A Zone-based Data Lake Architecture for Smart Crop Farming in Vietnam: A Strategic Perspective

Toan Pham Minh*, Huy Ha Quang, Tuan Nguyen Manh

UEH College of Technology and Design, UEH University, Ho Chi Minh City, Vietnam
toanphamminh03@gmail.com

Abstract. The rapid growth of smart crop farming technologies in Vietnam has resulted in an unforeseen surge in the volume, variety, and velocity of agricultural data. This paper suggests a zone-based data lake architecture to address the issues of managing, integrating, and analyzing the huge amount of heterogeneous data collected by several sources in smart crop farming. The proposed architecture utilizes big data technologies and its development follows a nine-step data lake architecture framework (DLAF). The paper covers the integration of batch and real-time processing technologies so as to develop predictive models, real-time monitoring, and data-driven decision support systems for smart crop farming. The paper illustrates the practical application of the proposed architecture in smart crop farming via many use cases, including crop yield estimation, disease detection, resource optimization, and climate risk assessment. It also covers the key aspects of metadata management, data security and data governance to meet the standards of data quality, lineage, and compliance.

Keywords: Big data, data lake, smart crop farming, zone.

1 Introduction

Agriculture remains a key driver of Vietnam's national stability and growth. Agriculture employs about one-third of the nation's workforce (World Bank, 2022). Farming was the major use of agricultural lands and it accounted for 39.4% of the total land area of Vietnam (World Bank, 2021). Vietnam ranks among the five countries most affected by climate change due to its remarkable coastline and the main settlement of its population and economic resources in coastal lowlands and deltas. Saltwater intrusion is anticipated to accelerate due to a projected 30 cm sea level rise by 2050 (Smajgl et al., 2015). An estimated 2 million hectares of agricultural land are at danger of saltwater intrusion each year (Loc et al., 2021). The ability to swiftly adapt and respond to environmental changes through intelligent data analysis is not merely beneficial but vital. Technologies that are employed to curb the use of resources, especially water, fertilizers, and pesticides for crop production will be the basis for improving agricultural sustainability (Misra et al., 2020).

Smart crop farming involves the use of the Internet of Things (IoT) sensors, drones, and machine learning algorithms. It relies heavily on large-scale, heterogeneous data collections from a range of sources, such as satellite images, soil sensors, and crop yields. However, quickly collecting and analyzing these big data sets from heter-

T. A. Trinh et al. (eds.), *Proceedings of the 2nd International Conference - Resilience by Technology and Design (RTD 2024)*, Advances in Intelligent Systems Research 186,
https://doi.org/10.2991/978-94-6463-583-6_4

ogeneous and large sources poses a big challenge especially in developing nations (Elijah et al., 2018). This necessitates a strong architecture capable of handling, processing, and extracting value from huge data pools - a role ideal for a data lake. The data lake is a repository where all types of data can be stored. Due to their flexibility and scalability, the data lake is well suited to the agricultural sectors of the emerging markets, enabling data-driven decisions to improve crop yields, resource use, and mitigate climate-related risks. This paper proposes a zone-based data lake architecture, with the goal of not only streamlining farm management, but also providing data integration, metadata management and advanced analytics that would be useful in data-driven decision-making in smart crop farming in Vietnam.

2 Literature Review

2.1 Data Lake Concept

Dixon (2010) firstly proposed the data lake concept to address the limitations of data marts, which are subsets of data warehouses that only answer certain issues. The data lake is a centralized storage repository that lets users store raw, heterogeneous data in its initial form to explore, retrieve and analyze data from external data sources. The data lake may be considered as a central repository where data of any kind is kept without a rigid schema for future investigations (Sawadogo & Darmont, 2021). Two fundamental features of the data lake serves as the foundation for this definition: The schema-on-read technique, which means that schema and data needs are not defined until the data is queried, and data variety (Sawadogo & Darmont, 2021). The data lake stores data in raw format using a flat architecture. A set of extended metadata tags and a unique identifier are assigned to each data entity in the lake (Miloslavskaya & Tolstoy, 2016).

Sawadogo et al. (2019) reviewed previous literature to define a complete definition of the data lake. The data lake is a scalable system for storing and analyzing any type of data in its initial form. It is mostly utilized by data professionals (statisticians, data scientists, or analysts) to extract knowledge. The characteristics of this system include a metadata catalog to ensure data quality, data governance policies and tools, accessibility to many types of users, integration of all data types, logical and physical organization, and scalability.

2.2 Data Lake Architecture Models

The initial flat data lake architecture made it possible to import heterogeneous data in its raw form at a cheap cost and closely tied to the Hadoop system. However, it prevents users from processing data and doesn't log any user operations (Ravat & Zhao, 2019). The data lake architecture predominantly employs two major approaches for managing pre-processed data: pond architecture and zone architecture (Giebler et al., 2019a).

The pond architecture by Inmon (2016) contains five data ponds. Each pond handles a specific type of data. Primarily, it restricts data availability to one specific pond

at a time, and as data transfers to other ponds, the raw data is lost. The process goes against the data lake's definition of ingesting all the raw data and processing them upon usage (Ravat & Zhao, 2019).

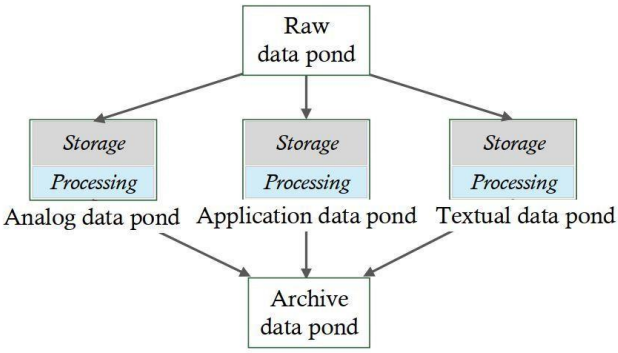

Fig. 1. Pond Architecture (Sawadogo & Darmont, 2021)

The basic concept behind dividing the data lake into various zones originates from the need to automate standardized pre-processing pipelines, organize the pre-processed data and deploy it for further processing (Wieder & Nolte, 2022). This is accomplished by allocating data to zones inside the data lake based on its processing level. The initial zone contains raw data that is ingested in its original format. There are several alternatives for zone architecture that are proposed and discussed in literature. An in-depth study of the zone architecture was carried out by Giebler et al. (2020), who examined the design differences, specific features, and use cases of five different data lakes based on the zone architecture (Gorelik, 2019; Madsen, 2015; Patel et al., 2017; Ravat & Zhao, 2019; Sharma, 2018; Zikopoulos, 2015), which led to the development of a generic meta-model for a zone and the specification of a zone reference model. This assessment shows that none of previous studies can satisfy all requirements. After designing the meta-model for zones and combining the widely accepted concepts from literature with the requirements from the latest assessment, Giebler et al. (2020) developed a zone architecture with six zones: Landing Zone, Raw Zone, Harmonized Zone, Distilled Zone, Explorative Zone, and Delivery Zone. In this model, Giebler et al. (2020) separate the zones into a raw zone and a harmonized zone, which are generic, and a distilled zone that is designed for individual use cases. The distilled zone serves data to the final delivery zone, enabling reporting and OLAP operations, while an explorative zone accommodates advanced analytics. Moreover, all zones have a protected area. This area is encrypted and safe. Data moves from one protected zone to another as shown in Fig. 2.

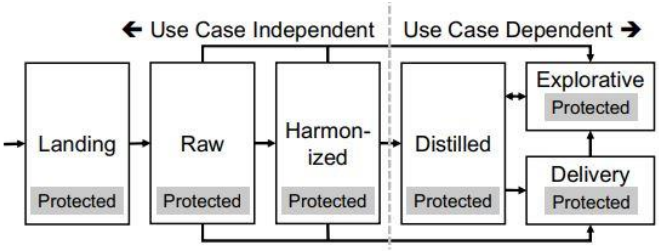

Fig. 2. Zone Architecture (Giebler et al., 2020)

2.3 Smart Crop Farming Concept

The smart farming concept formulated as the technology began to advance in the early 2010s (Rodríguez et al., 2019). Smart farming expanded upon from precision agriculture through the utilization of real-time data from smart sensors and IoT models to highlight a network for data exchange. Wolfert et al. (2017) define smart farming as a management strategy, which focuses mainly on the application of IoT, big data analytics, AI, and robotics to the most effective utilization of agricultural processes. A more comprehensive definition of smart farming was introduced as a data-driven approach that involved leveraging advanced technologies in order to maximize crop productivity, quality, and sustainability (Pivoto et al., 2018). Additionally, smart farming emphasizes the importance of extracting value and ensuring veracity from big data (Rodríguez et al., 2019).

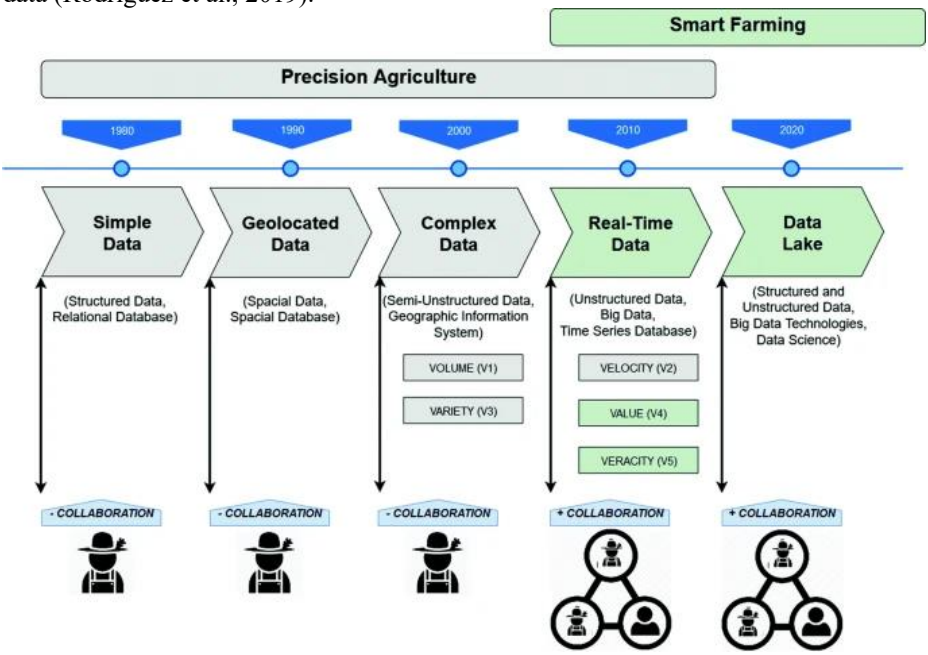

Fig. 3. Evolution of Smart Farming (Rodríguez et al., 2019)

Smart crop farming, as a subset of smart farming, is the use of smart technology and information systems to make the crop production process more efficient and environmentally friendly (Pivoto et al., 2018). Smart crop farming is primarily the application of technology to improve crop yield, quality, and resistance to external factors. This involves sending different sensors, such as soil moisture probes, and weather stations, to obtain real-time data on the growth, health and environmental conditions of crops. Through the utilization of these advanced algorithms and machine learning approaches, farmers will be provided with actionable insights that will help them optimize irrigation, fertilization, pest control and other critical factors of crop management. In addition, smart crop farming also considers the social, economic, and environmental factors that impact farming as well (Lioutas et al., 2019).

2.4 Overview of Crop Farming in Vietnam

There are various types of crop production in Vietnam, and each plays a crucial role in the agricultural industry. In 2020, Vietnam produced 42.7 million tons of rice, making the country one of the world's largest rice exporters (Ministry of Agriculture and Rural Development, 2021). Moreover, Vietnam is the second-largest coffee supplier in the world market (Nguyen et al., 2023). Additionally, Vietnam is known for its tropical fruit production, including mangos, dragon fruits, and lychees. Even though Vietnam is rated as a leading exporter of some types of agricultural crops, the country's production is distributed geographically. It means that each region has unique natural conditions, and their geography is distinctive with certain types of crops. The Mekong Delta is a region of rice growth, the Central Highlands specializes in coffee production, and fruit production is graphically diversified (Nguyen & Warr, 2020).

Vietnam's agricultural sector has experienced various technologies to improve productivity and efficiency. The use of machinery, chemical fertilizers, and pesticides has become more widespread, especially in large-scale farming operations. There is also a growing interest in smart farming technologies, such as remote sensing, GIS and GPS, which enable farmers to optimize resource use (Minh et al., 2019). The application of mobile apps and digital platforms has also facilitated access to market information and extension services for farmers (Hoang & Tran, 2023).

2.5 Related Work

Several studies have explored data lake architectures for various applications. Zhao et al. (2021) developed a zone-based data lake architecture for integrating IoT and big data. Their model has capabilities in managing heterogeneous data sources and implementing metadata management and data governance frameworks. However, it differentiates between batch and real-time data during the ingestion process, using internal datasets for batch operations and external ones for real-time data. Besides, Benjelloun et al. (2023) introduced a multi-zoned data lake architecture for Moroccan fish farming. Their model integrates data from diverse sources, including IoT sensors,

manual entries, and external APIs. The authors showcase the architecture's capabilities in monitoring water quality, feed management, and production forecasting.

The zone-based data lake architecture developed by Giebler et al. (2020) has been widely applied in various domains. For instance, Stach et al. (2022) utilized this model to manage large-scale biological datasets, particularly from protein sequence databases and mass spectrometry. This approach has streamlined the data handling process and made refined data readily available for analysis and interpretation. Parente (2021) also applied the Giebler's model to design a data lake for managing big data in healthcare. This approach addresses the challenges of integrating heterogeneous medical data, particularly medical images.

3 Methodology and Data

3.1 Data Lake Architecture Design

The data lake architecture framework (DLAF) was outlined by Giebler et al. (2021) as the comprehensive design of the data lake, including the infrastructure, data storage, data flow, data modeling, data organization, data processes, metadata management, data security and privacy, and data quality. This framework, by considering the interdependencies between aspects, ensures that the resulting architecture is cohesive, scalable and appropriate for data types, and the processing needs of the smart crop farming domain. There are nine steps that must be followed.

Step 1: Identify Scenario. The data collected is diverse in structure. The architecture must handle both real time stream processing for operational adjustments on the spot and batch processing approach for strategic planning. Data is utilized through a variety of advanced analytical techniques and operational processes that significantly enhance agricultural efficiency.

Step 2: Design Data Flow. The BRAID architecture, which enables both batch and stream processing, was chosen for this case. In the BRAID architecture, data ingested as a stream is immediately directed both to a stream processing engine and to persistent storage (Giebler et al., 2018). The BRAID architecture facilitates the utilization of batch processing results in stream processing activities as illustrated in Fig. 4.

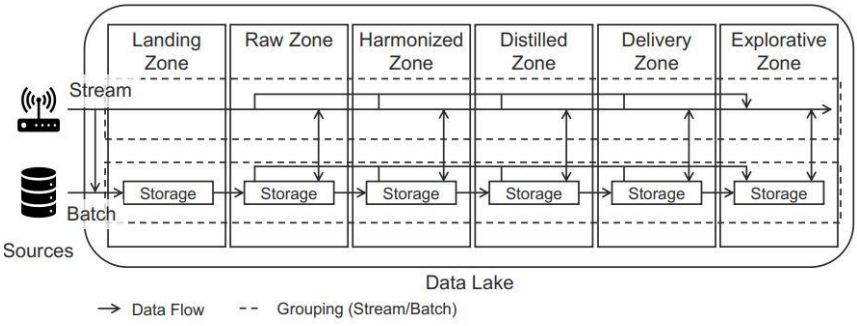

Fig. 4. Integration of Data Flow in the Zone Model (Giebler et al., 2021)

Step 3: Design Data Organization. As taking into account different types of data to get combined and raw data availability is essential, we opted to utilize zone architecture. We selected the zone architecture developed by Giebler et al. (2020), which is appropriate for our use case of handling both the data batch and real-time stream processing. Moreover, it considers data quality and data security concepts that maintain the privacy of sensitive agricultural data and access control. The zone architecture used and its data flow are shown in Fig. 4.

Step 4: Design Data Storage. The data storage approach makes use of multiple storage systems, each designed to meet certain data requirements to effectively handle various data types. It is also necessary to consider how the data will be used and its inherent characteristics.

Step 5: Design Infrastructure. As discussed in Step 4, it will be their task to determine and also to perform a complete range of big data technologies that have high proficiency in handling large-scale agricultural data as described in Section 3.3.

Step 6: Design Data Modeling. In the Landing Zone and the Raw Zone, data is kept in its raw form. The Harmonized Zone and Distilled Zone are both constructed using the Data Vault approach (Giebler et al., 2019b), which are exploited for the storing of structured data. While the Harmonized Zone employs Raw Vault modeling, the Distilled Zone uses Business Vault modeling in order for the incorporation of business logic and drawing of insights relevant to specific farming use cases. The connections of structured, semi-structured, and unstructured data are achieved by using link-based integration (Gröger et al., 2014). In the Delivery Zone and the Explorative Zone, data is built according to the end-users and analytics application needs.

Step 7: Design Metadata as Enabler. Since metadata is also data, the first six steps of the data lake design process should be followed to the metadata. Metadata can be structured or semi-structured, and is ingested similarly to data. The data flow concept remains the same for both metadata and data. In terms of data organization, metadata remains continuous and run through all the data lake zones. Because of highly connected metadata, it is stored using Neo4J, a graph database management system. The metadata is modeled using Apache Atlas.

Step 8: Design Data Processes. Data processes are divided into data lifecycle and data pipelining processes. Data lifecycle processes involved the data cycle from creation, storage, enhancement and disposal. During each stage of its lifecycle, the metadata that is relevant is captured and stored together with the data. The data pipelining processes begin from the bulk ingestion of raw data into the Landing Zone that serves as the temporary storage for the data before further processing. Data is moved from the Landing Zone to other zones via ETL processes. The concepts of data security, privacy, and quality are effectively applied in data lifecycle processing (e.g. access control and change management) and data pipelining processes (e.g. anonymization techniques).

Step 9: Design Metadata as a Feature. We leverage concepts that offer features beyond metadata as enabler, such as a data catalog to facilitate data access (Chessell et al., 2014).

3.2 Data Sources

The three main sources of data can be classified as DMP (Data Mediated by Process), DGM (Data Generated by Machine), and DHO (Data from Human Origin) (Ouafiq et al., 2022).

The data produced during agricultural activities and procedures is known as the DMP. It can be manually saved in file formats like CSV or automatically stored in information systems (IS). DMP is typically generated from IS, e.g., CRM, ERP, and Laboratory Information Management Systems (LIMS). It includes data such as crop yields, fertilizer applications, and irrigation schedules used in the fields. The data could be stored in relational database management systems (RDBMS).

The data created by IoT devices, smart machines, intelligent robots, or drones that are distributed around the farm and gather data using sensors, smart devices, and satellite imagery (e.g. by pH, temperature, moisture). This data is known as the DGM. This data mainly comes in the form of streaming data, micro-batches, and small batches that can be processed in real time.

The DHO is the data gathered from human inputs, observations, and experiences, relating to agricultural activities, primarily from social media and manual files, thus it can be considered a challenge to consider (e.g. field notes and survey responses) (Raif et al., 2022).

3.3 Big Data Technologies in Data Lake Architecture

It is critical to examine specific big data technologies that underpin our architecture by facilitating data flow, storage, processing, and analytics across all zones.

Data Storage: HDFS. Hadoop Distributed File System (HDFS) is a file system developed for storing large datasets reliably across multiple worker nodes in a cluster. HDFS can store unstructured, semi-structured, and structured data. It provides fault tolerance and high aggregate bandwidth as data blocks are replicated across nodes. HDFS is master-slave architecture using NameNode (master) to manage file system namespace, regulate client accesses, DataNodes (slaves) store data, serve read and write requests (White, 2012).

Data Ingestion: Sqoop and Flume. Apache Sqoop is useful in transferring large amounts of data between Hadoop and structured data stores like relational databases. It supports single table load or free-form SQL query as well as saved jobs that allow you to import updates made to a database since the last import (Gupta & Giri, 2018). Apache Flume is a distributed, fault-tolerant, and available service designed for collecting, aggregating, and moving large amounts of log data from different sources to a centralized repository. It uses a transaction-based messaging mode for data consistency and integrity in the whole process of data ingestion. It also supports batch ingestion (Gupta & Giri, 2018).

Data Processing: Spark. Apache Spark is an open-source data processing engine that is an efficient, large-scale data processing on clusters. It supports high-level language APIs in Java, Scala, Python, and R and an optimized engine that works with general computation graphs (White, 2012). Spark's key abstraction is the Resilient Distributed Dataset (RDD), which represents a read-only collection of objects partitioned across a cluster that can be rebuilt if a partition is lost. Spark supports batch processing, real-time streaming, machine learning, and graph processing (Gupta & Giri, 2018).

Data Exposition: Hive, HBase and Elasticsearch. Apache Hive, which is a data warehousing system built on Hadoop allows the process of reading, writing, and managing large datasets by SQL query. Hive is built around the concept of the table, which is a structured table corresponding to a directory in the HDFS. These directories are further organized into partitions, and each partition is split into buckets (Gupta & Giri, 2018). Apache HBase is a column-oriented, non-relational database managed by HDFS. It gives real-time access to large datasets. HBase follows a master-slave architecture, with a master node responsible for regulating access and updates to data, and multiple region servers handling data storage and retrieval (White, 2012). Elasticsearch is a distributed, RESTful search and analytics engine that can index and search parallel large volumes of data in near real-time. It supports full-text search of HTTP web interface and documents that are schema-free and written in JSON format. Elasticsearch supports search types from structured, semi-structured, and unstructured data.

Stream Processing: Apache Kafka. Kafka is a distributed stream processing platform that allows publishing and subscribing to streams of records. Kafka operates as a message queue and facilitates the transfer of reliable data between app systems and services (Gupta & Giri, 2018).

Workflow Scheduling: Oozie and Airflow. Apache Oozie is an open-source scheduling system used for managing Hadoop jobs. In Oozie, the job workflows are represented by a collection of control flow and action nodes presented as a directed acyclic graph (DAG) with the beginning and end control flow nodes for controlling the execution (White, 2012). Apache Airflow is the platform for creating, scheduling, and monitoring workflows programmatically. It can be used to define workflows as DAGs, in which tasks are executed by workers that run from the beginning to the end based on set dependencies.

Data Visualization: Superset. Apache Superset is an open-source web application for data visualization and exploration built on a comprehensive business intelligence platform. It offers a user-friendly interface for users to create and share interactive dashboards with various charts, tables, and maps.

User Interface and Management: Hue. Hue (Hadoop User Experience) is an open-source platform for web interfaces for interacting with Hadoop clusters and analyzing data. It offers a file browser for HDFS, a query editor for Hive and Impala, a shell, and an Oozie workflow/coordinator designer and dashboard.

Data Governance: Atlas. Apache Atlas is an open-source framework for data governance and metadata management that helps organizations manage their data assets from different platforms and formats. It is an extensible, scalable platform that

facilitates the management and storage of metadata which provides tools for data discovery, classification, and lineage.

4 Results and Discussions

4.1 Technical Architecture

The era of data-driven farming is associated with the phenomenal growth in the volume of data captured from different sources of smart crop farming. There is an urgent need to design a reliable and scalable data lake architecture. The proposed architecture provides a fundamental view of the data lifecycle from ingestion to consumption and how its key components interact while preserving the flexibility, scalability, and performance, as shown in Fig. 5.

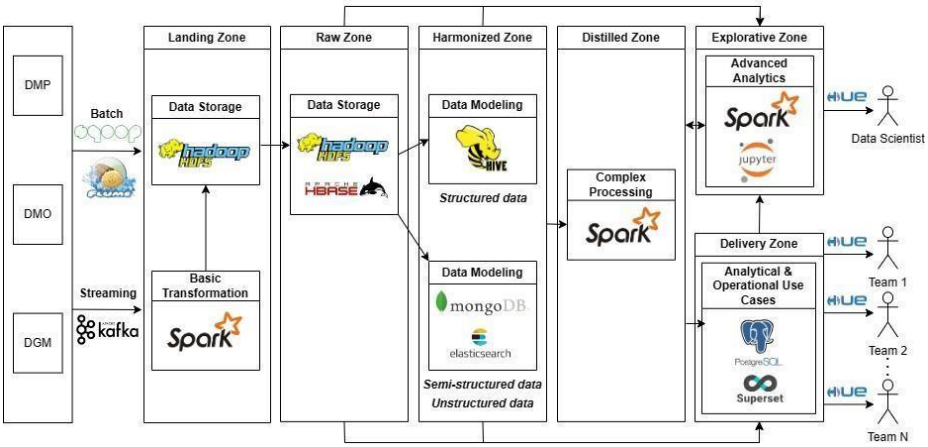

Fig. 5. Smart Crop Farming Data Lake Technical Architecture

Landing Zone. The Landing Zone acts as a buffer, enabling data to be quickly ingested and then moved toward the Raw Zone. The DMP and some types of the DHO are ingested into the Landing Zone using batch processing techniques (e.g. Sqoop for relational databases, Flume for manual files). Nevertheless, the DGM and certain types of the DHO are ingested through streaming processes (e.g. Kafka for sensors and real-time feedback). Data in this zone is mostly in its original raw format, retaining its granularity and schema from the source systems. However, basic transformations such as string character adjustments or timing formats may be used. The Landing Zone is managed to protect data quality and integrity but not designed for data history. After ingestion, HDFS is utilized to temporarily store data.

Raw Zone. The Data in the Landing Zone is transferred to the HDFS storage in the Raw Zone. The data is arranged and put in a hierarchical directory structure within HDFS. The Raw Zone also employs HBase to store and process the real-time data. The Raw Zone is the main repository for raw data ingested from various sources. For

properties, the Raw Zone data can not be modified or deleted to preserve data integrity and consistency. However, the rapid growth of data volume in sensor measurements and the legal terms require that some flexibility is offered in data manipulation and deletion. It means a compromise which can be made between space reduction, compliance, and data completeness. Data scientists are granted access to the Raw Zone. They can copy data from the Raw Zone to other zones. Updates to the data of the Raw Zone are saved as a new record with timestamps to track the changes.

Harmonized Zone. The Harmonized Zone is a part of the Raw Zone which is copied or viewed on demand without any changes to the original data. Data is critically linked to master data management, and the Harmonized Zone allows access to the master data once it has been cleansed and validated (Otto, 2012). Hive creates structured schemas and enables SQL-like querying for structured data. The data modeling in Hive consists of the schema definition, partitioning of the data based on relevant attributes, and the creation of external tables that point to the data stored in HDFS (Benjelloun et al., 2023). MongoDB is used for unstructured and semi-structured data storage and Elasticsearch indexes the data based on search queries. The same properties and the user access applied in the Raw Zone. Data from source systems are combined into a consolidated schema. Data Vault is adopted, where different partial schemas are built for data sources and contexts rather than an overarching schema.

Distilled Zone. The Distilled Zone gives analytical tasks the highest priority, but its primary focus is on data compilation. Data granularity can be changed. Complex processing techniques, including complex computations, data enrichment, or business logic implementation in a particular domain, are employed to transform the data. Moreover, the data schema in the Distilled Zone might be slightly modified to meet the requirements of particular use cases. This is applicable to data processing in both batches and streams. Data scientists, systems and processes can access both the Harmonized Zone and the Distilled Zone. Moreover, domain experts in different fields can also access the data in the Distilled Zone. The Distilled Zone is modeled using Data Vault. In-memory processing and batch and streaming data support provided by Spark allow us to design real-time analytics and decision support systems.

Explorative Zone. The Explorative Zone is a sandbox where only data scientists can freely explore, experiment with, and analyze data. Data scientists can transform data by modifying its granularity, schema, syntax, and semantics as they deem fit for the particular analytical purpose. However, access to sensitive data is strictly regulated. The Explorative Zone is not governed and there are no data logging requirements for this zone. Moreover, the Explorative Zone is non-persistent. Nonetheless, if an analysis yields favorable findings, it can be sent to the Distilled Zone for further refinement and integration before being removed from the Explorative Zone. Data scientists are not limited to any particular modeling method. Spark supports various programming languages including Python and Scala, and has a wide variety of libraries such as MLlib for machine learning and GraphX for graph processing (e.g. to experiment with algorithms, tune hyperparameters, and evaluate model performance). Jupyter Notebook performs iterative data exploration, prototyping, and visualization.

Delivery Zone. In the Delivery Zone, subsets of data are customized to the specific requirements of the analytical and operational use cases. Data in the Delivery Zone is

governed. The data is also stored persistently, unless the use case for which it was processed for is no longer relevant or required. The Delivery Zone is developed to provide users regardless of their technical literacy. The data in this zone is structured and presented in such a way that users find the data easily and export it to their analysis tools. The method in this zone is adaptable and can be adjusted to different use-case situations (e.g. star schema for OLAP). PostgreSQL efficiently stores and manages structured data in the Delivery Zone. It is a scalable base for data warehousing and reporting that enables users to run complex queries and retrieve data fast and efficiently. Superset is utilized to create interactive dashboards, reports, and visual analytics.

4.2 Discussions

The integration of various data sources enables the development of complex predictive models and decision support tools to help farmers adapt to changing and evolving climatic conditions. For example, machine learning algorithms can be deployed on the large repository of historical weather patterns, soil moisture levels, and crop yield data to adjust the irrigation schedule. Moreover, data lake architecture makes it possible to develop a warning system, and climate-related hazard analysis tools through the use of real-time data.

In addition, BRAID hybrid processing architecture enables the utilization of batch processing findings to stream processing. For instance, a crop disease detection model based on machine learning can be trained by analyzing crop images and will be used to identify potential outbreaks in real-time streams of drone imagery. The results can be stored persistently in the data lake for later use. The application of a zone-based data lake architecture complements existing IoT infrastructure, sensors, and analytics capabilities (Munshi & Yasser, 2017). The data lake provides the entire data cycle: ingestion, storage, processing, analysis, and visualization. Furthermore, it integrates with advanced technologies (e.g. AI, deep learning, and machine learning) to find hidden patterns, predict crop yields, detect crop diseases, and optimize resource allocation.

Metadata management helps solve the problem of standardization and consistency across data sources (e.g. different sensors use different units of measurement for the same parameter). Metadata management provides a structured and standardized way to describe and annotate the raw data with relevant context and semantics. Using graph technologies such as Neo4j and Apache Atlas, metadata enables flexible search and intuitive exploration of relationships between heterogeneous datasets from different sources (Hai et al., 2016). Besides, including the ETL processes and data lineage in the metadata facilitates the automation of data preprocessing tasks.

5 Conclusion

Smart crop farming focuses on advanced technologies and information systems to improve crop cultivation process and enhance yield, quality, and minimize environ-

mental impacts. The proposed zone-based data lake architecture adopts the zone reference model by Giebler et al. (2020). The development of this architecture follows the nine-step DLAF, which ensures a comprehensive design by considering the interdependencies between various aspects. This architecture addresses the challenges associated with managing, integrating, and analyzing the vast amounts of heterogeneous data generated by offering a dependable, flexible, and secure environment for data-driven decisions. This architecture effectively integrates both batch and real-time processing capabilities through BRAID architecture to support the development of predictive models, real-time monitoring systems, and decision-support tools tailored to the specific needs of Vietnamese farmers. An integrated metadata management system also ensures compliance, data quality and lineage to facilitate efficient and useful agricultural data.

Addressing the immense challenges posed by climate change and ensuring food security for the nation, the proposed data lake architecture enables crop yield estimation, disease detection, resource optimization, and climate risk assessment. It incorporates robust data governance policies and tools, safeguarding sensitive agricultural data through access controls and ensuring adherence to standards. Different big data technologies are incorporated into a proposed technical architecture, which serves as a foundation for initiating a data-driven strategy in the smart crop farming domain in Vietnam.

Acknowledgments. This study was funded by the UEH College of Technology and Design, UEH University.

References

1. Benjelloun, S., El Aissi, M. E. M., Lakhrissi, Y., & El Haj Ben Ali, S. (2023). Data lake architecture for smart fish farming data-driven strategy. *Applied System Innovation*, *6*(1), 8.
2. Chessell, M., Scheepers, F., Nguyen, N., van Kessel, R., & van der Starre, R. (2014). Governing and managing big data for analytics and decision makers. *IBM Redguides for Business Leaders*, *252*.
3. Dixon, J. (2010). *Pentaho, Hadoop, and Data Lakes*. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/
4. Elijah, O., Rahman, T. A., Orikumhi, I., Leow, C. Y., & Hindia, M. N. (2018). An overview of Internet of Things (IoT) and data analytics in agriculture: Benefits and challenges. *IEEE Internet of things Journal*, *5*(5), 3758-3773.
5. Giebler, C., Gröger, C., Hoos, E., Eichler, R., Schwarz, H., & Mitschang, B. (2021). The data lake architecture framework. In *BTW 2021* (pp. 351-370). Gesellschaft für Informatik, Bonn.
6. Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2019). Leveraging the data lake: current state and challenges. In *Big Data Analytics and Knowledge Discovery: 21st International Conference, DaWaK 2019, Linz, Austria, August 26–29, 2019, Proceedings 21* (pp. 179-188). Springer International Publishing.
7. Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2019b). Modeling data lakes with data vault: practical experiences, assessment, and lessons learned.

In *Conceptual Modeling: 38th International Conference, ER 2019, Salvador, Brazil, November 4–7, 2019, Proceedings 38* (pp. 63-77). Springer International Publishing.

8. Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2020, October). A zone reference model for enterprise-grade data lake management. In *2020 IEEE 24th International Enterprise Distributed Object Computing Conference (EDOC)* (pp. 57-66). IEEE.
9. Giebler, C., Stach, C., Schwarz, H., & Mitschang, B. (2018). A Hybrid Processing Architecture for Big Data.
10. Gorelik, A. (2019). *The enterprise big data lake: Delivering the promise of big data and data science*. O'Reilly Media.
11. Gröger, C., Schwarz, H., & Mitschang, B. (2014, September). The deep data warehouse: link-based integration and enrichment of warehouse data and unstructured content. In *2014 IEEE 18th International Enterprise Distributed Object Computing Conference* (pp. 210-217). IEEE.
12. Gupta, S., & Giri, V. (2018). *Practical Enterprise Data Lake Insights: Handle Data-Driven Challenges in an Enterprise Big Data Lake*. Apress.
13. Hai, R., Geisler, S., & Quix, C. (2016, June). Constance: An intelligent data lake system. In *Proceedings of the 2016 international conference on management of data* (pp. 2097-2100).
14. Hoang, H. G., & Tran, H. D. (2023). Smallholder farmers' perception and adoption of digital agricultural technologies: An empirical evidence from Vietnam. *Outlook on Agriculture*, *52*(4), 457-468.
15. Inmon, B. (2016). *Data Lake Architecture: Designing the Data Lake and avoiding the garbage dump*. Technics Publications, LLC.
16. Lioutas, E. D., Charatsari, C., La Rocca, G., & De Rosa, M. (2019). Key questions on the use of big data in farming: An activity theory approach. *NJAS-Wageningen Journal of Life Sciences*, *90*, 100297.
17. Loc, H. H., Van Binh, D., Park, E., Shrestha, S., Dung, T. D., Son, V. H., . . . Seijger, C. (2021). Intensifying saline water intrusion and drought in the Mekong Delta: From physical evidence to policy outlooks. *Science of the Total Environment*, *757*, 143919.
18. Madsen, M. (2015). How to Build an enterprise data lake: important considerations before jumping in. *Third Nature Inc*, 13-17.
19. Miloslavskaya, N., & Tolstoy, A. (2016). Big data, fast data and data lake concepts. *Procedia Computer Science*, *88*, 300-305.
20. Minh, N. D., Son, D. H., & Van Trinh, M. (2019). A Review of Precision Agriculture in Rice Production in Vietnam.
21. Ministry of Agriculture and Rural Development. (2021). *The hallmark of Vietnamese rice*. https://www.mard.gov.vn/en/Pages/the-hallmark-of-vietnamese-rice.aspx
22. Misra, N., Dixit, Y., Al-Mallahi, A., Bhullar, M. S., Upadhyay, R., & Martynenko, A. (2020). IoT, big data, and artificial intelligence in agriculture and food industry. *IEEE Internet of things Journal*, *9*(9), 6305-6324.
23. Munshi, A. A., & Yasser, A.-R. M. (2017). Big data framework for analytics in smart grids. *Electric Power Systems Research*, *151*, 369-380.
24. Nguyen, H. Q., & Warr, P. (2020). Land consolidation as technical change: Economic impacts in rural Vietnam. *World Development*, *127*, 104750.
25. Nguyen, V. D., Pham, T. C., Le, C. H., Huynh, T. T., Le, T. H., & Packianather, M. (2023). An innovative and smart agriculture platform for improving the coffee value chain and supply chain. In *Machine Learning and Mechanics Based Soft Computing Applications* (pp. 185-197). Springer.

26. Otto, B. (2012). How to design the master data architecture: Findings from a case study at Bosch. *International journal of information management*, *32*(4), 337-346.
27. Ouafiq, E. M., Saadane, R., & Chehri, A. (2022). Data management and integration of low power consumption embedded devices IoT for transforming smart agriculture into actionable knowledge. *Agriculture*, *12*(3), 329.
28. Patel, P., Wood, G., & Diaz, A. (2017). Data lake governance best practices. *The DZone Guide to Big Data-Data Science Advanced Analytics*, *4*, 6-7.
29. Pivoto, D., Waquil, P. D., Talamini, E., Finocchio, C. P. S., Dalla Corte, V. F., & de Vargas Mores, G. (2018). Scientific development of smart farming technologies and their application in Brazil. *Information processing in agriculture*, *5*(1), 21-32.
30. Parente, S. (2020). *The design of a data lake architecture for the healthcare use case: problems and solutions.*
31. Raif, M., Chehri, A., & Saadane, R. (2022). Data architecture and big data analytics in smart cities. *Procedia Computer Science*, *207*, 4123-4131.
32. Ravat, F., & Zhao, Y. (2019). Data lakes: Trends and perspectives. In *Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30* (pp. 304-313). Springer International Publishing.
33. Rodríguez, M. A., Cuenca, L., & Ortiz, Á. (2019). Big data transformation in agriculture: From precision agriculture towards smart farming. In *Collaborative Networks and Digital Transformation: 20th IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2019, Turin, Italy, September 23–25, 2019, Proceedings 20* (pp. 467-474). Springer International Publishing.
34. Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, *56*(1), 97-120.
35. Sawadogo, P. N., Scholly, E., Favre, C., Ferey, E., Loudcher, S., & Darmont, J. (2019). Metadata systems for data lakes: models and features. In *New Trends in Databases and Information Systems: ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8–11, 2019, Proceedings 23* (pp. 440-451). Springer International Publishing.
36. Sharma, B. (2018). *Architecting data lakes: data management architectures for advanced business use cases*. O'Reilly Media.
37. Smajgl, A., Toan, T. Q., Nhan, D. K., Ward, J., Trung, N. H., Tri, L., . . . Vu, P. (2015). Responding to rising sea levels in the Mekong Delta. *Nature Climate Change*, *5*(2), 167-174.
38. Stach, C., Bräcker, J., Eichler, R., Giebler, C., & Mitschang, B. (2022). *Simplified Specification of Data Requirements for Demand-Actuated Big Data Refinement*. J. Data Intell., 3(3), 366-400.
39. White, T. (2012). *Hadoop: The definitive guide*. O'Reilly Media, Inc.
40. Wieder, P., & Nolte, H. (2022). Toward data lakes as central building blocks for data management and analysis. *Frontiers in big Data*, *5*, 945720.
41. Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big data in smart farming–a review. *Agricultural systems*, *153*, 69-80.
42. World Bank. (2021). *Agricultural land (% of land area)*. https://data.worldbank.org/indicator/AG.LND.AGRI.ZS
43. World Bank. (2022). *Employment in agriculture (% of total employment) (modeled ILO estimate)*. https://data.worldbank.org/indicator/SL.AGR.EMPL.ZS
44. Zhao, Y., Megdiche, I., Ravat, F., & Dang, V. N. (2021, July). *A zone-based data lake architecture for IoT, small and big data*. In Proceedings of the 25th International Database Engineering & Applications Symposium (pp. 94-102).

45. Zikopoulos, P. (2015). *Big data beyond the hype: A guide to conversations for today's data center*. McGraw-Hill Education.