data technologies are defined as a set of software designed to collect, store, process and analyze big datasets with complex structures that cannot be handled by traditional data processing technologies. There are four types of big data technologies for analytical big data: data storage, data mining, data analytics and data visualization. These four types are shown in Figure 3.

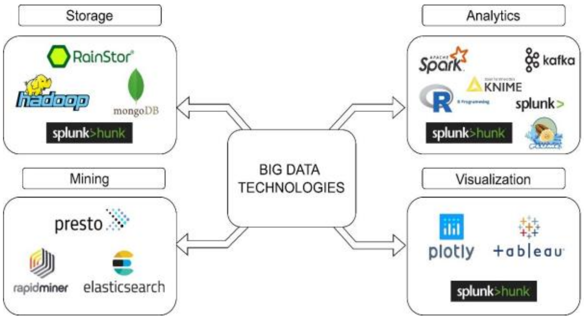

Figure 3. Big data technologies by type

3.2 Data management systems

Adopting a data-driven strategy is possible by considering data as the backbone of the domain by making it in the middle so that all the decisions are based on it. To do so, it is necessary to have an adequate data management system capable of handling tremendous amounts of generated data.

At this level, it is imperative to evaluate the different data management systems by shedding light on the key differences, advantages and inconveniences of each one of them, in order to choose the adequate data management architecture allowing an efficient data-driven fish farming strategy.

The data generated in the world has been increasing exponentially which led to the apparition of new data management systems, with the main goal of handling, exploring, and extracting valuable information and hidden patterns from the data [27]. Mainly, we can distinguish between two concepts of data management systems [28]. Nevertheless, there are many other architectures but they can be categorized as one of these two.

The first concept is the data warehouse, which is designed to allow querying and analyzing data. A data warehouse handles structured historical data gathered from many transactional databases, and the data handling is subject oriented [29]. Usually, when using a data warehouse architecture, data undergoes the process of data cleansing before storing it to guarantee a high data quality [30]. These key features of the data warehouse are usually referred to as the ETL process (extract, transform and load).

The ETL process involves three main steps. First, the data extraction step is where the data is gathered from homogeneous or heterogeneous sources. The second step is data transformation, where data is cleaned, formatted and transformed in order to respect a specific defined structure. The last step is the data loading, which refers to ingesting data into the model.

The second concept is the data lake which is considered as an efficient big data architecture allowing handling huge amounts of data with different types: Structured, semi-structured and unstructured data [31]. The data lake offers a high flexibility in terms of handling and interacting with data, since it does not require a predefined schema or a model to respect the gathered data [32]. In the same optic, the data lake relies on ELT (extract, load, transform) process rather than the ETL as presented in data warehouse architecture. This is because it gives more importance for extracting and loading data from different sources as much as possible. The last step is the data transformation where data consumers can build their own model based on the already existing data depending on the use case [18].

In Table 1, we present the key differences between the two architectures along with the key dimensions that our study is based on.

Table 1. Data warehouse architecture vs. data lake architecture

Dimension	Data Warehouse	Data Lake
Workload	Heavy	Medium
Schema Definition	Before Data Collection	After Data Collection
Data Access	Standard SQL	Custom Programs
Data Stored	Cleaned and formatted	Raw

The dimensions used to study the differences between the two architectures are chosen based on the properties of a data-based fish farming system. Workload for how heavy transformations is, schema definition for when data schema is defined, data access is for tools used to access data in the system and finally, data stored is for which form is data stored in (transformed or raw).

In sum, the data lake architecture offers multiple advantages for handling massive data. In term of scalability, the data lake it is relatively inexpensive compared to the scalability in traditional data warehouses. Furthermore, data lake provides native data storage format as it stores raw data directly coming from the source systems. Moreover, using a data lake solution implies a high schema flexibility, as the data lake decouples schema definition from data which is perfect for advanced analytics as we can create multiple schemas for the same dataset depending on the use case. Finally, the data stored in a data lake architecture could be processed using multiple languages which allows performing fast data analytics in both, batch and real-time mode.

4. POTENTIAL DATA SOURCES FOR FISH FARMING SYSTEM

A data source is the initial location where data is first created or where physical information is first digitized. However, even refined data can be used as a data source as long as another system or process accesses it and utilizes it. Concretely, a data source may be a database, flat files, measurements from physical devices, web data or any of the many static and streaming data services available on the internet. In general, these data sources fall under three types: machine data source, transactional data source and social data source [33].

In a fish farming production system, data is generated continuously by either transaction data sources or machine data sources.

·**Sensors:** They are implanted generally in tanks in order to get precise and valuable measures about the environmental data or the feeding intake. They can measure: temperature, dissolved oxygen, salinity, potential of hydrogen (pH), etc. The sensors generate data as streams of real-time data.

·**APIs:** They are sought to access data available on: raw material prices, market data, weather forecast, product market prices.

·**Flat files:** They contain additional or complementary data that cannot be automated or have not been automated yet. They can be in the format of separated values or size delimited values and are manually constructed. It can be data about feed types, raw material quantities, marketing strategy or sales data.

Data is a set of qualitative or quantitative variables, it can be structured or unstructured. There are dimensions or characteristics that distinguish data from big data. In 2001, the analytics firm Metagroup (now Gartner) introduced data scientists and analysts to the 3V's of three-dimensional (3D) data, which are Volume, Velocity and Variety [34]. Over a period of time, there was a rampant change in how data is captured and processed. Data was growing so rapidly in size that it came to be known as big data [35]. With the astronomical growth of data, two new V's - value and veracity- have been added by Gartner to the data processing concepts.

·Volume defines the size of data that is produced. It is the V most associated with big data. Volumes of data can reach unprecedented heights. As a result, it is common for large organizations to have Terabytes and even Petabytes of data in storage devices and servers. The more data available, the more insights can be retrieved and patterns discovered.

·Velocity refers to the speed of which data is generated and how quickly that data moves. This factor is important for organizations in order to have data available at the right time to make the best business decisions possible.

·Variety in big data entails processing diverse data types collected to varied data sources [36]. Generally, data is classified as structured, semi-structured and unstructured data. Structured data had defined format, length and size. Semi-structured data is one that may partially conform to a specific data format. Unstructured data is unorganized and doesn't conform to traditional data format. To illustrate, Comma Separated Values (CSV) files or data from relational databases are structured data; JavaScript Object Notation (JSON), extensible markup language (xml) or other markup languages are semi-structured; images, videos and social data fall under the uncategorized data type.

·Value is how worthy the data is of positively impacting the organization. Because collecting data that is produced in large volumes is of no use and the storing and aggregating of data is not equal to value addition. Instead, the insights extracted from the data is what matters [37]. With the help of advanced data analytics, useful insights can be derived from collected data. These insights, in return, help in the decision-making process. To ensure the value of big data is worth investing time and effort into, a cost VS benefit analysis can be conducted, organizations can then decide whether or not big data analytics will actually add any value.

·Veracity or validity is the assurance of quality and credibility of the collected data. It has to be credible enough to trust the insights garnered from this data in order to base the decision making on them. Dirty data, also known as rogue data, are inaccurate, incomplete or inconsistent data that contains erroneous information. Cleaning dirty data helps enhance its veracity and prevent misinformed and misleading decision-making. The common types of dirty data are: Duplicate data, outdated data, non-compliant data, incomplete data, inaccurate data. To prevent dirty data, data health assessments can be held with the data provider, mixing data sources (first party, third party and intent data), cleansing data regularly and filling gaps and ongoing data management.

In fish farming, data is certainly big data since it respects all 5V's that characterize big data. To demonstrate, data captured from sensors has volume and velocity since the IOT devices generate data constantly and continuously across the tanks of fish farms (Ex.: Temperature, Ph, Oxygen…) [38]. Moreover, data is gathered from multiple data sources, data from sensors are semi-structured while APIs and manual data are structured; it represents the variety of big data (Ex.: weather data from APIs, sales data from flat files…). There are insights and patterns that can be discovered from fish farming data so it has Value for the domain (Ex.: Calculating the turbidity rate in the fish tanks, feed conversion rate, sales, and market KPIs). As per Veracity, manual data and sensor data can be cleaned and managed through the data values chain while APIs data is clean data. As a consequence, the most suitable architecture to manage fish farming data is the data lake architecture.

5. FISH FARMING DATA VALUE CHAIN

Adopting a fish farming data-driven strategy becomes possible after identifying all the elements that make data the backbone of this approach.

The purpose is to conceive an architecture based on big data in order to constitute enough structured and unstructured fish farming data to be able to exploit it using analytics and AI. However, only adopting a big data architecture without a large perspective is not sufficient to exploit the data to its true potential. Thus, defining a fish farming big data value chain becomes a necessity.

The fish farming big data value chain sheds light on the different steps of the data life cycle. After listing all the elements necessary in the process of extracting value from data, we constructed a fish farming data value chain.

The Figure 4 illustrates an end-to-end fish farming data flow from the moment it is captured to data exploration and exploitation.

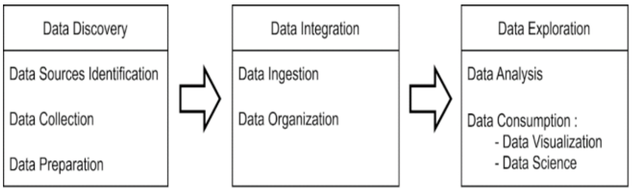

Figure 4. Data value chain for fish farming data

5.1 Data discovery

Before starting the process of data discovery, it is mandatory to state clearly the objectives in order to understand the need and get suitable data that will be used for extracting valuable information on behalf of business people. The process contains three main steps, the first step is identifying the potential different data sources with the main objective of creating an inventory of metadata that describes the data as-is in the source, then comes the data collection step where determining the data acquisition approaches remains the most important task, finally, the in the data preparation step, we prepare all the prerequisites for data access, like enabling the access to the data sources and setting up access-control rules [39].

5.2 Data integration

At this phase, we first need to organize the data by identifying its type (structured, semi-structured, unstructured) which allows a smooth data ingestion process. Depending on the data type, we can define the adequate parameters and specifications for making data consumable. Even if the organization step is usually at the level of data discovery, we consider that it should be included in the data integration process, since it allows data modelers, mainly, to have a wider view about the available data.

The following step is the data ingestion, where the data is being moved from the staging area to the data lake storage, where it can be further processed and analyzed. Depending on the data velocity, the data ingestion step can be performed in two different modes, batch or real time.

The batch mode is usually adopted when we need data to be imported at scheduled intervals. This can be useful when the processes run on a schedule (daily, weekly, etc.) such as daily reports. The real-time ingestion is used when data is time sensitive and needs to be imported as soon as it is generated in order to keep the very last image of data; this can be useful in monitoring use cases [40].

5.3 Data exploration

The final phase of the fish farming big data value chain is the data exploration, where data is used to unveil insights for decision making. The first step is data analysis where data is cleaned, transformed and modeled in order to discover decision-making insights [41]. The two primary methods for data analysis are qualitative data analysis techniques and quantitative data analysis techniques. These techniques can be used separately or in combination depending on the need.

The second and last step consists of using the generated insights from the analysis. At this level, we can distinguish between two levels of data consumption: data visualization and data science.

At the data visualization level, the objective is to build a comprehensive layout for decision-making containing mainly the key performance indicators (KPIs) extracted in the data analysis step.

Whereas at the data science level, the objective is to discover trends and patterns that cannot be seen using only data visualization techniques. With the help of artificial intelligence techniques, the decision-making process is elevated and supported by predictions [42].

6. MANAGING FISH FARMING BIG DATA USING DATA LAKE: END-TO-END FUNCTIONAL ARCHITECTURE

A dedicated data-based architecture in the use case of fish farming shows the overall design of the system and the logical interrelationships between its components. It also will be the ground for future works to propose the technical architecture detailing all the technologies and software for a successful fish farming data-driven strategy. After that, it is possible to produce a proof of concept for a fish farming data management system.

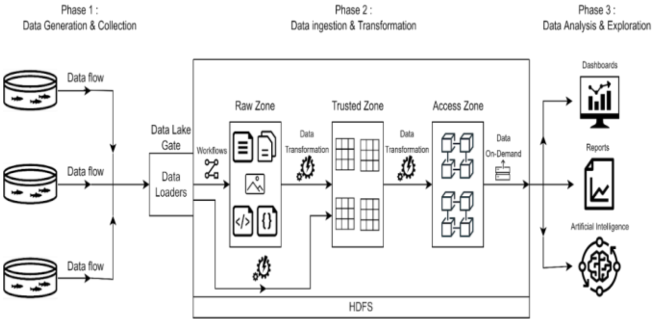

Figure 5. Data lake architecture for managing fish farming data

The data life cycle is a specific sequence of phases that each single unit of data goes through from its generation until its consumption [43]. Adopting a smart fish farming data-driven strategy leads to following a specific data life cycle where data is never lost or deleted, unlike other life cycles where the last phase is usually deletion of data [44]. Therefore, the previous data value chain (Figure 4) helps in the process of proposing a dedicated data lake architecture for managing fish farming data. The figure below (Figure 5) presents the architecture proposed and shows the path that data goes through starting from its generation in farms, passing through the data lake and arriving at the phase where it's consumed.

6.1 Data generation and collection

Fish farming data is generated from multiple farms. This data falls under different types (structured, semi, unstructured) depending on what it represents and its source. Each fish farm generates data mainly in two modes: batch and streaming.

Batch data represents the data produced with a certain frequency, since it needs time to be gathered or simply when the source cannot deliver a continuous stream, usually it regards data with large volume. On the other hand, stream data is a continuous data flow characterized by its velocity rather than volume [45, 46]. When generated, both batch and streaming data are sent to two different staging areas which are considered as the entry points of the data lake.

The batch staging area is a file system where different files are stored temporarily and where each data file respects a unified nomenclature so it can be easily identified in the following phase.

The streaming staging area consists of a message managing system where a single unit of data is represented as a message and messages are organized under different topics [47].

6.2 Data ingestion and transformation

After the data is staged, comes the role of the data loaders. A data loader is a program responsible for copying data into the Hadoop Distributed File System (HDFS) and then archiving it. This program is able to distinguish between the different data types and data sources and, depending on each data file's characteristics, stores it at the level of the raw zone.

The raw zone, as its name implies, contains data as-is in its raw form. It means data without any curation or structure. Regardless of what will be used, all data is ingested. The benefit of the raw zone is allowing a quick data ingestion process.

Depending on the data type, each file is stored in the appropriate HDFS repository. Structured data is stored with every field as a string in order to avoid loss of pieces of data which means data is not suitable for consumption yet. For unstructured and semi-structured data, there are dedicated repositories allowing the organization of data files by type, source and domain.

The trusted zone contains conformed data. After being stored at the level of the raw zone, data undergoes some cleaning processes and transformations which make it ready to use. The cleaning regards all the operations that allow building the golden data. When talking about a data lake structure, we are directly talking about the ELT process, where we give more importance to the loading phase. As a consequence, we usually end up with a big challenge of data quality. To tackle this challenge, applying some cleaning is required since it allows avoiding keeping data with low quality at the level of the trusted zone. However, it has to be noted that the cleaned data is not removed definitively from the data lake since a copy is still available in the raw zone.

Obtaining clean data is not enough to consider it eligible for the trusted zone. Along with this, applying some transformations to data is necessary in order to make it trustful. The type of transformations applied will not affect directly the data value, but defines the adequate structure and types of each data unit in order to ensure efficient use in different use cases.

In the access zone, data is preprocessed. For each use case specification, the required data is exposed in this zone, then the accesses are granted to different users since the available data in the fish farming data lake is not limited only for data specialists but also as data in demand for other users.

6.3 Data analysis and exploration

This phase consists of data exploration. One of the main reasons behind adopting the data lake architecture is allowing a flexible connection to different external platforms such as dashboarding and reporting tools. In addition, data scientists, mainly, can connect to the access zone to perform machine learning, deep learning and other artificial intelligence manipulations. Thus, the objective of the access zone is to expose preprocessed data and make it available for future analysis. With this architecture, we ensure that the different data consumers are able to access all the data available in the lake and combine it in order to extract valuable information without going through the complexity of gathering and preprocessing data. However, the presence of multiple parties manipulating the data at the level of the access zone becomes a real challenge since data can be altered and thus affecting another party that uses the same set of data. To tackle this challenge, multiple abstract layers are created with the required data for each party then, accessing these layers is granted and controlled by the data lake administrator.

7. CONCLUSIONS

After adopting big data technologies on multiple domains and acknowledging its benefits, it's undeniable that using these technologies in the fish farming domain will have just as big of an impact. Therefore, we propose a dedicated data lake architecture for handling fish farming data to initiate the adoption of a data-driven strategy in order to avail from the big data technologies advantages.

In this paper, we present the fish farming state of art, where we reveal the state of fish production worldwide and the gap between production and consumption in Morocco as well as the challenges faced in this domain. Then, after asserting the need of adopting a data-driven strategy, we explain the different data handling systems -data warehouses and data lakes- by focusing on the characteristics of each system. Moreover, we expose the different fish farming data sources and how it is classified as big data. Next, we present the fish farming data value chain that contains three major phases: Data Discovery, Data Integration and Data Exploration. After that, we propose an end-to-end data lake functional architecture which promotes the effective consumption of all the data gathered from the different sources, by providing layers of specific data for data consumers depending on the need.

Adopting this proposed data lake architecture for the fish farming data-driven strategy will not only optimize the production of a fish farm but it will greatly impact the evolution of the fish farming domain in a country or region. It has to be noted that this strategy is flexible to be applied in other agriculture domains without compromising its impact.

However, applying this strategy cannot be possible without proposing a specific technical architecture that defines the interactions between the different data lake architecture components and the used technologies. In this perspective, our future work will focus on proposing the adequate set of tools that should be implemented across with the previously proposed data-driven strategy. Also, providing a proof of concept that simulates the data flow starting from the data sources to data consumption using AI algorithms.

REFERENCES

[1] Sarker, M.N.I., Wu, M., Chanthamith, B., Yusufzada, S., Li, D., Zhang, J. (2019). Big data driven smart agriculture: Pathway for sustainable development. In 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 60-65. https://doi.org/10.1109/ICAIBD.2019.8836982

[2] The World Bank. (2016). Aquaculture Production (Metric Tons)|Data. www.Data.worldbank.org/indicator/ER.FSH.AQUA.MT?start=2012, accessed on Mar. 26, 2021.

[3] Holth, M., Van der Meer, A. (2018). Aquaculture business opportunities in Morocco for Dutch entrepreneurs. https://www.rvo.nl/sites/default/files/2018/06/Aquaculture-Business-Opportunities-Morocco.pdf, accessed on February 27, 2022.

[4] Lezoche, M., Hernandez, J.E., Díaz, M.D.M.E.A., Panetto, H., Kacprzyk, J. (2020). Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture. Computers in Industry, 117: 103187. http://dx.doi.org/10.1016/j.compind.2020.103187

[5] Sonka, S. (2016). Big data: Fueling the next evolution of agricultural innovation. Journal of Innovation Management, 4(1): 114-136. http://dx.doi.org/10.24840/2183-0606_004.001_0008

[6] Liu, S. (2020). Big data-statistics & facts. statista. https://www.statista.com/topics/1464/big-Data/, accessed on Mar. 20, 2022.

[7] Rodríguez-Mazahua, L., Rodríguez-Enríquez, C.A., Sánchez-Cervantes, J.L., Cervantes, J., García-Alcaraz, J.L., Alor-Hernández, G. (2016). A general perspective of big data: Applications, tools, challenges and trends. The Journal of Supercomputing, 72(8): 3073-3113. http://dx.doi.org/10.1007/s11227-015-1501-1

[8] Saiz-Rubio, V., Rovira-Más, F. (2020). From smart farming towards agriculture 5.0: A review on crop data management. Agronomy, 10(2): 207. http://dx.doi.org/10.3390/agronomy10020207

[9] Sarker, M.N.I., Islam, M.S., Murmu, H., Rozario, E. (2020). Role of big data on digital farming. Int J. Sci. Technol Res., 9(4): 1222-1225.

[10] Astill, J., Dara, R.A., Fraser, E.D., Roberts, B., Sharif, S. (2020). Smart poultry management: Smart sensors, big data, and the internet of things. Computers and Electronics in Agriculture, 170: 105291. http://dx.doi.org/10.1016/j.compag.2020.105291

[11] Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F.X. (2017). A review on the practice of big data analysis in agriculture. Computers and Electronics in Agriculture, 143: 23-37. http://dx.doi.org/10.1016/j.compag.2017.09.037

[12] FAO. (2020). The State of World Fisheries and Aquaculture (SOFIA). https://www.fao.org/state-of-fisheries-aquaculture.

[13] The World Bank. (n.d.). Aquaculture production (metric tons) - Morocco|Data. https://data.worldbank.org/indicator/ER.FSH.AQUA.MT?locations=MA, accessed August 11, 2022.

[14] Our World in Data. (2018). Fish and seafood consumption per capita, 1991 to 2017. https://ourworldinData.org/grapher/fish-and-seafood-consumption-per-capita?tab=chart&time=1991..latest&country=~MAR, accessed on February 27, 2022.

[15] Trade Map. (2020). Fishery products imported by Morocco. www.trademap.org/Product_SelCountry_TS.aspx, accessed on October 19, 2022.

[16] FAO. (2016). The state of food and agriculture: Climate change, agriculture and food security. Rome, Italy: Author. ISBN 978-92-5-109374-0.

[17] Devi, P.A., Padmavathy, P., Aanand, S., Aruljothi, K. (2017). Review on water quality parameters in freshwater cage fish culture. International Journal of Applied Research, 3(5): 114-120.

[18] Mouzakitis, S., Tsapelas, G., Pelekis, S., Ntanopoulos, S., Askounis, D., Osinga, S., Athanasiadis, I.N. (2020, February). Investigation of common big data analytics and decision-making requirements across diverse precision agriculture and livestock farming use cases. In International Symposium on Environmental Software Systems, pp. 139-150. http://dx.doi.org/10.1007/978-3-030-39815-6_14

[19] Yang, X., Zhang, S., Liu, J., Gao, Q., Dong, S., Zhou, C. (2021). Deep learning for smart fish farming: applications, opportunities and challenges. Reviews in Aquaculture, 13(1): 66-90. https://doi.org/10.1111/raq.12464

[20] Jayashankar, P., Johnston, W.J., Nilakanta, S., Burres, R. (2019). Co-creation of value-in-use through big data technology-a B2B agricultural perspective. Journal of business & industrial marketing. http://dx.doi.org/10.1108/JBIM-12-2018-0411

[21] Lokers, R., Knapen, R., Janssen, S., van Randen, Y., Jansen, J. (2016). Analysis of big data technologies for use in agro-environmental science. Environmental Modelling & Software, 84: 494-504. http://dx.doi.org/10.1016/j.envsoft.2016.07.017

[22] Kamble, S.S., Gunasekaran, A., Gawankar, S.A. (2020). Achieving sustainable performance in a data-driven agriculture supply chain: A review for research and applications. International Journal of Production Economics, 219: 179-194. http://dx.doi.org/10.1016/j.ijpe.2019.05.022

[23] Kuo, Y.H., Kusiak, A. (2019). From data to big data in production research: The past and future trends. International Journal of Production Research, 57(15-16): 4828-4853. http://dx.doi.org/10.1080/00207543.2018.1443230

[24] Coleman, S., Göb, R., Manco, G., Pievatolo, A., Tort‐Martorell, X., Reis, M.S. (2016). How can SMEs benefit from big data? Challenges and a path forward. Quality and Reliability Engineering International, 32(6): 2151-2164. https://doi.org/10.1002/qre.2008

[25] Statista. (2022). Data-driven decision-making in organizations worldwide in 2018 and 2020. https://www.statista.com/statistics/1235409/worldwide-data-driven-decision-making-companies/, accessed on

August 25, 2022.

[26] Statista. (2022). Data-driven decision-making in organizations worldwide as of 2020, by sector. https://www.statista.com/statistics/1235436/worldwide-data-driven-decision-making-organizations-by-sector/, accessed on August 25, 2022.

[27] Klerkx, L., Jakku, E., Labarthe, P. (2019). A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda. NJAS-Wageningen Journal of Life Sciences, 90: 100315. http://dx.doi.org/10.1016/j.njas.2019.100315

[28] Aissi, E., El Mehdi, M., Benjelloun, S., Loukili, Y., Lakhrissi, Y., Boushaki, A.E., Elhaj Ben Ali, S. (2022). Data lake versus data warehouse architecture: A comparative study. In WITS 2020, pp. 201-210. http://dx.doi.org/10.1007/978-981-33-6893-4_19

[29] Inmon, W.H. (1996). The data warehouse and data mining. Communications of the ACM, 39(11): 49-51. http://dx.doi.org/10.1145/240455.240470

[30] Naeem, M., Jamal, T., Diaz-Martinez, J., Butt, S.A., Montesano, N., Tariq, M.I., De-La-Hoz-Valdiris, E. (2022). Trends and future perspective challenges in big data. In Advances in intelligent data analysis and applications, pp. 309-325. http://dx.doi.org/10.1007/978-981-16-5036-9_30

[31] Nargesian, F., Zhu, E., Miller, R.J., Pu, K.Q., Arocena, P.C. (2019). Data lake management: challenges and opportunities. Proceedings of the VLDB Endowment, 12(12): 1986-1989. http://dx.doi.org/10.14778/3352063.3352116

[32] Naqvi, R., Soomro, T.R., Alzoubi, H.M., Ghazal, T.M., Alshurideh, M. (2021). The nexus between big data and decision-making: A study of big data techniques and technologies. In The International Conference on Artificial Intelligence and Computer Vision, pp. 838-853. http://dx.doi.org/10.1007/978-3-030-76346-6_73

[33] Dong, X.L., Srivastava, D. (2013). Big data integration. In 2013 IEEE 29th International Conference on Data Engineering (ICDE), pp. 1245-1248. https://doi.org/10.1109/ICDE.2013.6544914

[34] Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META group research note, 6(70): 1. https://www.gartner.com/en/blog, accessed on Sept. 29, 2022.

[35] Oussous, A., Benjelloun, F.Z., Lahcen, A.A., Belfkih, S. (2018). Big Data technologies: A survey. Journal of King Saud University-Computer and Information Sciences, 30(4): 431-448. http://dx.doi.org/10.1016/j.jksuci.2017.06.001

[36] Jagadish, H.V. (2015). Big data and science: Myths and reality. Big Data Research, 2(2): 49-52. http://dx.doi.org/10.1016/j.bdr.2015.01.005

[37] Wysel, M., Baker, D., Billingsley, W. (2021). Data sharing platforms: How value is created from agricultural data. Agricultural Systems, 193: 103241. http://dx.doi.org/10.1016/j.agsy.2021.103241

[38] Yang, X.P., Cao, D.M., Chen, J., Xiao, Z.P., Daowd, A. (2020). AI and IoT-based collaborative business ecosystem: A case in Chinese fish farming industry. International Journal of Technology Management, 82(2): 151-171. http://dx.doi.org/10.1504/IJTM.2020.107856

[39] Ang, L.M., Seng, K.P. (2016). Big sensor data applications in urban environments. Big Data Research, 4: 1-12. https://doi.org/10.1016/j.bdr.2015.12.003

[40] Curry, E. (2016). The big data value chain: definitions, concepts, and theoretical approaches. In New horizons for a data-driven economy, pp. 29-37. https://doi.org/10.1007/978-3-319-21569-3_3

[41] Jony, R.I., Rony, R.I., Rahman, M., Rahat, A. (2016). Big data characteristics, value chain and challenges. In Proceedings of the 1st International Conference on Advanced Information and Communication Technology. Bangladesh.

[42] Cockburn, M. (2020). Application and prospective discussion of machine learning for the management of dairy farms. Animals, 10(9): 1690. https://doi.org/10.3390/ani10091690

[43] Rahul, K., Banyal, R.K. (2020). Data life cycle management in big data analytics. Procedia Computer Science, 173: 364-371. http://dx.doi.org/10.1016/j.procs.2020.06.042

[44] Cui, Y., Kara, S., Chan, K.C. (2020). Manufacturing big data ecosystem: A systematic literature review. Robotics and computer-integrated Manufacturing, 62: 101861. https://doi.org/10.1016/j.rcim.2019.101861

[45] Benjelloun, S., El Aissi, M.E.M., Loukili, Y., Lakhrissi, Y., Ali, S.E.B., Chougrad, H., El Boushaki, A. (2020). Big data processing: batch-based processing and stream-based processing. In 2020 Fourth International Conference on Intelligent Computing in Data Sciences (ICDS), pp. 1-6. https://doi.org/10.1109/ICDS50568.2020.9268684

[46] Subbiah, S.S., Chinnappan, J. (2021). Opportunities and challenges of feature selection methods for high dimensional data: A review. Ingénierie des Systèmes d'Information, 26(1): 67-77. https://doi.org/10.18280/isi.260107

[47] Chennouk, H., Ziyati, E.H., El Bhiri, B. (2022). Business value creation through project management based on big data approach. Ingénierie des Systèmes d'Information, 27(5): 823-828. https://doi.org/10.18280/isi.270516